ФЕДЕРАЛЬНАЯ СЛУЖБА ГОСУДАРСТВЕННОЙ СТАТИСТИКИ

ЭНЦИКЛОПЕДИЯ СТАТИСТИЧЕСКИХ ТЕРМИНОВ

В 8 ТОМАХ

ИНСТРУМЕНТАЛЬНЫЕ МЕТОДЫ СТАТИСТИКИ

ТОМ 2 РАЗДЕЛ

МОСКВА 2011

Аннотация

2 том Энциклопедии статистических терминов описывает фундаментальные вероятностно-статистические понятия, категории и термины, а также инструментальные методы математической и прикладной статистики, эконометрики, которые должны рассматриваться и применяться как своеобразный унифицированный язык, пригодный для общения специалистов разных стран мира. Термины (их длина, формулировка, последовательность используемых слов, символы их представления, условные обозначения, сокращения и т.д.) унифицированы в формате международных библиотечных и издательских стандартов.

Содержание

Подраздел 2.1. Теория вероятностей и математическая статистика

Рубрика 2.1.1. Вероятностные методы

А

АКСИОМАТИКА КОЛМОГОРОВА (СИСТЕМА АКСИОМ КОЛМОГОРОВА)

система аксиом, лежащих в основе построения вероятностных моделей экспериментов с исходами (явлениями) из множества Ω. В А.К. некоторое произвольное множество Ω принимается за множество элементарных событий. Подмножество A множества Ω отождествляется с событием A. При этом сумма событий A и B понимается как объединение подмножеств A и B. Произведение событий A и B – пересечение подмножеств A и B, а противоположное событие $\overline{A}$ – дополнение к подмножеству A (в Ω).

Пусть задано некоторое множество Ω, которое назовём множеством элементарных событий. Фиксируем некоторую систему F подмножеств множества Ω; эти подмножества названы просто событиями. Потребуем, чтобы выполнялись следующие условия:

если A – событие, то $\overline{A}$ – тоже событие;

если $A_1, A_2, A_3, \ldots$ – события,

то $A_1 + A_2 + A_3 + \ldots$ – тоже событие.

Множество Ω, согласно приведённым условиям, будет являться событием. Система F в этом случае является σ-алгеброй. Отметим, что пара (Ω, F) в этом случае задает измеримое пространство.

Примем определения: события A и B называются несовместными, если A и B не имеют (как подмножества) общих элементов; множество Ω называется достоверным событием, множество $\overline{\Omega} = \varnothing$ – невозможным событием.

Сформулируем аксиомы, задающие понятие вероятности – аксиома 1: каждому событию A поставлено в соответствие неотрицательное число $P(A)$, называемое вероятностью события A; аксиома 2: если события $A_1, A_2, \ldots$ попарно несовместны, то

$$P(A_1 + A_2 + \ldots) = P(A_1) + P(A_2) + \ldots.$$

Данная аксиома носит название аксиомы счётной аддитивности; аксиома 3: $P(\Omega) = 1$. Из введённых аксиом видно, что аксиоматика *теории вероятностей* существенно опирается на аппарат теории множеств и теории меры.

Аксиомы Колмогорова дают весьма удобную математическую схему для исследования конкретных теоретико-вероятностных задач, точнее, для описания опытов со случайными исходами. Вероятностная схема, для которой принято обозначение $(\Omega, F, P(A))$, включает три объекта: множество Ω, называемое пространством элементарных событий; систему F подмножеств множества Ω (σ-алгебра подмножеств Ω), удовлетворяющих условиям аксиом 1 и 2; функцию $P(A)$, определённую на множестве событий и удовлетворяющую аксиомам 1, 2, 3.

При построении моделей экспериментов, призванных описывать вероятностные связи «явлений с условиями», следует оговаривать, при каком «комплексе условий» эти эксперименты рассматриваются. Однако следует отметить, что при изложении теоретических положений теории вероятностей «комплекс условий» не упоминается и по умолчанию предполагается данным. Заметим, что различные комплексы условий для одного и того же эксперимента могут приводить к различным вероятностным моделям.

Пример построения вероятностной модели. Пусть осуществляется трёхкратное подбрасывание монеты. Пространство элементарных событий (исходов) Ω состоит из восьми точек: $\Omega = \{ГГГ, ГГР, \ldots, РРР\}$, и если «комплекс условий» позволяет записать (зафиксировать, «измерить» и т.п.) результаты всех трёх подбрасываний, то, скажем, множество $A = \{ГГГ, ГГР, ГРГ, РГГ\}$ является событием, состоящим в том, что выпадет по крайней мере два «герба». Однако если «комплекс условий» позволяет зафиксировать лишь только результат первого подбрасывания, то рассматриваемое множество A уже нельзя назвать событи-

ем, поскольку нельзя дать ни утвердительного, ни отрицательного ответа на вопрос о том, принадлежит ли конкретный исход ω множеству A. Т.о., определение системы подмножеств F, являющейся σ-алгеброй, фиксирует «комплекс условий» при которых рассматривается эксперимент.

А.К. позволяет доказать ряд свойств вероятности (вероятностной меры):

если $\varnothing$ – пустое множество, то $P(\varnothing)=0$;

если $A,B\in F$,

то $P(A\cup B)=P(A)+P(B)-P(A\cap B)$;

если $A,B\in F$ и $A\subseteq B$, то $P(A)\le P(B)$.

Опираясь на А.К. можно дать определение статистической структуры, играющей важную роль в математическом обосновании математической статистики: пусть $\wp$ – семейство вероятностных мер (распределений) на измеримом пространстве (Ω,F); статистической структурой называется тройка $(\Omega,F,\wp)$. Пространство Ω в этом случае имеет смысл пространства наблюдений, и предполагается, что эти наблюдения отвечают некоторой случайной величине, распределение вероятностей которой априори считается принадлежащим известному семейству $\wp$. Часто используют другое обозначение для семейства $\wp$ с помощью введения индекса θ, называемого параметром: $\wp=\{P_\theta,\theta\in\Theta\}$.

АЛГЕБРАИЧЕСКОЕ ДОПОЛНЕНИЕ

элемента a_{ij} матрицы A – число

$$A_{ij}=(-1)^{i+j}M^i_j,$$

где M_j – минор, определитель матрицы, получающейся из матрицы A путём вычёркивания i-й строки и j-го столбца. Название «А.д.» связано с формулами разложения определителя матрицы по строке (по столбцу):

$$\det A=\sum_{j=1}^{n}a_{ij}A_{ij}=\sum_{i=1}^{n}a_{ij}A_{ij}\ .$$

Понятие А.д. используется, напр., при вычислении *коэффициента детерминации* и *частных коэффициентов корреляции* по корреляционной матрице R исследуемого многомерного признака.

АНАЛИЗ КАНОНИЧЕСКИХ КОРРЕЛЯЦИЙ

общая теория и практическое применение структуры связей между двумя совокупностями *случайных величин* (между двумя случайными векторами) $X^{(1)}$ и $X^{(2)}$. А.к.к. позволяет находить макс. корреляционные связи между двумя группами случайных величин. Эта зависимость определяется при помощи новых аргументов – канонических величин, вычисленных как линейные комбинации исходных признаков. Новые канонические величины выбираются т.о., чтобы новые координаты непосредственно указывали значение корреляции. В каждой группе отыскиваются линейные комбинации исходных величин, имеющие макс. корреляцию, они и являются первыми координатами новых систем. Затем в каждой группе рассматриваются следующие линейные комбинации, у которых корреляции больше, чем между любыми другими линейными комбинациями, некоррелированными с первыми линейными комбинациями. Построение продолжается до тех пор, пока не будут полностью получены две новые координатные системы.

А.к.к. реализуется в форме задачи нахождения собственных значений и собственных векторов от некоторой функции корреляционной матрицы исходных признаков $B=R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$, где R_{11} R_{22} – корреляционные матрицы признаков $X^{(1)}$ и показателей $X^{(2)}$, размерности которых соответственно равны $(p_1\times p_1)$ u $(p_2\times p_2)$; R_{12} – матрица взаимных корреляций первой и второй групп $(R_{21}=R_{12}^T)$. Собственные значения B, ранжированные по убыванию, равняются квадратам канонических корреляций ρ_k^2; левые и правые собственные векторы – соответствующим каноническим переменным групп исходных признаков $X^{(1)}$ и $X^{(2)}$. С точки зрения А.к.к. обе группы равноценны. Для разрешимости задачи требуется, чтобы корреляционные матрицы R_{11} и R_{22} были положительно определены. В противном случае следует один или

несколько признаков-показателей из рассмотрения исключить.

Свойства канонических переменных: являются линейными комбинациями исходных показателей соответствующих групп; канонические переменные одной группы взаимно некоррелированны; канонические переменные выбраны т.о., чтобы соответствующие канонические корреляции были максимальны; канонические переменные упорядочены по мере убывания соответствующих корреляций; число используемых канонических корреляций обычно значительно меньше числа исследуемых показателей p_2.

Канонические корреляции всегда неотрицательны, причём их осн. свойства совпадают со свойствами *множественных коэффициентов корреляции*. Чем больше канонические корреляции, тем сильнее связаны рассматриваемые группы признаков $X^{(1)}$ и показателей $X^{(2)}$.

Значимость канонических переменных проверяется при помощи критерия χ^2. Если вычислено p_1 канонических корреляций, то для каждого m (m=1,2,…,p_1) следует проверить гипотезы:

$H_0^m : \rho_m = \rho_{m+1} = ... = \rho_{p_1} = 0,$

$H_1^m : \rho_m \neq 0$ (по крайней мере ρ_m отличается от нуля).

При этом учитывается, что

$\rho_m > \rho_{m+1} > ... > \rho_{p_1}$.

Наблюдаемое значение статистики рассчитывается по формуле:

$$\chi^2_{набл} = -\left\{ N - m - \frac{1}{2}(p_1 + p_2 + 1) + \sum_{k=1}^{m-1} r_k^2 \right\} \ln \prod_{k=m}^{p_1} (1 - r_k^2),$$

где N– объём выборки, r_k^2 – оценка канонического коэффициента детерминации.

Если значение

$\chi^2_{набл} > \chi^2_{кр}(\alpha, \nu = [(p_2 - m + 1)(p_1 - m + 1)]),$

то принимается гипотеза $H_1^m : \rho_m \neq 0$. Процедура повторяется для следующей (m+1) канонической корреляции. Если вычисленное значение χ^2 – статистики меньше соответствующего табличного значения, то нулевая гипотеза H_0^{m+1} не отвергается, т.е. зависимость между группами исследуемых признаков уже описана каноническими переменными с индексами 1, 2, …, m. Если при некотором значении m_0 нулевая гипотеза не отвергается, т.е. если каноническая корреляция ρ_{m_0} равна нулю, то равны нулю и все последующие ρ_m при $m=m_0+1, m_0+2,\ldots,p_1$. Следует интерпретировать только такие канонические переменные, которые соответствуют значимым каноническим корреляциям.

В процессе А.к.к. исходные данные $X^{(1)}$ и $X^{(2)}$ приводятся к стандартизованному виду, поэтому коэффициенты в выражениях для канонических переменных характеризуют силу влияния соответствующих исходных признаков и показателей, что позволяет получить их ранжированные последовательности. Отсев несущественных переменных может осуществляться на основе многошаговой процедуры, при которой на каждом шаге отбрасывается только одна переменная, наименее существенная в исходной последовательности. Для сравнения канонических корреляций исходного ρ_k и ρ_{k+1} наборов факторов используется z-преобразование Фишера:

$$t_{набл} = (z_k - z_{k+1})\sqrt{\frac{N-3}{2}}, \text{ где } z_l = \frac{1}{2}\ln\frac{1+r_l}{1-r_l}$$

$t_{набл}$ – имеет нормированный нормальный закон распределения; l –индекс, относящийся к признакам вектора $X^{(1)}$. Если канонические корреляции ρ_k и ρ_{k+1} отличаются незначимо, процесс сокращения продолжается.

При А.к.к. процедура отсева учитывает всю сложность структуры связей как внутри групп признаков и показателей, так и между этими группами. Признак, значимо влияющий хотя бы на один показатель и являющийся значимым для других, уже не может быть отброшен. Процеду-

ра отсева основывается на принципе дополнительности: признаки $X^{(2)}$ исключаются с учётом того, какие показатели $X^{(1)}$ исключаются.

А.к.к. на практике облегчает интерпретацию структуры взаимосвязей, особенно в тех случаях, когда взаимосвязь между двумя множествами наблюдаемых величин достаточно полно описывается корреляцией между несколькими каноническими случайными величинами. А.к.к. используется для научного обоснования системы показателей при проведении *многомерного статистического анализа*, а также как осн. инструментарий в каноническом *факторном анализе*.

АПОСТЕРИОРНОЕ РАСПРЕДЕЛЕНИЕ

условное распределение параметра θ (вероятностная мера на измеримом пространстве (Θ, T)) как случайной величины при наблюдении выборки из множества выборок Ω определённое по *теореме* (*формуле*) *Байеса*.

А.р. позволяет получить *байесовскую оценку* как условное математическое ожидание случайной величины $\theta \in \Theta$ при наблюдении $\omega \in \Omega$. А.р. можно интерпретировать как уточнённое знание о значениях параметра θ после получения дополнительной информации в виде значения *выборки* из выборочного пространства Ω.

АПРИОРНОЕ РАСПРЕДЕЛЕНИЕ

предполагаемое распределение параметра $\theta \in \Theta$ (вероятностная мера на измеримом пространстве (Θ, T)) как случайной величины.

А.р. – важное понятие при использовании *байесовского подхода к оцениванию* параметров в *математической статистике*. А.р. должно быть согласованно с вероятностным пространством наблюдений (Ω, F) в том смысле, что функция правдоподобия должна быть измерима на $(\Omega \times \Theta, F \times T)$. Введение А.р. в статистической задаче имеет целью задание информации перед началом эксперимента или предварительных данных о неизвестном параметре $\theta \in \Theta$. Выбор А.р. в некоторых задачах можно обосновать, но иногда он является достаточно произвольным, в этом случае чаще всего говорят о субъективной вероятности как мнении исследователя о возможности получения того или иного значения из Θ. На основании заданного А.р. учитывая наблюдения $\omega \in \Omega$ с использованием *Байеса теоремы* (*формулы)* определяют *апостериорное распределение*.

АСИМПТОТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ

– распределение, к которому стремится закон распределения последовательности случайных величин $\{X_n\}_{n=1}^{\infty}$ (в том или ином смысле) при $n \to \infty$. Примером А.р. может служить нормальный закон распределения в *центральной предельной теореме*. А.р. играет большую роль в *математической статистике*, в которой рассматриваются статистики при различных объёмах *выборки* как последовательность случайных величин. В этом случае можно говорить о А.р. статистики при бесконечном увеличении объёма выборки.

Б

БАЙЕСА ТЕОРЕМА (ФОРМУЛА)

см. в ст. Теорема (формула) Байеса.

БЕРНУЛЛИ ИСПЫТАНИЯ

последовательность независимых испытаний с двумя случайными исходами («удачей» и «неудачей»), вероятности которых не изменяются от испытания к испытанию. Пусть p – вероятность удачи и q=1 – p – вероятность неудачи, и пусть 1 обозначает наступление удачи, а 0 – наступление неудачи. Тогда вероятность определённого чередования удач и неудач, напр., 10001101011...0 равна

$$pqqqppqpqpp \ldots q = p^m q^{n-1},$$

где m – число удач в рассматриваемом ряду n испытаний. Б.с.и. – важнейшая схема, рассматриваемая в *теории вероятностей*. Схема названа в честь Я.Бернулли, доказавшего свою теорему, для такой последовательности испытаний. Со схемой Б.с.и. связаны многие распространенные дискретные распределения вероят-

ностей. Пусть S_n – случайная величина, равная числу удач в n Б.с.и. Тогда вероятность появления k удач, т.е. события $S_n=k$ равна:

$$P_n(k) = C_n^k p^k q^{n-k}, \quad k = 0,1,\ldots,n,$$

т.е S_n имеет биномиальное распределение. Последнее при $n \to \infty$ аппроксимируется нормальным распределением или распределением Пуассона. Пусть Y – число испытаний до первой удачи, тогда вероятность события Y=k равна:

$$P_n(Y = k) = p\ q^k, \quad k = 0,1,\ldots,n,$$

т.е Y имеет геометрическое распределение. Если Z_s – число неудач, предшествующих s-му появлению удачи, то Z_s имеет отрицательное биномиальное распределение:

$$P(Z_s = k) = C_{s-1}^{k-1} p^k q^{s-k}, \quad s = k, k+1,\ldots$$

См. также *Бернулли теорема.*

БЕРНУЛЛИ ТЕОРЕМА

первая из предельных теорем *теории вероятностей*; является простейшей формой *закона больших чисел*. Б.т. утверждает, что при большом числе *n* повторных независимых испытаний, в которых вероятность наступления некоторого случайного события постоянна, практически достоверно, что относительная частота (частость) *m/n* наступления этого события – величина случайная, как угодно мало отличается от неслучайной величины *p* – вероятности события, т. е практически перестаёт быть случайной. Формулировка Б.т.: частость события в *n* независимых испытаниях, в каждом из которых оно может произойти с одной и той же вероятностью *p*, при неограниченном увеличении числа *n* сходится по вероятности к вероятности *p* этого события в отдельном испытании:

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| \prec \varepsilon\right) = 1.$$

Б.т. дает теоретическое обоснование замены неизвестной вероятности события его частостью, полученной в n повторных независимых испытаниях, проводимых при одном и том же комплексе условий. Напр., если вероятность рождения мальчика нам не известна, то в качестве её значения мы можем принять относительную частоту этого события, которая, как известно по многолетним статистическим данным, составляет приближенно 0,515. Б.т. является звеном, позволяющим связать формальное аксиоматическое определение вероятности с эмпирическим (опытным) законом постоянства относительной частоты. Б.т. даёт возможность обосновать широкое применение на практике вероятностных методов исследования. Б.т. впервые опубликована в книге Я. Бернулли «Искусство предположений» в 1713, переведена на рус. язык в 1913 и в 1986. Первое доказательство Б.т., данное Я. Бернулли, было основано на изучении характера убывания вероятностей в биномиальном распределении и требовало сложных вычислений. Лишь в середине 19 в. П.Л.Чебышев нашёл изящное и краткое её доказательство, основанное на простой оценке, частном случае неравенства Чебышева:

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| \prec \varepsilon\right) \succ 1 - \frac{p(p-1)}{n\varepsilon^2}.$$

БЕТА-РАСПРЕДЕЛЕНИЕ – непрерывная случайная величина X имеет Б.-р. с параметрами (α, β) $(\alpha > 0, \beta > 0)$, если функция плотности определяется выражением:

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} = \dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & x \in (0,1); \end{cases}$$

где $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ – гамма функция;

$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}$ – бета функция Эйлера.

Функция распределения выражается через неполную бета-функцию:

$$F(x)=\frac{1}{B(\alpha,\beta)}\int_0^x u^{\alpha-1}(1-u)^{\beta-1}du\,.$$

Вид функции плотности сильно зависит от параметров распределения α и β (см. рис. 1).

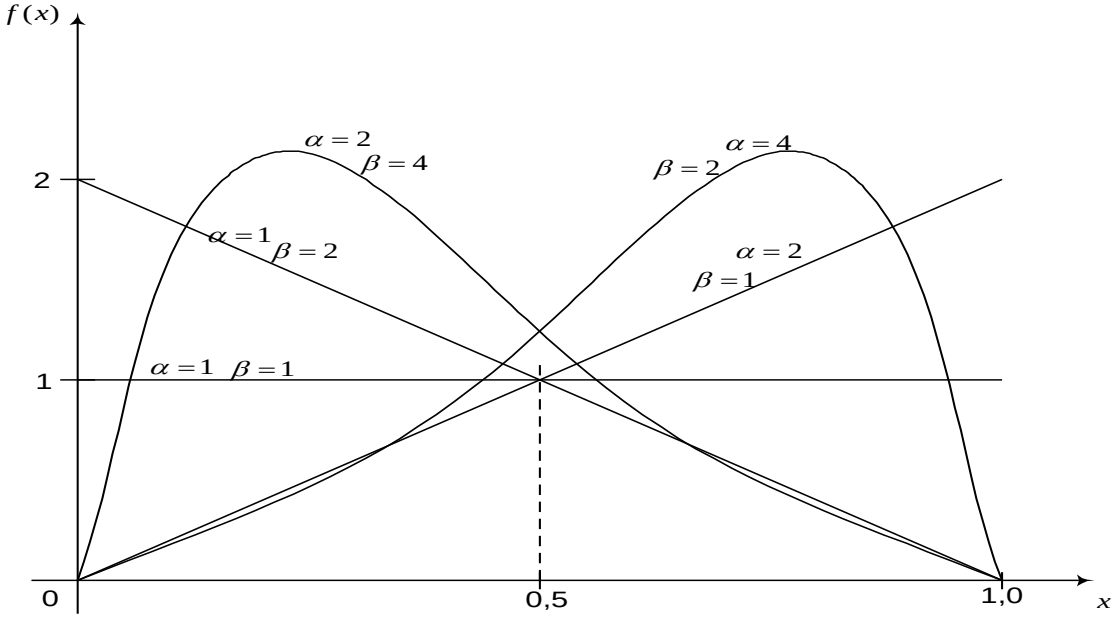


Рис. 1. **График функции плотности Б.-р.**

Числовые характеристики: Среднее: $M[X]=\dfrac{\alpha}{\alpha+\beta}$;

Мода: $M_0=\dfrac{\alpha-1}{\alpha+\beta-2}$, $\alpha>1,\ \beta>1$;

Дисперсия: $D[X]=\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$; Асимметрия: $A_x=\dfrac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$;

Эксцесс: $\varepsilon_x=6\dfrac{\alpha^3-\alpha^2(2\beta-1)+\beta^2(\beta+1)-2\alpha\beta(\beta+2)}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}$.

Распределения многих порядковых статистик сводятся к Б.-р. Если $\alpha=1$, $\beta=1$, то Б.-р. совпадает с равномерным на интервале (0,1).

При $\beta=\alpha+1$ бета – распределение называется обобщённым распределением арксинуса, а при $\alpha=\beta=\dfrac{1}{2}$ – распределением арксинуса.

$$\tilde{\alpha}=\bar{x}\left\{\left[\bar{x}(1-\bar{x})/S^2\right]-1\right\};$$
$$\tilde{\beta}=(1-\bar{x})\left\{\left[\bar{x}(1-\bar{x})/S^2\right]-1\right\},$$

Один из важных случаев возникновения Б.-р. такой: если x_1 и x_2 независимы и имеют *гамма - распределение* с параметрами α и β соответственно, то величина $\dfrac{x_1}{x_1+x_2}$ имеет Б.-р. с плотностью $b_{\alpha,\beta}(x)$. Этим объясняется роль, которую Б.-р. играет в приложениях, так как распределения многих важнейших статистик сводятся к Б.-р.

Статистические оценки параметров распределения определяются по формулам:

где $\bar{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i$; $S^2=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

БУЛЕВА МАТРИЦА ПАРНЫХ СРАВНЕНИЙ

матрица $\{a_{ij}\}$, отражающая результаты сравнений объектов в наиболее простой форме. Её элемент a_{ij} равен единице, если i-й объект не уступает j-му в определённом смысле, и нулю – в противном случае. Если производится сравнение объектов по принадлежности к однородной группе, то единичное значение присваивается в случае их отнесения к одному и тому же классу и нулевое – при отнесении к разным классам. Так как диагональными элементами Б.м.п.с. являются единицы, число её независимых элементов при общем числе сравниваемых объектов N равно N(N-1)/2.

Б.м.п.с. широко используются в экспертном оценивании. Формирование Б.м.п.с. является наиболее простым для эксперта. Считается, что гораздо легче сделать качественное сравнение двух объектов, чем выражать свои предпочтения в балльной или ранговой шкале. Однако разные пары объектов иногда сопоставляются респондентами по разным критериям, что приводит к нетранзитивности предпочтений. Кроме того, трудоёмкость процедуры сбора данных существенно возрастает при увеличении числа сравниваемых объектов. Тем не менее, данные экспертного оценивания при выявлении предпочтений, полученные в виде Б.м.п.с., – наиболее надежны.

Если противоречивость во мнениях о предпочтении одних объектов другим незначительна, степень предпочтения объекта каждого объекта всем анализируемым можно определить как отношение суммы единиц в соответствующей строке матрицы к общему числу сравниваемых объектов. При использовании мнений ряда экспертов и отсутствии существенных противоречий в их суждениях, оценку предпочтений для каждого объекта можно получить как взвешенную среднюю, полученную по всем формируемым Б.м.п.с. Весовыми коэффициентами при этом могут выступать уровни компетентности соответствующих экспертов.

Расчёт собственного вектора Б.м.п.с., соответствующего макс. собственному числу позволяет повысить уровень шкалы изменений признака, по которому производились сравнения. Компоненты этого вектора в отсутствие существенной несогласованности парных сравнений можно использовать как балльные оценки объектов в количественной шкале. При этом степень непротиворечивости можно оценить во мнениях эксперта по отличию макс. собственного числа от потенциально достижимого.

В

ВЕРОЯТНОСТНОЕ ПРОСТРАНСТВО

заданная тройка (Ω, F, P), удовлетворяющая *аксиоматике Колмогорова*.

Если в качестве пространства элементарных исходов Ω взять дискретное (конечное или счётное) множество, а в качестве F – все подмножества множества Ω, то задание некоторой числовой функции $p(\omega)$ на Ω, удовлетворяющей условиям: неотрицательности ($p(\omega) \geq 0$ для любого $\omega \in \Omega$); нормированности ($\sum_{\omega \in \Omega} p(\omega) = 1$), позволяет говорить о задании дискретного В.п. Пусть $\Omega = \{\omega\}$ – *евклидово пространство* или область в евклидовом пространстве и F – σ-алгебра, порождённая измеримыми областями в Ω. Каждому $\omega \in \Omega$ поставим в соответствие число $p(\omega)$, т.е. зададим на Ω числовую функцию, удовлетворяющую условиям: неотрицательности ($p(\omega) \geq 0$ для любого $\omega \in \Omega$) и нормированности ($\int_{\omega \in \Omega} p(\omega) d\omega = 1$). Т.о. определённая тройка (Ω, F, P) – непрерывное В.п.

ВЕРОЯТНОСТЬ

число, заключённое между нулём и единицей, характеризующее, измеряющее степень наступления случайного события (P{A}). Со случайным событием связан случайный эксперимент, испытание в результате повторения которого событие A может появиться или не появиться. Множество исходов испытания, наблюдения составляет пространство элементарных событий. В результате эксперимента может осуществляться лишь одно из элементарных событий, исходов, случаев. Случайное событие является подмножеством элементар-

ных событий и происходит тогда, когда случай благоприятствует ему, т.е. принадлежит подмножеству, определяющему случайное событие A.

В случае конечного множества элементарных исходов в предположении их равновозможности, равновероятности и условия нормировки (сумма В. каждого исхода равна единице), можно заключить, что В. каждого исхода будет равна 1/N, где N – число элементарных событий. В такой ситуации В. случайного события A определяется по формуле:

$$P(A) = \frac{N_1}{N},$$

где N_1 – число случаев, благоприятствующих появлению события A.

В более сложных ситуациях, напр., когда пространство исходов бесконечно или континуально, требуется аксиоматический подход к определению В. случайного события и самого случайного события. Формула, приведённая выше, обычно называется классическим определением В.

Для примера расчёта по этой формуле вычислим В. появления пяти очков при выбрасывании двух игральных костей. Здесь элементарным исходом является любая пара чисел, каждое из которых может по предположению (кости не фальшивые) равновозможно принимать от одного до шести очков. Общее число элементарных событий равно числу различных пар чисел выпадающих на верхних гранях, это число N = 6 x 6 = 36. Число благоприятствующих событию A (выпадение пяти очков) составит $N_1 = 4$, а именно: (1,4), (2,3), (3,2), (4,1) – четыре пары, каждая из которых суммарно дает пять очков.

Отметим, что В. невозможного события (событие, которому соответствует пустое множество исходов) равна нулю, а В. достоверного события (когда событию благоприятствует любой исход из пространства элементарных исходов) равна единице, но не наоборот. В случае бесконечного множества элементарных исходов В. наступления случайного события может оказаться равной нулю (хотя это событие не является невозможным), и В. недостоверного события может равняться единице.

ВЕРОЯТНОСТЬ УСЛОВНАЯ

вероятность события A, вычисленная при условии, что событие B произошло; обозначается $P(A/B)$.

Пример. В урне пять шаров, из них два черных и три белых. Производится испытание: два раза вынимается шар из урны без возвращения. Определить В.у. события A – появление белого шара при втором вынимании; событие B – появление белого шара при первом изъятии.

Если первым вынутым шаром оказался белый, то после этого в ящике осталось два белых шара, а всего – четыре. Поэтому $P(A/B) = \frac{2}{4} = \frac{1}{2}$. Если первым вынутым шаром оказался черный (произошло событие $\overline{B}$), то $P(A/\overline{B}) = \frac{3}{4}$.

ВЕРОЯТНОСТНАЯ БУМАГА

нормальная, специальным образом разграфленная бумага, построенная так, что график функции нормального распределения изображается на ней прямой линией.

Это достигается изменением шкалы на вертикальной оси (см. рис.1). На свойстве «выпрямления» основан простой способ проверки гипотезы о принадлежности данной выборки к нормальной совокупности: если построенная на В.б. эмпирическая функция распределения хорошо аппроксимируется прямой линией, то можно с основанием полагать, что совокупность, из которой взята *выборка*, является приближенно нормальной. Достоинство этого метода состоит в том, что вывод о принадлежности к нормальной совокупности можно сделать без знания численных значений параметров гипотетического распределения.

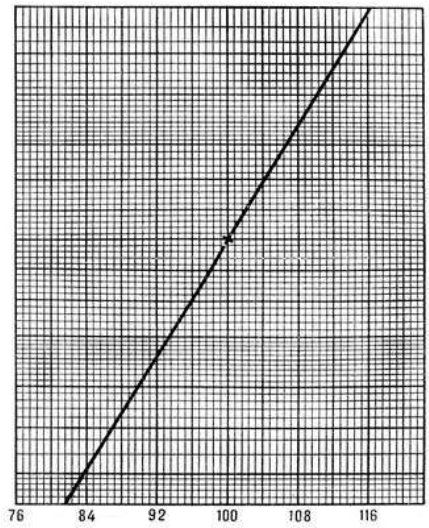


Рис. 1. **Образец В.б. Проведённая линия – график функции нормального распределения со средним 100 и стандартным отклонением 8.**

Этот простой графический метод часто используют для первоначальной прикидки, правдоподобно ли предположение о том, что независимая выборка $x = (x_1,...,x_n)$ взята из нормального распределения. Эта прикидка осуществляется в буквальном смысле «на глазок», поэтому здесь не идёт речь о количественных показателях, таких как вероятность ошибки и т.п.

ВЕРОЯТНОСТНАЯ МЕРА

см. в ст. Вероятность

ВОЗМОЖНЫЕ ЗНАЧЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

возможные значения функции, определённой на множестве элементарных событий.

Случайные величины обозначаются прописными буквами *X, Y, Z,* а их возможные значения — соответствующими строчными буквами *x, y, z.* Например, если *случайная величина X* имеет три возможных значения, то они будут обозначены так: x_1, x_2, x_3.

Возможные значения и их общее число определяются структурой соответствующего *пространства элементарных событий* Ω: каждому элементарному событию ω соответствует свое возможное значение ξ. При этом, правда, может быть, что нескольким элементарным исходам соответствует одно и то же возможное значение анализируемой *случайной величины,* так что, вообще говоря, в конечном *дискретном вероятностном пространстве* число возможных значений случайной величины всегда меньше или равно числу различных элементарных исходов.

Следует отличать теоретически возможные значения *случайной величины* (обозначим их $x_1^0, x_2^0,..., x_i^0,...$ в дискретном случае и просто *x* — в непрерывном) от практически осуществившихся в экспериментах, т.е. от наблюденных ее значений (последние обозначим $x_1, x_2,..., x_n$).

Г

ГАММА-РАСПРЕДЕЛЕНИЕ

непрерывная случайная величина X имеет Г.-р. с плотностью вероятности:

$$f(x) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, & x > 0; \\ 0, & x \le 0, \end{cases},$$

где α и β параметры распределения ($\alpha > 0,\ \beta > 0$);

$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$ – гамма – функция Эйлера.

Функция распределения определяется зависимостью:

$$F(x)=\begin{cases}\dfrac{1}{\beta^{\alpha}\Gamma(\alpha)}\displaystyle\int_0^x u^{\alpha-1}e^{-\frac{u}{\beta}}du, & x>0;\\ 0, & x\le 0.\end{cases}$$

График функции плотности Г.-р. имеет вид (см. рис. 1):

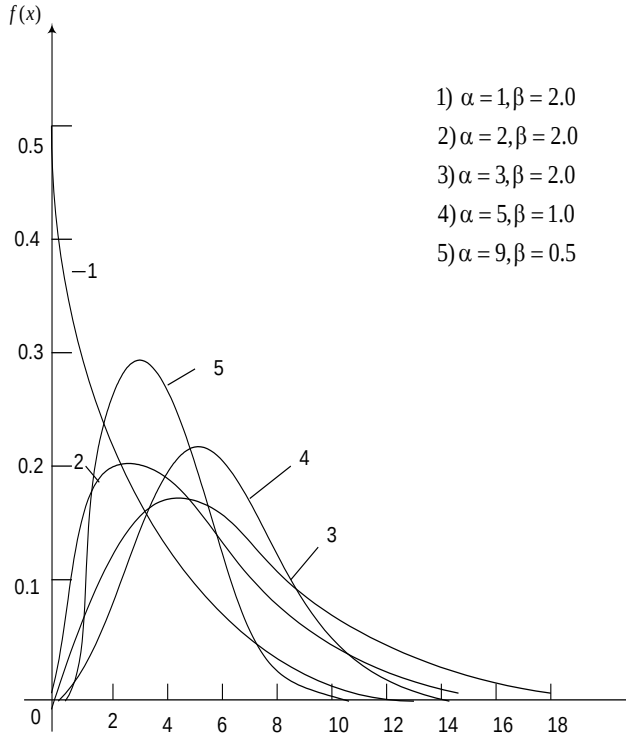


Рис. 1. **График функции плотности Г.-р.**

Числовые характеристики:

Среднее: $M[x]=\beta\alpha$; Мода: $M_0=\beta(\alpha-1),\quad \alpha\ge 1$; Дисперсия: $D[x]=\beta^2\alpha$; Асимметрия: $A_x=\dfrac{2}{\sqrt{\alpha}}$; Эксцесс: $\varepsilon_x=3+\dfrac{6}{\alpha}$.

Г.-р. является непрерывным аналогом отрицательного биномиального распределения. При $\alpha=1$ совпадает с экспоненциальным с $\lambda=\dfrac{1}{\beta}$, а при $\alpha=\dfrac{n}{2}$, $\beta=\dfrac{1}{2}$ – с χ^2 – распределением с n – степенями свободы.

При $\beta=n\mu$ и $\alpha=n$ называется эрланговым распределением с параметрами (n,μ) и описывает распределение длительности интервала времени до появления n событий процесса Пуассона с параметром μ, используемым в теории массового обслуживания и теории надежности.

При $\beta=1$ получаем стандартное Г.-р. с функцией плотности $f(x)=\dfrac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)},\ x>0$.

Сумма любого числа независимых случайных величин, имеющих Г.-р. (с одинаковым параметром масштаба β) $\gamma(\alpha_1,\beta)+\gamma(\alpha_2,\beta)+\ldots+\gamma(\alpha_n,\beta)$ также подчиняется Г.-р., с параметрами $(\alpha_1+\alpha_2+\ldots+\alpha_n,\beta)$. Статистические оценки параметров распределения определяются по формулам:

$$\tilde{\beta}=\frac{S^2}{\bar{x}};\ \tilde{\alpha}=\left(\frac{\bar{x}}{S}\right)^2,\ \text{где}\ \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i;\ S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

ГАММА-ФУНКЦИЯ ЭЙЛЕРА

(Γ – функция, $\Gamma(x)$) – одна из важнейших специальных функций, обобщающая понятие факториала; для целых положительных n равна $\Gamma(n)=(n-1)!=1\cdot 2\ldots(n-1)$; впервые введена Л. Эйлером в 1729. Г.-ф.Э. для действительных $x>0$ определяется равенством

$$\Gamma(x)=\int_0^{+\infty}t^{x-1}e^{-t}dt;$$ другое обозначение:

$$\Gamma(x+1)=\pi(x)=x!$$

Осн. соотношения для Г.-ф.Э.:

$\Gamma(x+1)=x\Gamma(x)$ (функциональное уравнение);

$\Gamma(x)\Gamma(1-x)=\pi/\sin\pi x$ (формула дополнения);

$\Gamma(x)\Gamma(x+1/2)=2^{1-2x}\sqrt{\pi}(2x).$

Частные значения:

$\Gamma(1) = 0! = 1;$ $\Gamma(1/2) = \sqrt{\pi}$.

При больших x справедлива асимптотическая формула Стирлинга: $\Gamma(x+1) \approx \sqrt{2\pi} x^x e^{-x}$.

Через Г.-ф.Э. выражаются некоторые распределения случайных величин, большое число определённых интегралов, бесконечных произведений и сумм рядов. Г.-ф.Э. распространяется и на комплексные значения аргумента.

Д

ДИСКРЕТНОЕ ВЕРОЯТНОСТНОЕ ПРОСТРАНСТВО

см. в ст. Вероятностное пространство.

ДИСПЕРСИЯ ГЕНЕРАЛЬНАЯ

дисперсия одномерной ген. совокупности или, что то же самое, случайной величины x, которая характеризует вариацию значении x относительно математического ожидания MX и определяется по формуле:

$$\boldsymbol{DX} = \begin{cases} \sum (\boldsymbol{x_i} - \boldsymbol{MX})^2 \boldsymbol{p_i} \quad (\boldsymbol{X} - \text{дискретная случайная величина}), \\ \int\limits_{-\infty}^{\infty} (\boldsymbol{x} - \boldsymbol{MX})^2 \boldsymbol{p}(\boldsymbol{x}) \boldsymbol{dx} \quad (\boldsymbol{X} - \text{непрерывная}), \end{cases}$$

$$\boldsymbol{DX} = \begin{cases} \sum\limits_{\boldsymbol{i}} (\boldsymbol{x_i})^2 \boldsymbol{p_i} - (\boldsymbol{MX})^2 \quad (\boldsymbol{X} - \text{дискретная}), \\ \int\limits_{-\infty}^{\infty} \boldsymbol{x}^2 \boldsymbol{p}(\boldsymbol{x}) - (\boldsymbol{MX})^2 \quad (\boldsymbol{X} - \text{непрерывная}). \end{cases}$$

Т.о. определяется как математическое ожидание квадрата отклонения значения признака от среднего значения этого признака в ген. совокупности.

Д.г. линейной функции случайной величины (признака) равна произведению квадрата коэффициента пропорциональности на дисперсию случайной величины. Д.г. признака Y в регрессионном анализе равна сумме дисперсии, связанной с регрессией признака Y на другие переменные и остаточной дисперсии, связанной с «собственным» рассеянием Y около линии регрессии или объясняемой действием неучтенных в регрессии случайных факторов: DY=DYрегр.+DYост.

Д.г. обладает свойством минимизации. Математическое ожидание квадрата отклонения случайной величины X от числа $\boldsymbol{a}$ имеет вид: $M(X-a)^2=DX+(MX-a)^2$ и достигает минимума равного Д.г. при $\boldsymbol{a}$=MX.

В качестве Д.г. может выступать обобщённая дисперсия – величина определителя ковариационной матрицы многомерного признака.

Часто в качестве характеристики рассеяния многомерной случайной величины используется след её ковариационной матрицы, т.е. сумма её диагональных элементов

$Sp(\Sigma)=\sigma_{11}+\sigma_{22}+\ldots+\sigma_{pp}$,

где $\sigma_{11}, \ldots, \sigma_{pp}$ – дисперсии компонент p-мерного случайного вектора, Σ – ковариационная матрица. Д.г. обозначается DX,σ^2.

ДИСПЕРСИЯ УСЛОВНАЯ

для многомерной случайной величины (X,Y) дисперсия случайной величины Y при условии, что X приняла значение x. По определению

$$D(Y \mid X = x) \equiv \sigma_Y^2(x) \equiv M\left[(y - m_Y(X))^2 \mid X = x\right],$$

где $m_Y(X)$ есть условное *математическое ожидание* случайной величины Y относительно случайной величины X. Д.у. иногда называют скедастической функцией Y на $X = x$.

Весьма часто в прикладных исследованиях вместо детерминированной (не случайной) функции – условной дисперсии – рассматривают функцию случайной величины, формально полагая, что $D(Y \mid X) = \sigma_Y^2(x)\Big|_{x=X}$. При таком понимании Д.у. говорят о скедастической функции Y на X.

ДОСТОВЕРНОЕ СОБЫТИЕ

см. в ст. Случайное событие

З

ЗАВИСИМЫЕ И НЕЗАВИСИМЫЕ СОБЫТИЯ

см. в ст. Случайное событие

ЗАКОН БОЛЬШИХ ЧИСЕЛ

ряд теорем, в каждой из которых устанавливается факт приближения средних характеристик большого числа опытов к некоторым определенным постоянным.

Теоремы, выражающие З.б.ч., относятся к предельным теоремам *теории вероятностей*. Первоначально был сформулирован З.б.ч. для схемы Бернулли, который получил название *Бернулли теорема*. Доказательство в этом случае строится на прямом анализе допредельных функций распределения, которые просто выражаются через биномиальные вероятности.

В основе доказательства теорем, выражающих З.б.ч., для произвольно распределённых независимых случайных величин, лежит важное неравенство, установленное в 1845 русским математиком П.Л. Чебышевым.

Лемма (Неравенство Чебышева): пусть имеется случайная величина X с математическим ожиданием m и дисперсией D. Каково бы ни было положительное число ε, вероятность того, что величина X отклонится от своего математического ожидания не меньше чем на ε, ограничена сверху числом $\frac{D}{\varepsilon^2}$:

$$P\left(|X - m| \geq \varepsilon\right) \leq \frac{D}{\varepsilon^2}.$$

Доказательство строится на использовании другого неравенства: если случайная величина X, для которой существует математическое ожидание m, может принимать только неотрицательные значения, то вероятность того, что принятое ею значение окажется не меньше единицы, не превосходит m, т.е.:

$$P\left(X \geq 1\right) \leq m).$$

Неравенство Чебышева можно записать в эквивалентной форме:

$$P\left(|X - m| < \varepsilon\right) > 1 - \frac{D}{\varepsilon^2}.$$

Рассмотрим ряд теорем, выражающих З.б.ч.

Теорема Чебышева: пусть имеется бесконечная последовательность $X_1, X_2, \ldots$ независимых случайных величин заданных на некотором вероятностном пространстве (Ω, F, P) с одним и тем же математическим ожиданием m и с дисперсиями, ограниченными одной и той же постоянной:

$$M[X_1] = M[X_2] = \ldots = m,$$

$$D[X_1] < c,\ D[X_2] < c, \ldots$$

Тогда, каково бы ни было $\varepsilon > 0$, вероятность события

$$\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - m\right| > \varepsilon$$

стремиться к 0 при $n \to \infty$.

Утверждение теоремы есть ничто иное, как сходимость по вероятности (мере):

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{P} m.$$

Теорема Чебышева обосновывает рекомендуемый в практической деятельности способ получения более точных результатов измерений: одна и та же величина измеряется многократно, и в качестве ее значения берется среднее арифметическое полученных результатов измерений.

Теорема, выражающая З.б.ч. может быть сформулирована в терминах сходимости почти наверное (с вероятностью единица). В этой формулировке закон получил название – З.б.ч. усиленный или З.б.ч. в форме А.Н. Колмогорова.

Теорема (З.б.ч. усиленный): пусть имеется бесконечная последовательность $X_1, X_2, \ldots$ независимых случайных величин заданных на некотором вероятностном пространстве (Ω, F, P) с одним и тем же математическим ожиданием m и с дисперсиями, ограниченными одной и той же постоянной:

$$M[X_1] = M[X_2] = \ldots = m,$$
$$D[X_1] < c,\ D[X_2] < c, \ldots$$

Тогда вероятность события, что $\frac{X_1 + X_2 + \ldots + X_n}{n}$ не стремится к m при $n \to \infty$, равна нулю.

Кратко можно записать

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to m \ (P\text{-п.н.})$$

или $\frac{X_1 + X_2 + \ldots + X_n}{n} \xrightarrow{\text{п.н.}} m$.

Следует отметить справедливость следующей теоремы: для того чтобы последовательность случайных величин сходилась почти наверное (с вероятностью единица) необходимо и достаточно, чтобы для любого $\varepsilon > 0$

$$P\left\{\sup_{k \geq n} |X_n - X| \geq 0\right\} \to 0,\ n \to \infty.$$

Эта теорема позволяет доказать утверждение, что если последовательность сходится почти наверное, то она сходится и по вероятности, т.е. из усиленного З.б.ч. следует З.б.ч. в форме Чебышева.

Если при доказательстве З.б.ч. использовать аппарат характеристических функций, оказывается, что требование ограниченности вторых моментов не обязательна, т.е. имеет место З.б.ч. в форме Хинчина.

Теорема Хинчина: пусть имеется бесконечная последовательность $X_1, X_2, \ldots$ независимых одинаково распределённых случайных величин заданных на некотором вероятностном пространстве (Ω, F, P) с математическим ожиданием

$m < \infty$, $S_n = X_1 + \ldots + X_n$.

Тогда $\frac{S_n}{n} \xrightarrow{P} m$, т.е. для всякого $\varepsilon > 0$

$$P\left\{\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - m\right| \geq \varepsilon\right\} \to 0,\ n \to \infty.$$

При доказательстве сходимости по вероятности в теореме Чебышева получается оценка, для вероятности среднего арифметического любого числа независимых и одинаково распределённых величин, отличаться от m более чем на заданное число:

$$P\left\{\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - m\right| \geq \varepsilon\right\} \leq \frac{D[X_1]}{n\varepsilon^2}.$$

Именно это неравенство имеет наибольшее применение на практике.

Еще одним выражением З.б.ч. является теорема, которую сформулировал и доказал Я. Бернулли.

Теорема Бернулли: для всех $\varepsilon > 0$

$$P\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) \to 1 \text{ при } n \to \infty,$$

где S_n – число наступлений события A в n независимых опытах, p – вероятность появления A в одном опыте.

Используя неравенство Чебышева, можно получить оценку вероятности, о которой идёт речь в теореме Бернулли:

$$P\left(\left|\frac{V_n}{n} - p\right| < \varepsilon\right) > 1 - \frac{p(1-p)}{\varepsilon^2}.$$

Отметим, что эта оценка является весьма грубой. Более точные оценки можно получить с использованием *теорем Муавра–Лапласа*.

ЗАКОН БОЛЬШИХ ЧИСЕЛ УСИЛЕННЫЙ

см. в ст. Закон больших чисел

ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

случайной величины – всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими вероятностями. Зная распределение вероятностей между возможными значениями случайной величины, можно до опыта судить о том, какие значения случайной величины будут появляться чаще и какие реже.

Способы или формы представления З.р.в. случайной величины – различны. В качестве З.р.в. дискретной случайной величины используют ряд распределения и функцию распределения.

Для непрерывной случайной величины нельзя построить ряд распределения, ибо невозможно пересчитать все возможные значения для такой случайной величины. Поэтому для непрерывной случайной величины в качестве законов распределения используют функцию распределения и *функцию плотности вероятностей* (плотность распределения вероятности).

И

ИСПЫТАНИЯ БЕРНУЛЛИ

см. в ст. Бернулли испытания

К

КВАНТИЛЬ

(от лат. quantum – сколько) – одна из числовых характеристик *распределения вероятностей*. В математической статистике при построении статистических критериев, интервальных оценок неизвестных параметров широко используются понятия К., процентных точек распределения. К. уровня p или p-К. непрерывной случайной величины X с функцией распределения $F(x)$ называется такое значение d_p этой случайной величины, для которого вероятность события $X < d_p$ равна заданной величине p: $P(X < d_p) = p$.

Из определения следует, что d_p есть решение уравнения $F(d_p) = p$, $0 < p < 1$. Если функция $F(x)$ строго монотонна, то указанное уравнение имеет единственный корень $d_p = F^{-1}(p)$, где $F^{-1}(p)$ – обратная к $F(x)$ функция. В противном случае при некоторых p уравнению $F(d_p) = p$ удовлетворяют многие значения d, заполняющие целый интервал, тогда в качестве d_p берут миним. из значений d, удовлетворяющих уравнению. Частным случаем К. уровня $1/2$ является медиана X, К. $d_{1/4}$ и $d_{3/4}$ называются квартилями, а $d_{0,1}, d_{0,2}, \ldots, d_{0,9}$ – децилями. Знание К. для некоторого множества значений p даёт представление о расположении и рассеянии значений случайной величины, и в частности о виде функции распределения. Напр., для стандартного нормального распределения $N(0;1)$ представление о графике функции распределения:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2}\, dz$$

дают децили:

$$d_{0,1} = -1{,}28; \quad d_{0,2} = -0{,}84; \quad d_{0,3} = -0{,}52; \quad d_{0,4} = -0{,}25; \quad d_{0,5} = 0; \quad d_{0,6} = 0{,}25;$$

$d_{0,7} = 0{,}52$; $d_{0,8} = 0{,}84$; $d_{0,9} = 1{,}28$ (см. рис. 1). К. этого нормального распределения равны $d_{1/4} = -0{,}67$; $d_{3/4} = 0{,}67$. Для ряда наиболее часто встречающихся в статистической практике законов распределения составлены специальные табл. квантилей.

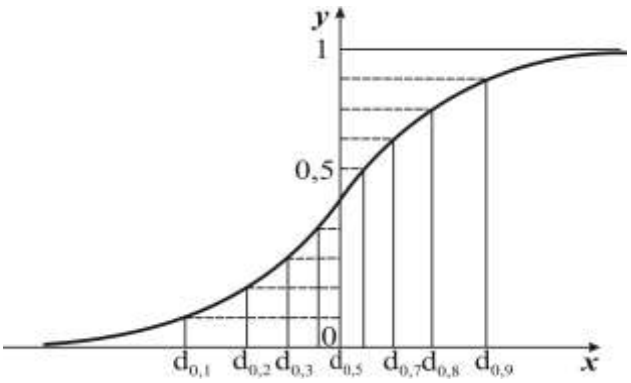


Рис. 1 **График функции распределения**

Для дискретной случайной величины функция распределения $F(x)$ меняется скачками, и следовательно, существуют такие значения уровня p, для каждго из которых не найдется возможного значения d_p, точно удовлетворяющего уравнению $F(d_p) = p$. Поэтому в дискретном случае p - К. определяется как любое число, лежащее между двумя соседними возможными значениями x_i и x_{i+1}, такое, что $F(x_i) < p$, но $F(x_{i+1}) \geq p$. Часто вместо понятия К. используют тесно связанное с ним понятие процентной точки. Процентной точкой уровня q, $0 \leq q \leq 100$, или $q\%$-й точкой непрерывной случайной величины X с функцией распределения $F(x)$ называется такое значение ν_q этой случайной величины, при котором вероятность события $X \geq \nu_q$ равна заданной величине $q/100$, т. е.

$P(X \geq \nu_q) = 1 - F(\nu_q) = q/100$. Геометрически $q\%$-ая точка – значение случайной величины, при котором пл. криволинейной трапеции, ограниченная кривой плотности распределения $f(x)$, осью абсцисс и лежащая правее $x = \nu_q$, равна $q/100$. Из определения К. и процентных точек вытекает простое соотношение, связывающее их: $d_p = \nu_{(1-p)}$. Квантильные характеристики иногда играют и самостоятельную роль. Напр., широко распространенной характеристикой степени случайного рассеяния при изучении законов распределения заработной платы и доходов являются т.н. квантильные коэффициенты дифференциации $K_d(q)$, определяемые соотношением:

$$K_d(q) = \frac{d_{1-q}}{d_q} \quad (0 < q \leq 0{,}25).$$

Наиболее распространённые среди них – децильные коэффициенты дифферен-циации, когда $q = 0{,}1$. При анализе модельных законов распределения К. и процентные точки используются для обозначения практических границ диапазона изменения соответствующего признака. Напр., К. 0,005 и 0,995 иногда определяют соответственно миним. и макс. уровни заработной платы работников в соответствующей системе показателей.

КВАНТИЛЬНАЯ ТОЧКА

см. в ст. Квантиль

КОВАРИАЦИЯ

совместный центральный момент порядков 1 и 1 или мера линейной зависимости *случайных величин*.

К. характеризует рассеивание и взаимную зависимость этих *случайных* величин, имеет размерность равную произведению размерностей случайных величин; обозначение $\operatorname{cov}(X, Y)$.

Пусть X, Y – две случайные величины, определённые на одном и том же вероятностном пространстве. Тогда их К. определяется: $\operatorname{cov}(X, Y) = M\left[(X - \mu_x)(Y - \mu_v)\right]$, в предположении, что все *математические ожидания* в правой части определены.

Свойства : К. симметрична

$\operatorname{cov}(X,Y)=\operatorname{cov}(Y,X)$; в силу линейности *математического ожидания*, К. может быть записана как

$\operatorname{cov}(X,Y)=M[XY]-M[X]\cdot M[Y]$;

К. случайной величины с собой равна *дисперсии*: $\operatorname{cov}(X,Y)=D[X]$; если X и Y независимые случайные величины, то $\operatorname{cov}(X,Y)=0$, обратное, вообще говоря, неверно.

КОВАРИАЦИОННАЯ ФУНКЦИЯ

автоковариационная функция случайной функции $X(t)$ – функция $K_x(t',t'')$, которая при каждой паре t_j и t_l допустимых значений аргумента t равна *коэффициенту ковариации* $K[X(t_j),X(t_l)]$ *случайных величин* $X(t_j),X(t_l)$ – ординат случайной функции: $K_x(t_j,t_l)=K[X(t_j),X(t_l)]$.

Выборочной К.ф. $\hat{K}_x(t',t'')$ называется зависимость между парами зафиксированных значений аргумента t случайной функции и выборочными коэффициентами ковариации соответствующих ординат. При $t'=t_i$ и $t''=t_l$ значение

$$\hat{K}_x(t',t'')=\frac{1}{n}\sum_{i=1}^{n}[x_i(t_j)-\bar{x}(t_j)]\cdot[x_i(t_l)-\bar{x}(t_l)],$$

где $x_i(t_i)$ – значение ординаты $X(t_i)$ в i-м наблюдении, n – число наблюдений. Обычно значения аргумента t задают равноотстоящими.

КОМБИНАТОРИКА

раздел математики, изучающий методы решения задач для подсчёта числа различных комбинаций; термин К. введён в математический обиход немецким философом и математиком Г. Лейбницем, который в 1666 опубликовал свой труд «Рассуждения о комбинаторном искусстве»; широко используется в *теории вероятностей* и её приложениях, а также во многих других научных областях.

В К. есть два важных правила, часто применяемых при решении комбинаторных задач – правила умножения и сложения.

Правило умножения. Пусть требуется выполнить одно за другим какие-то k действий, причём 1-ое действие можно выполнить n_1 способами, 2-ое – n_2 способами и т.д. до k-го действия, которое можно выполнить n_k способами. Тогда все k действий вместе могут быть выполнены $n_1\cdot n_2\cdot\ldots\cdot n_k$ способами.

Правило сложения. Если k действий взаимно исключают друг друга, причём одно из них можно выполнить n_1 способами, а другое - n_2 способами, и т.д. до k-го действия, которое можно выполнить n_k способами, то выполнить одно любое из этих действий можно $n_1+n_2+\ldots+n_k$ способами.

Пусть дано множество из n различных элементов и из него мы выбираем (составляем) случайным образом его m-элементные подмножества ($m\geq 0$). Эти m-элементные подмножества могут отличаться: составом элементов; порядком следования элементов в подмножестве; объёмом подмножества; возможностью повтора элементов (так называемый выбор без возвращения и с возвращением каждого выбранного элемента обратно в исходное множество). В соответствие с этим в К. выделяют следующие виды подмножеств.

А). Выбор без возвращения (входящие в состав комбинаций элементы не могут повторяться).

1. Размещения – упорядоченные m-элементные подмножества n-элементного множества, которые отличаются и составом, и порядком следования элементов. Число всех размещений из n элементов по m (где $m\leq n$), определяется по формуле:

$$A_n^m=\frac{n!}{(n-m)!},$$

где $n!=n\cdot(n-1)\cdot\ldots\cdot3\cdot2\cdot1$ – факториал числа n; n – целое неотрицательное число. $0!=1$.

2. Перестановки – любые упорядоченные множества, в которые входят по одному все n различных элементов исходного множества. Число всех перестановок из n элементов определяется по формуле: $P_n=A_n^n=n!$

3. Сочетания – m-элементные подмножества n-элементного множества, которые отличаются только составом элементов (по-

рядок их следования не важен). Число всех сочетаний из n элементов по m (где $m \le n$), определяется по формуле:

$$C_n^m = \frac{A_n^m}{m!} = \frac{n!}{m!(n-m)!}$$

Числа сочетаний называются также биномиальными коэффициентами, т.к. являются коэффициентами в разложении бинома Ньютона:

$$(a+b)^n = C_n^0 a^n + C_n^1 a^{n-1} b + ... + C_n^m a^{n-m} b^m + ... + C_n^n b^n$$

.

Также в математике известен треугольник Паскаля – арифметический треугольник, образованный биномиальными коэффициентами.

Сочетания удовлетворяет следующим соотношениям:
$C_n^m = C_n^{n-m}$; $C_n^0 = C_n^n = 1$; $C_n^1 = C_n^{n-1} = n$;
$C_n^m = C_{n-1}^{m-1} + C_{n-1}^m$; $1 \le m < n$ (правило Паскаля);

$C_n^0 + C_n^1 + C_n^2 + C_n^3 + ... + C_n^n = 2^n$ Встречается также следующее обозначение числа сочетаний из n по m:

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

Б). Выбор с возвращением (входящие в состав комбинаций элементы могут повторяться).

4. Размещения с повторениями – упорядоченные m-элементные подмножества n-элементного множества, которые отличаются и элементами, и порядком, и возможностью повтора (поэтому m может быть больше n). Число всех размещений с повторениями из n элементов по m определяется в соответствие с правилом умножения комбинаторики по формуле:

$\hat{A}_n^m = n^m$.

5. Сочетания с повторениями – m-элементные подмножества n-элементного множества, которые отличаются только элементами и возможностью повтора (m может быть больше n). Число всех сочетаний с повторениями из n элементов по m определяется по формуле:

$$\hat{C}_n^m = C_{n+m-1}^m = \frac{(n+m-1)!}{m!\ (n-1)!}$$

6. Перестановки с повторениями – упорядоченные n-элементные подмножества, в которых элемент a_1 повторяется n_1 раз, a_2 повторяется n_2 раз,…, a_k повторяется n_k раз, причем $n = n_1+n_2+ … +n_k$. Число всех перестановок с повторениями определяется по формуле:

$$\hat{P}_n(n_1, n_2, ...n_k) = \frac{n!}{n_1! n_2! ... n_k!}.$$

Осн. формулы комбинаторики

Параметры	A_n^m	C_n^m	P_n	$\hat{A}_n^m$	$\hat{C}_n^m$	$\hat{P}_n(n_1, n_2, ..., n_k)$
Элементы	+	+		+	+	
Порядок	+		+	+		+
Повторения				+	+	+
Объём	$m \le n$	$m \le n$	$m = n$			$n=n_1+n_2+…+n_k$
Формула	$\frac{n!}{(n-m)!}$	$\frac{n!}{m!(n-m)!}$	$n!$	n^m	$\frac{(n+m-1)!}{m!(n-1)!}$	$\frac{n!}{n_1! n_2! ... n_k!}$

Рассматривая конкретную задачу, необходимо выяснить, каким требованиям удовлетворяют комбинации элементов. Только после этого можно использовать нужные вычислительные формулы, комбинируя их с правилами сложения и умножения.

Пример. Пусть имеется трехэлементное множество {a,b,c,} (n=3). Рассмотрим на этом множестве все 6 типов осн. понятий К. Требуется найти:

1. Число размещений из этих 3-х элементов по 2 элемента (используется, когда нам важно,

какие элементы выбраны и в каком порядке они следуют):

{ab};{ac};{ba};{bc};{ca};{cb};

$$A_3^2 = \frac{3!}{1!} = 3 \cdot 2 \cdot 1 = 6.$$

2. Число перестановок из данных 3-х элементов (когда выбираются все исходные элементы множества и переставляются между собой):

{abc};{acb};{bac};{bca};{cab};{cba};

$$P_3 = 3! = 3 \cdot 2 \cdot 1 = 6.$$

3. Число сочетаний из данных 3-х элементов по 2 элемента (когда нам важно, какие элементы выбраны, но все равно, в каком порядке они следуют):

{ab};{ac};{bc};

$$C_3^2 = \frac{3!}{2! \cdot 1!} = \frac{3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 1} = 3.$$

4. Число размещений с повторениями из этих 3-х элементов по 2 элемента (когда нам важно, какие элементы выбраны и в каком порядке они следуют, и элементы могут повторяться):

{ab};{ac};{ba};{bc};{ca};{cb};{aa};{bb};{cc}

$$\hat{A}_3^2 = 3^2 = 9.$$

5. Число сочетаний с повторениями из данных 3-х элементов по 2 элемента (когда нам важно, какие элементы выбраны, но все равно, в каком порядке они следуют, и элементы могут повторяться):

{ab};{ac};{bc};{aa};{bb};{cc}

$$\hat{C}_3^2 = C_4^2 = \frac{4!}{2! \cdot 2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6.$$

6. Число перестановок из 6 элементов, в которых a повторяется 3 раза, b – 1 раз, c – 2 раза (когда выбираются все исходные элементы множества, повторяются заданное число раз и переставляются между собой):

{aaabcc};{aabacc};{baacac};{cacaab} и т.д.

$$\hat{P}_6(3,1,2) = \frac{6!}{3! \cdot 1! \cdot 2!} = \frac{6 \cdot 5 \cdot 4 \cdot 3!}{3! \cdot 1 \cdot 2 \cdot 1} = 60.$$

КОРРЕЛИРОВАННЫЕ ВЕЛИЧИНЫ

случайные величины, коэффициент корреляции которых не равен 0.

Если коэффициент корреляции равен нулю, случайные величины X и Y называются некоррелированными, если ρ не равен нулю – коррелированными. Из независимости случайных величин следует их некоррелированность, но из некоррелированности случайных величин ($\rho = 0$) ещё не вытекает их независимость. Если $\rho = 0$, это означает только отсутствие линейной связи между случайными величинами; любой другой вид связи может при этом присутствовать.

См. также *Корреляция*.

КОРРЕЛЯЦИЯ

зависимость между *случайными величинами*, не имеющая строго функционального характера, при которой изменение одной из случайных величин приводит в случае корреляционной зависимости к изменению только математического ожидания другой, а в общем случае стохастической зависимости – к изменению закона распределения последней. Характеристику степени корреляционной зависимости выбирают в зависимости от шкалы измерения анализируемых величин. Если все признаки количественные, то степень линейной зависимости определяют с помощью *парных, частных и множественных коэффициентов корреляции* или их квадратов – *коэффициента детерминации*. Адекватность коэффициентов корреляций тем выше, чем точнее выполняется условие линейности связей. Обычно это сводится к требованию, чтобы совместное распределение анализируемых признаков подчинялось многомерному нормальному закону.

Степень нелинейной корреляционной зависимости между двумя количественными признаками определяют с помощью корреляционного отношения. Величина распределения между корреляционными отношениями и коэффициентом детерминации свидетельствует о нелинейности связи.

Для тесноты связи между двумя признаками, измеренными в шкале порядка, используют ранговые коэффициенты корреляции Спирмена и Кендалла. В качестве измерителя тесноты связи между несколькими порядковыми переменными используют *коэффициент конкордации* (согласованности).

Тесноту связи между двумя номинальными (атрибутивными) признаками, значения которых можно классифицировать, но не ранжировать (напр., профессия работающего), используют коэффициент квадратичной сопряжённости или информационную меру связи.

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

мера линейной зависимости исследуемых количественных признаков, подчинённых совместному многомерному нормальному закону. Смешанный момент второго порядка (ковариация): 1 M{[X-MX][Y-MY]}= K_{xy} – характеризует связь X и Y, а также разброс этой двумерной величины вокруг своего центра. Недостатком является существенное влияние масштаба на значение этого момента. Нормирование устраняет этот недостаток: $\rho_{xy} = \frac{K_{xy}}{\sigma_x \cdot \sigma_y}$. Построенный т.о. К.к. *парный* является параметром ген. совокупности, характеризующим взаимосвязь признаков. Согласно неравенству Коши-Буняковского -1 $\le \rho_{xy} \le$ +1. $\rho_{xy} = \rho_{yx}$, т.е. насколько X влияет на Y, настолько же и Y влияет на X . Знак коэффициента указывает направление связи, а модуль коэффициента характеризует тесноту связи. Ген. коэффициент исследователю недоступен, поэтому о его свойствах судят по выборочному коэффициенту:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2] \cdot [\sum_{i=1}^{n}(y_i - \bar{y})^2]}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y}.$$

Выборочный коэффициент обладает всеми свойствами ген., поэтому является его хорошей оценкой и позволяет получить адекватное представление о ген. коэффициенте.

Если рассмотреть два вектора в многомерном пространстве: $\vec{X} = \{(x_1 - \bar{x}),...(x_n - \bar{x})\}^T$ и $\vec{Y} = \{(y_1 - \bar{y}),...,(y_n - \bar{y})\}^T$, то в терминах этих векторов:

$r_{xy} = \frac{\langle \vec{X} \cdot \vec{Y} \rangle}{|\vec{X}| \cdot |\vec{Y}|} = \cos\varphi$, т.е. парный

К.к. можно трактовать, как косинус угла между многомерными векторами, что хорошо объясняет свойства этого коэффициента.

На основе выборочного парного коэффициента проверяется значимость соответствующего ген. коэффициента, т.е. гипотезу: $H_0 : \rho_{xy} = 0$, при отклонении которой делается вывод о наличии связи и строится доверительный интервал. Проверка значимости осуществляется любым из трёх *критериев*: Фишера – Иейтса, *Стьюдента* или Фишера – Снедекора, которые всегда приводят к одинаковому результату. Для построения доверительного интервала сначала используют Z-преобразование Фишера, тогда Z(r) распределён нормально:

N(0, $1/\sqrt{n-3}$),

что позволяет построить интервал и с помощью обратного Z-преобразования получить границы искомого доверительного интервала.

Частный К.к. характеризует тесноту и направление связи между двумя признаками при исключении влияния на эту связь третьего признака (или группы таких признаков). Он обладает всеми свойствами парного коэффициента и для него решаются те же задачи, но число степеней свободы уменьшается на число зафиксированных признаков.

Множественный К.к., в отличие от парного и частного, характеризует только тесноту связи одного результативного признака с двумя или несколькими объясняющими. Для него проверяется значимость по критерию Фишера – Снедекора. Множественный коэффициент можно трактовать как обобщение парного. Ищется модуль парного коэффициента результативного признака с линейной комбинацией объясняющих, причём весовые коэффициенты этой комбинации определяются так, чтобы обеспечить наиболее тесную связь с результативным.

КОЭФФИЦИЕНТ СВЯЗИ КАЧЕСТВЕННЫХ ПРИЗНАКОВ

числовая величина, характеризующая статистическую связь двух признаков, т.е. возможность прогноза значений одного признака по значениям другого.

Для признаков, измеренных в порядковых шкалах, наиболее известны *коэффициенты ранговой корреляции* Спирмена и Кендалла, а также расстояние Кемени. Коэффициент Спирмена – обычный *коэффициент корреляции* между векторами рангов, соответствующих рассматриваемым показателям. Если признак строго ранжирует данное множество объектов, то ранги – номера объектов в порядке возрастания градаций. Если же имеются т.н. связанные ранги, т.е. эквивалентные объекты, им приписывается один и тот же ранг, равный среднему арифметическому значению номеров этих объектов в ранжированном вариационном ряду. Если связанных рангов нет, коэффициент Спирмена вычисляется по формуле:

$$\rho = 1 - \frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)},$$

где $\sum d^2 = (Y - X)^2$ – сумма квадратов разности рангов Y и X; n – число ранжированных единиц.

Коэффициент Спирмена изменяется от +1 (полная корреляция рангов, в этом случае $\sum d^2 = 0$) до -1 (полная обратная корреляция рангов, в этом случае

$$\frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)} = 2).$$

При $\rho = 0$, когда

$$\frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)} = 1,$$

корреляция рангов отсутствует. Значимость коэффициента корреляции рангов Спирмена проверяется на основе t-критерия Стьюдента по формуле:

$$t_p = \rho \cdot \sqrt{\frac{n-2}{1-\rho^2}}.$$

Значение коэффициента корреляции считается статистически существенным,
если $t_p > t_{табличное}$ (α, k = n - 2).

При расчёте коэффициента Кендалла и расстояния Кемени исходят из квадратных матриц величин $a_{i,j}, i, j = 1,\ldots,N$, где N – число объектов, сопоставляемых с каждым признаком по правилу: $a_{i,j} = \begin{cases} 1, \text{i предшествует j;} \\ 0, \text{i эквивалентно j;} \\ -1, \text{i следует за j.} \end{cases}$

Коэффициент Кендалла $\tau(a,b)$ рассчитывается как коэффициент корреляции соответствующих матриц (a_{ij}) и (b_{ij}):

$$\tau = \sum (a_{ij} b_{ij}) \Big/ \sqrt{\sum a_{ij}^2 \sum b_{ij}^2},$$

а расстояние Кемени $d(a,b)$ - как:

$$d = \frac{N(N-1)}{2}(1 - \tau).$$

При случайном и независимом появлении объектов в условиях гипотезы независимости признаков распределения r и τ асимптотически (по N) нормальны с нулевым средним.

Для признаков, измеренных в номинальной шкале, наиболее известны коэффициент Пирсона Φ^2 и Чупрова T^2, где:

$$\Phi^2 = \sum_{s,t} \frac{(\hat{p}_{st} - \hat{p}_{s.}\hat{p}_{.t})^2}{\hat{p}_{s.}\hat{p}_{.t}},$$

$$T^2 = \Phi^2 \Big/ \sqrt{(m_1 - 1)(m_2 - 1)},$$

$\hat{p}_{st}$ – доля объектов, имеющих s-е значение одного и t-е значение другого признака:

$$\hat{p}_{s.} = \sum_t \hat{p}_{st}; \ \hat{p}_{.t} = \sum_s \hat{p}_{st} \ (s=1,\ldots,m_1, \ t_1,\ldots,m_2).$$

Эти коэффициенты неотрицательны и принимают нулевое значение тогда и только тогда, когда признаки статистически независимы. В условиях случайного и независимого порождения объектов при гипотезе статистической независимости признаков величина $N\Phi^2$ имеет асимптотически распределение

χ^2 с $(m_1 - 1)(m_2 - 1)$

степенями свободы. Справедлива формула

$$\Phi^2 = \sum_{s,t} \frac{\hat{p}^2{}_{st}{}^2}{\hat{p}_{s.}\hat{p}_{.t}} - 1,$$

показывающая, что Φ^2 – среднее значение величин $\frac{p_{st}/p_{s.} - p_{.t}}{\hat{p}}$, характеризующих относительное улучшение прогноза значений t, когда становится известным s.

Величины τ и d могут быть рассчитаны на основе *табл. сопряжённости*, тогда как коэффициенты Φ^2 и T^2 имеют смысл коэффициентов ковариации и корреляции соответствующих матриц (a_{ij}) (см. табл. 1).

Таблица 1

Четырёхклеточная табл. сопряжённости

Группы	Подгруппы		Всего
	1	2	
А	*a*	*b*	*a* + *b*
Б	*c*	*d*	*c* + *d*
Итого	*a* + *c*	*b* + *d*	

Для данной табл. можно рассчитать коэффициент ассоциации:

$$K_a = \frac{a \cdot b - b \cdot c}{a \cdot d + b \cdot c},$$

где a, b , c , d – частоты «табл. четырёх полей». Изменяется от -1 до +1. Чем ближе этот показатель к 1 или -1, тем сильнее связаны между собой изучаемые признаки. Если коэффициент ассоциации не ниже 0.3, можно говорить о наличии существенной связи между признаками.

Для четырёхклеточной табл. сопряжённости

$\Phi^2 = T^2 = \rho^2 = \tau^2$,

где

$$K_k = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (a+c) \cdot (d+b) \cdot (d+c)}} -$$

обычный коэффициент корреляции соответствующих дихотомических признаков, заданных как альтернативные признаки. Этот коэффициент также называют коэффициентом контингенции и применяют в том случае, когда хотя бы одно значение из четырёх показателей в «табл. четырёх полей» отсутствует. По абсолютной величине коэффициент контингенции всегда меньше коэффициента ассоциации. Изменяется от -1 до +1. Чем ближе к 1 или -1, тем сильнее связаны изучаемые признаки.

Кроме указанных выше коэффициентов можно использовать т.н. биссериальный коэффициент корреляции. Коэффициент позволяет изучить связь между качественным альтернативным и количественным варьирующим признаками и определяется по формуле:

$$r = \frac{\left|\bar{Y}_2 - \bar{Y}_1\right|}{\sigma_\gamma} \cdot \frac{pq}{Z},$$

где $\bar{Y}_2, \bar{Y}_1$ – средние значения признака в группах; σ_γ – стандартное отклонение фактических значений признака от среднего уровня; p – доля первой группы в совокупности; q – доля второй группы; Z – табличные значения Z-распределения в зависимости от p.

КОЭФФИЦИЕНТ ЭКСЦЕССА

мера остроты пика распределения *случайной величины.*

Поведение *плотности (полигона) распределения* в районе его модального значения обуславливает геометрическую форму соответствующей кривой в окрестности точки её максимума, её *островершинность*. Это свойство описывается с помощью величины, называемой К.э., и обозначается ε_x. К.э. рассчитывается по зависимости:

$$\varepsilon_x = \frac{m_4^{(0)}}{\left(m_2^{(0)}\right)^2} - 3.$$

Число 3 вычитается из отношения

$$\frac{m_4^{(0)}}{\left(m_2^{(0)}\right)^2}$$

потому, что для наиболее распространённого *нормального закона распределения*

$$\frac{m_4^{(0)}}{\left(m_2^{(0)}\right)^2} = 3.$$

Кривая нормального распределения, для которого К.э. равен нулю, принята за эталон, с которым сравниваются другие распределения. Более островершинные кривые имеют положительный К.э., более плосковершинные – отрицательный К.э. (см. рис. 1).

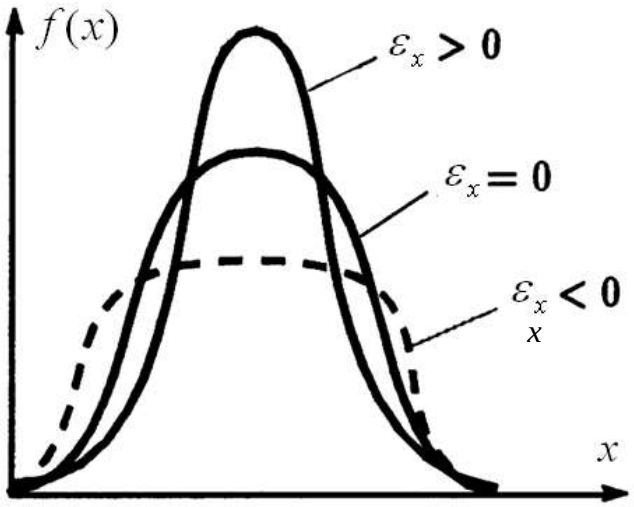


Рис. 1. **Характеристика остроты пика распределения**

К.э. оказывается полезной характеристикой при решении ряда задач, напр., при определении общего вида исследуемого распределения, или при его аппроксимации с помощью некоторых специальных разложений.

Л

ЛОГАРИФМИЧЕСКИ-НОРМАЛЬНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ

распределение непрерывной *случайной величины X*, при котором её логарифм $Y = \ln X$ распределён по нормальному закону с функцией плотности:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-m_y)^2}{2\sigma_y^2}},$$

где $y = \ln x$; $m_y = M[Y]$; $\sigma_y = \sqrt{D[Y]}$.

Случайная величина X имеет функцию плотности:

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma_y} e^{-\frac{(\ln x - m_y)^2}{2\sigma_y^2}},$$

где $m_x = e^{\frac{1}{2}\sigma_y^2 + m_y}$; $\sigma_x^2 = e^{\sigma_y^2 + 2m_y}(e^{\sigma_y^2} - 1)$.

График функции плотности имеет вид (см. рис. 1):

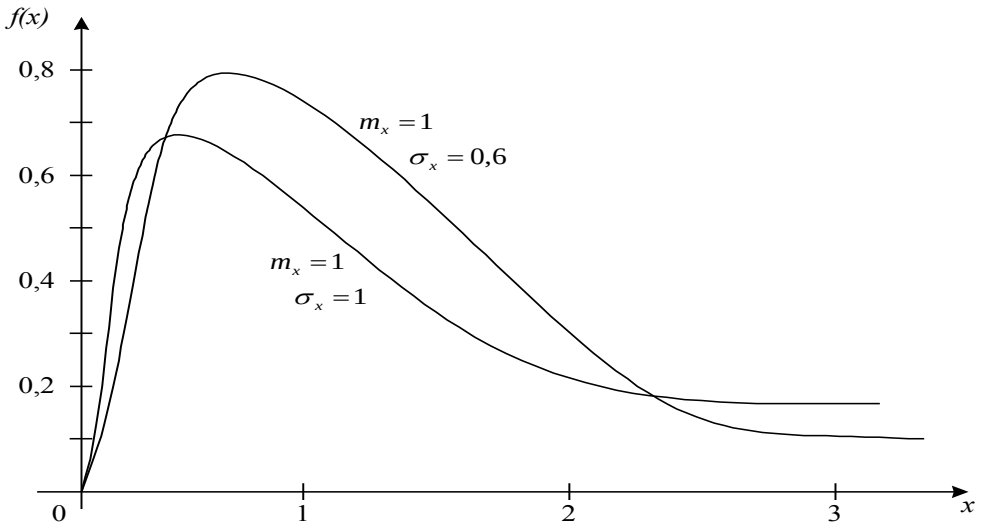


Рис. 1. **График функции плотности логнормального распределения**

Логарифмически нормальное распределение используется при моделировании таких случайных величин, как доход семьи, возраст новобрачных, допустимое отклонение от стандарта вредных веществ, зарплата работника, долговечность изделий и др. Логнормальная величина получается в результате многократных умножений независимых величин, также как нормальная случайная величина есть результат многократного суммирования.

Функция распределения логнормального распределения:

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma_y}\int_0^x \frac{1}{x} e^{-\frac{(\ln x - m_y)^2}{2\sigma_y^2}} dx$$

Числовые характеристики: Среднее: $M[x] = e^{\frac{1}{2}\sigma_y^2 + m_y}$; Мода: $M_0 = e^{m_y - \sigma_y^2}$;

Дисперсия: $D[x] = e^{\sigma_y^2 + 2m_y}(e^{\sigma_y^2} - 1)$; Асимметрия: $A_x = e^{-\frac{(m_y + \sigma_y^2)}{2}}(e^{\sigma_y^2} + 2)\sqrt{e^{\sigma_y^2} - 1}$;

Эксцесс: $\varepsilon_x = e^{4\sigma_y^2} + 2e^{3\sigma_y^2} + 3e^{2\sigma_y^2} - 6.$

Статистические оценки параметров распределения определяются по формулам:

$$\tilde{m}_y = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i \; ; \quad \tilde{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\ln x_i - \frac{1}{n}\sum_{i=1}^{n}\ln x_i\right)^2 .$$

М

МАРКОВСКАЯ ЦЕПЬ

обобщение схемы независимых испытаний (схемы Бернулли) – последовательность испытаний, в которой условная вероятность в $s+1$–м испытании ($s = 1, 2, \ldots$) осуществиться событию A_j^{s+1} (j =1,…,k), зависит только от того, каким было событие, произошедшее в s–м испытании, и не изменяется от добавочных сведений о том, какие события происходили в более ранних испытаниях.

Будем считать что *условная вероятность* появления события A_j^{s+1} в $s+1$–м испытании при условии, что в s-м испытании осуществилось событие A_i^s, не зависит от номера испытания. В этом случае М.ц. называется однородной. Вероятность перехода из i в j состояние назовём вероятностью перехода и обозначим p_{ij}.

Полная вероятностная картина возможных изменений системы, осуществляющихся при переходе от одного испытания к непосредственно следующему, задается матрицей:

$$\pi = \begin{pmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & \ddots & \vdots \\ p_{k1} & \cdots & p_{kk} \end{pmatrix},$$

составленной из вероятностей перехода, которая называется матрицей перехода.

Элементы матрицы перехода должны быть заключены в интервале (0,1), при этом, исходя из того, что при переходе из состояния A_i^s перед $(s+1)$-м испытанием система обязательно переходит в одно и только в одно из состояний $A_1,\ldots,A_k$ после s-го состояния, вытекает равенство

$$\sum_{j=1}^{k} p_{ij} = 1,\ (i = 1,\ldots,k).$$

Матрица, удовлетворяющая этому свойству, называется стохастической.

Вероятность перехода из состояния A_i^s в s-м испытании в состояние A_j^{s+n} через n испытаний вычисляется по формуле

$$P_{ij}(n) = \sum_{r=1}^{k} P_{ir}(m) P_{rj}(n-m).$$

Равенство получило название равенство Маркова.

Обозначим через π_n матрицу перехода через n испытаний. Между матрицами π_s с различными индексами существует соотношение $\pi_n = \pi_{n-m} \cdot \pi_m$ $(0 < m < n)$. В частности, при $n = 2$ находим: $\pi_2 = \pi_1 \cdot \pi_1 = \pi^2$ и вообще при любом n $\pi_n = \pi^n$. Предлагаемая здесь классификация состояний была описана А.Н. Колмогоровым для М.ц. со счётным множеством состояний и В. Деблином для М.ц. с конечным множеством состояний.

Состояние A_i называется несущественным, если существуют такое A_j и такое n, что $P_{ij}(n) > 0$, но $P_{ji}(m) = 0$ для всех m. Т.о., несущественное состояние обладает тем свойством, что из него можно попасть с положительной вероятностью в некоторое другое состояние, но из этого другого состояния вновь попасть в первоначальное (несущественное) состояние уже нельзя.

Все состояния, отличные от несущественных, называются существенными. Если состояния A_i и A_i существенны и существует такое положительное n, что $P_{ij}(n) > 0$, то существует положительное m, для которого $P_{ji}(m) > 0$. Если состояния A_i и A_j таковы, что для них при некоторых m и n выполнены оба только что указанных неравенства, то они называются сообщающимися. Очевидно, что если A_i сообщается с A_j, а A_j сообщается с A_k, то A_i сообщается также с A_k. Т.о., все существенные состояния разбиваются на классы так, что все состояния, принадлежащие одному классу, сообщаются, а принадлежащие различным классам – не сообщаются между собой. Т.к. для существенного состояния A_i и несущественного A_j при любом m имеет место равенство $P_{ij}(m) = 0$, то мы можем сделать вывод о том, что если система попала в одно из состояний определенного класса существенных состояний, то она уже не может выйти за пределы этого класса.

Рассмотрим какое-нибудь существенное состояние A_i и обозначим через M_i множество всех целых чисел m, для которых $P_{ii}(m) > 0$. Это множество не может быть пустым в силу определения существенного состояния. Очевидно, что если числа m и n входят в множество M_i, то их сумма $m+n$ также входит в это множество. Обозначим через d_i общий наибольший делитель всех чисел множества M_i. Число d_i называется периодом состояния A_i. Все состояния одного и того же класса имеют один и тот же период. Т.о., можно говорить о периоде класса существенных состояний. Если период класса $d > 1$, то класс называется периодическим. Сформулированный результат позволяет нам сделать заключение: для двух состояний A_i и A_j, принадлежащих одному классу, неравенства $P_{ij}(m) > 0$ и $P_{ji}(n) > 0$ могут выполняться лишь в том случае, когда m и $-n$ сравнимы по модулю d. Если мы выберем определённое состояние A_α изучаемого класса, то для каждого состояния A_i этого класса мы сможем поставить в соответствие определённое число β_i ($\beta_i = 1,\ldots,d$) такое, что неравенство $P_{\alpha i}(n) > 0$ возможно лишь для значений n, которые удовлетворяют равенству $n \equiv \beta_i (\mathrm{mod}\, d)$. Все состояния A_i, которым поставлено в соответствие число β, мы объединим в подкласс S_β. Т.о., класс существенных состояний оказывается разбит на d подклассов. Эти подклассы обладают

тем свойством, что при каждом шаге система из состояния, принадлежащего подклассу S_β, может перейти только в одно из состояний подкласса $S_{\beta+1}$. Если же $\beta = d$, то система переходит в одно из состояний подкласса S_1.

Если при некотором $s > 0$ все элементы матрицы перехода π_s положительны, то существуют такие постоянные числа p_j $(j = 1,\ldots,k)$ (), что независимо от индекса i имеют место равенства $\lim_{n\to\infty} P_{ij}(n) = p_j$.

Если возможные состояния системы образуют один существенный класс, то в этом случае М.ц. называется неприводимой. При этом, если неприводимая М.ц. –периодический класс, то М.ц. является периодической.

Важное обобщение рассмотренного – определение М.ц. с непрерывным временем. Семейство дискретных случайных величин $\{X_t\}_{t\geq 0}$ называется М.ц. (с непрерывным временем), если

$$P\{X_{t+h} = x_{t+h} | X_s = x_s,\ 0 \prec s \leq t\} = P\{X_{t+h} = x_{t+h} | X_t = x_t\}.$$

М.ц. с непрерывным временем называется однородной, если

$$P\{X_{t+h} = x_{t+h} \mid X_t = x_t\} = P\{X_h = x_h \mid X_0 = x_0\}.$$

Аналогично случаю дискретного времени конечномерные распределения однородной М.ц. с непрерывным временем полностью определены начальным распределением

$(p_1, p_2, \ldots)^T$, $p_i = P\{X_0 = i\}$, $i = 1, 2, \ldots$

и матрицей переходных функций (переходных вероятностей)

$$P(h) = \left(P_{ij}(h)\right) = P\{X_h = j \mid X_0 = i\}.$$

Матрица переходных вероятностей удовлетворяет уравнению Колмогорова - Чепмена:

$$P(t+s) = P(t)P(s).$$

По определению, матрица интенсивностей

$$Q = \lim_{h\to 0} \frac{P(h) - I}{h}$$

или, что эквивалентно,

$$Q = \left(q_{ij}\right) = \left(\frac{dP_{ij}(h)}{dh}\right)_{h=0}.$$

Для М.ц. с непрерывным временем строится ориентированный граф переходов по правилам: множество вершин графа совпадает со множеством состояний цепи; вершины i, j $(i \neq j)$ соединяются ориентированным ребром, если $q_{ij} > 0$ (т.е. интенсивность потока из i-го состояния в j-е положительна).

Пять свойств конечной М.ц. эквивалентны, цепи, обладающие ими, называют эргодическими: граф переходов цепи ориентированно связан; нулевое собственное число матрицы Q невырождено и ему соответствует строго положительный левый собственный вектор (равновесное распределение); для некоторого $t > 0$ матрица $P(t)$ строго положительна (то есть $P_{ij}(t) > 0$ для всех i,j); для всех $t > 0$ матрица $P(t)$ строго положительна; при $t \to \infty$ матрица $P(t)$ стремится к строго положительной матрице, у которой все строки совпадают и совпадают, очевидно, с равновесным распределением.

МАРКОВСКИЙ ПРОЦЕСС (ПРОЦЕСС МАРКОВА)

случайный процесс, в котором будущее зависит от прошлого только через настоящее.

Пусть $X(t)$ – случайный процесс, $P\{X \mid t_n\}$ – *вероятность условная* (в зависимости от условий, обозначенных в скобках). Случайный процесс $X(t)$, $t \in T$ называется М.п., если для любых n моментов времени $t_1, t_2, \ldots, t_n$ из множества T условная *функция распределения* последнего значения $X(t_n)$ при фиксированных значениях $X(t_1), \ldots, X(t_{n-1})$ зависит только от $X(t_{n-1})$. Следовательно, при известных значениях $x_1, x_2, \ldots, x_n$ справедливым будет соотношение:

$$P\{X(t_n) \le x_n \mid X(t_1) = x_1, X(t_2) = x_2, \ldots, X(t_{n-1}) = x_{n-1}\} = P\{X(t_n) \le x_n \mid X(t_{n-1}) = x_{n-1}\}.$$

М.п., определённый на множестве T, состоящем только из целых чисел, принимающий только дискретное множество значений называют *Марковской цепью*.

МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ (СРЕДНЕЕ ЗНАЧЕНИЕ)

функции y(X) от *случайной величины одномерной, дискретной или непрерывной* X определяется формулой:

$$My(X) = \begin{cases} \sum_i y(x_i)p_i & (X - \text{дискретная}), \\ \int_{-\infty}^{\infty} y(x)p(x)dx & (X - \text{непрерывная}), \end{cases}$$

если эти выражения существуют в смысле абсолютной сходимости.

При y (X)=X получаем формулу М.о. случайной величины X:

$$MX = \begin{cases} \sum_i x_i p_i & (X - \text{дискретная}), \\ \int_{-\infty}^{\infty} x \cdot p(x)dx & (X - \text{непрерывная}). \end{cases}$$

Здесь x_i – значение случайной величины X; p_i= P (X=x_i) – вероятность события X=x_i; p(x) – плотность распределения вероятностей случайной величины X.

MX является функционалом (но не функцией X), описывающим свойство случайной величины X, характеризующее ее положение (расположение на оси Ox).

М.о. линейной функции случайной величины равно этой линейной функции от М.о. случайной величины. В частности M(cX) = cMX; Mc = c. М.о. линейной комбинации случайных компонент случайной величины *многомерной* равно той же линейной комбинации их М.о. В частности M(X±Y) = MX±MY. М.о. произведения двух случайных компонент XY находится по формуле:

M(X·Y) = MX·MY+cov(X,Y), где cov(X,Y) – *коэффициент ковариации* X и Y; если X и Y некоррелированы, то M(X·Y) = MX · MY.

В статистике М.о. играет существенную роль в определении такого свойства оценки параметра, как несмещённость. В качестве примера можно показать, что М.о. закона *распределения биномиального* случайной величины X=m равно Mm=np, где p – вероятность появления события A в каждом из n испытаний, m – число появлений события A (частота события A). Эта величина может не совпадать ни с одним значением числа m появлений события A, являющимся случайной величиной. Если y(X)=(X-MX)^2, получим формулу дисперсии случайной величины X. М.о. обозначается MX, EX или μ.

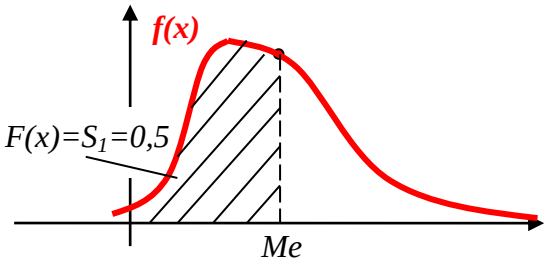
МЕДИАНА

(от лат. mediana – средняя) – в *теории вероятностей* одна из числовых характеристик случайной величины; термин введён Гальтоном в 1882.

Для *случайной величины непрерывной*, имеющей строго монотонную функцию распределения $F(x)$, М. определяется как единственный корень уравнения:

$$F(x) = \frac{1}{2},$$

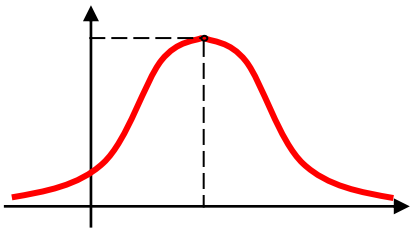
т.е. М. – такое число Me, что X принимает с вероятностью 0,5 как значения больше Me, так и значения меньше Me.



Геометрически вертикальная прямая x=Me делит пл. под кривой плотности вероятности f(x) пополам. В случае, если функция плотности вероятностей абсолютно симметрична (как, напр., у нормального распределения N(μ,σ)) медиана совпадает с *модой* и *математическим ожиданием*.

М. случайной величины *дискретной* называется любое из чисел таких, что:

$$P(X \le Me) \ge \frac{1}{2}; \quad P(X \ge Me) \ge \frac{1}{2}.$$



М. дискретной случайной величины всегда существует, но не всегда единственна. Она может определяться как любая точка некоторого интервала $(x_i; x_j]$, для которого $F(x)=\frac{1}{2}$, и тогда за М. часто принимают середину этого интервала. Если у функции распределения дискретной случайной величины нет значения

$$F(x)=\frac{1}{2},$$

то М. $Me=x_j$ разделяет два интервала значений случайной величины, $(x_i; x_j]$ и $(x_j; x_k]$, для которых, соответственно,

$$F(x)<\frac{1}{2} \text{ и } F(x)>\frac{1}{2}.$$

М. случайной величины X минимизирует средний модуль отклонения значений случайной величины от любого числа $a \neq Me$:

$$M\,|X-Me|=\min_a M\,|X-a|.$$

В *математической статистике* для оценки М. случайной величины по независимым результатам её наблюдений $x_1, x_2, \ldots, x_n$ используют выборочную медиану – середину ранжированного ряда наблюдений, соответствующего *вариационного ряда* $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. М. $\hat{Me}=x_{(k)}$, если имеется нечётное число наблюдений $n=2k+1$. Если объём выборки равен чётному числу $n = 2k$, по определению, М. может быть любое число, лежащее между $x_{(k)}$ и $x_{(k+1)}$; на практике чаще всего используют среднее арифметическое двух средних значений, т.е.:

$$\hat{Me}=\frac{x_{(k)}+x_{(k+1)}}{2}.$$

Выборочная М. является несмещённой ($M(\hat{Me})=Me$) и состоятельной оценкой. Если объём выборки стремится к бесконечности, выборочная М. сходится по вероятности к истинной М., т.е. для всякого $\varepsilon>0$:

$$P\left(\left|\hat{Me}-Me\right|>\varepsilon\right)\xrightarrow{n\to\infty}0.$$

Если выборочная совокупность наблюдений сгруппирована с помощью интервального вариационного ряда, то М. вычисляют по формуле:

$$\hat{Me}=a_{Me}+h\cdot\frac{\frac{n}{2}-m_{н(Me-1)}}{m_{Me}},$$

где a_{Me} – нижняя граница медианного интервала; h – шаг интервального ряда (ширина интервала); n – объём выборки; m_{Me} – частота встречаемости признака в медианном интервале; $m_{н(Me\text{-}1)}$ – накопленная частота интервала, предшествующего медианному.

Достоинство М. как меры центральной тенденции состоит в том, что на М. не влияют изменения крайних членов вариационного ряда при условии, что не меняется их положение относительно М. В этом смысле М. – наиболее устойчивая характеристика среднего значения вариационного ряда.

М. употребляется реже, чем математическое ожидание и чаще, чем мода. М. – это 50-й процентиль, вторая квартиль или *квантиль* уровня 0,5.

МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

измеритель тесноты статистической связи между результирующим показателем y и набором объясняющих переменных x(1), x(2), …, x(p) при линейной форме регрессионной зависимости f(X). М.к.к. $R_{y,x}$ вычисляется по (p + 1) × (p + 1) – матрице *парных коэффициентов корреляции*

$R = (\rho_{j,k})$, j, k = 0, 1, …, p:

$$R_{y,X}=1-\frac{|R|}{R_{00}},$$

где R_{00} – *алгебраическое дополнение* элемента $\rho_{00} = 1$ в матрице R.

Квадрат М.к.к. называется *коэффициентом детерминации*, который определяет долю общей вариации результирующего признака y, объясняемую влиянием соответствующих факторных признаков. М.к.к. изменяется в пределах от 0 до 1. Равенство $R_{y,X}$ единице свидетельствует о функциональной зависимости

между результирующим показателем y и набором объясняющих переменных x(1), x(2), …, x(p), а равенство его нулю свидетельствует об отсутствии линейной зависимости между указанными выше признаками.

Выборочное значение М.к.к. $\hat{R}_{y,x}$ рассчитывается по вышеприведённым формулам с заменой соответствующих *коэффициентов корреляции*, участвующих в правых частях этих выражений, их выборочными аналогами.

Проверка гипотезы H0: $R_{y,x} = 0$, (т.е. проверка гипотезы об отсутствии линейной связи между y, с одной стороны, и совокупностью переменных x(1), x(2), …, x(p), – с другой) осуществляется с помощью критической статистики $\gamma_n = \frac{n-p-1}{p} \cdot \frac{\hat{R}^2_{y,X}}{1-\bar{R}^2_{y,X}}$, которая в условиях справедливости проверяемой гипотезы должна «вести себя» как F(p; n – p – 1) – распределённая случайная величина. Поэтому, если окажется, что $\frac{n-p-1}{p} \cdot \frac{\hat{R}^2_{y,X}}{1-\hat{R}^2_{y,X}} > F_{\alpha;\,p;\,n-p-1}$, то проверяемая гипотеза отвергается в пользу *альтернативной гипотезы* с вероятностью ошибки, равной α. Величина F(α; p; n – p – 1) – это 100α%-ная точка F(p; n – p – 1) – распределения.

МУАВРА – ЛАПЛАСА ТЕОРЕМА

см. в ст. Теорема Муавра – Лапласа

Н

НАЧАЛЬНЫЙ МОМЕНТ СЛУЧАЙНОЙ ВЕЛИЧИНЫ (МОМЕНТ ПОРЯДКА Q ОТНОСИТЕЛЬНО НАЧАЛА ОТСЧТА)

математическое ожидание случайной величины в степени *q* для одномерного распределения $M\left[X^q\right]$.

Для *случайной величины дискретной* начальный момент выражается суммой

$$m_q = \sum_{i=1}^{n} \left(x_i^0\right)^q p_i,$$

а для непрерывной – интегралом

$$m_q = \int_{-\infty}^{\infty} x^q f(x)dx.$$

Из Н.м.с.в. особое значение имеет момент первого порядка, который представляет собой математическое ожидание случайной величины $m_1 = M[X]$.

Начальные моменты высших порядков используются гл. обр. для вычисления центральных моментов.

НАЧАЛЬНЫЕ МОМЕНТЫ ВЫБОРОЧНЫЕ

эмпирические, выборочные аналоги начальных моментов *случайной величин*, определяются по формуле

$$\hat{m}_k = \frac{\sum_{i=1}^{n} x_i^k}{n}.$$

Для группированных выборочных данных, когда *n* наблюдений разбиты на *P* групп с n_i наблюдениями в *i-й* группе (*i=1,2,..,p*) н.м.в. определяются по формуле –

$$\hat{m}_k = \frac{\sum_{i=1}^{n} n_i x_i^k}{n}.$$

Н.м.в. первого порядка представляет собой среднюю арифметическую:

$$\bar{x} = \hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

а нулевого порядка равен единице $\hat{m}_0 = 1$. При большом объёме *выборки* (при $n \to \infty$) каждый Н.м.в. распределён асимптотически нормально.

НЕВОЗМОЖНОЕ СОБЫТИЕ

событие, которое никогда не наступает в результате данного эксперимента; вероятность невозможного события равна 0.

Существует также другое определение – событие называется невозможным, если оно не содержит ни одного элементарного события, иначе – это пустое подмножество *пространства элементарных событий*. Н.с. никогда не происходит.

НЕЗАВИСИМЫЕ СОБЫТИЯ

события, вероятность появления или непоявления которых не изменяется от появления или непоявления других событий.

Для двух событий: событие A называется независимым от события B, если величина вероятности события A не изменяется при появлении или непоявлении события B.

Условие независимости записывается в виде $P(A) = P(A/B) = P(A/\overline{B})$.

Событие A называется зависимым от события B, если вероятность события A изменяется при появлении или не появлении события B.

Условие зависимости записывается в виде $P(A) \neq P(A/B) \neq P(A/\overline{B})$.

Зависимость и независимость событий являются взаимными. Это значит, что, если событие A зависит от события B, то и событие B зависит от события A.

НЕЗАВИСИМОСТЬ СОБЫТИЙ

см. в ст. Независимые события

НЕЗАВИСИМОСТЬ СЛУЧАЙНЫХ ВЕЛИЧИН

две *случайные величины* *X* и *Y* независимы, если их функции распределения представлены как

$$F(x,y) = F(x,\infty)F(\infty,y) = G(x)H(y),$$

где $F(x,\infty) = G(x)$ и $F(\infty,y) = H(y)$ – маргинальные функции распределения *X* и *Y*, соответственно, для всех пар (x,y).

Для непрерывной независимой случайной величины её плотность распределения, если она существует, выражают как

$$f(x,y) = g(x)h(y),$$

где $g(x)$ и $h(y)$ – маргинальные плотности распределения *X* и *Y*, соответственно, для всех пар (x,y).

Для дискретной независимой случайной величины её вероятности выражают как

$$P(X = x_i; Y = y_j) = P(X = x_i)P(Y = y_j)$$

для всех пар (x_i, y_j).

С понятием независимости случайных величин тесно связано понятие независимости σ-алгебр.

НЕПРЕРЫВНОЕ ВЕРОЯТНОСТНОЕ ПРОСТРАНСТВО

см. в ст. Вероятностное пространство

НЕСОВМЕСТНОЕ СОБЫТИЕ

событие, появление которого исключает появление другого события. Напр., в магазин поступила партия товара одной номенклатуры, но разного цвета. Событие A – наудачу взятая коробка с товаром чёрного цвета, событие B – коробка с товаром коричневого цвета. A и B – Н.с.

С.с. $A_1, A_2, ..., A_n$ называются несовместными, если события, входящие в группу попарно, несовместны.

Напр., производится бросок игральной кости (кубика с пронумерованными боковыми гранями от 1 до 6). A_1 – выпадение одного очка, A_2 – выпадение двух очков, A_3 – выпадение трёх очков, A_4 – выпадение четырёх очков, A_5 – выпадение пяти очков, A_6 – выпадение шести очков. $A_1, A_2, A_3, A_4, A_5, A_6$ образуют группу Н.с.

О

ОБЪЕДИНЕНИЕ (СУММА) СОБЫТИЙ

событие, состоящее в наступлении хотя бы одного из этих событий. Сумма C событий $A_1, A_2, ..., A_n$ обозначается:

$$C = A_1 + A_2 + ... + A_n.$$

Напр., бросаем один раз игральную кость. Событие A – выпадение чётного числа очков, событие B – выпадение числа очков, большего четырёх. Событие C = A + B состоит в том, что выпало либо чётное число очков, либо число очков большее четырёх, т.е. произошло либо событие A, либо событие B, либо оба события вместе.

См. также *Случайное событие.*

ОТКЛОНЕНИЕ СРЕДНЕКВАДРАТИЧЕСКОЕ

Среднеквадратическим отклонением σ_X *случайной величины* X называется положительный корень из дисперсии

$\sigma_X = \sqrt{D(X)}$, где дисперсия $D(X) = \begin{cases} \sum_{i=1}^{n} (x_i - M(X))^2 p_i, & \text{если } X \text{ дискретна}; \\ \int_{-\infty}^{+\infty} (x - M(X))^2 f(x)dx, & \text{если } X \text{ непрерывна}; \end{cases}$

математическое ожидание:

$$M(X) = \begin{cases} \sum_{i=1}^{n} (x_i p_i), & \text{если } X \text{ дискретна}; \\ \int_{-\infty}^{+\infty} xf(x)dx, & \text{если } X \text{ непрерывна}; \end{cases}$$

где p_i – вероятность того, что случайная величина X примет значение x_i; $f(x)$ – функция плотности вероятностей случайной величины X; n – число возможных значений случайной величины X.

О.с. σ_X используется, наряду с дисперсией, для характеристики степени рассеивания случайной величины и оказывается в ряде случаев более удобным и естественным, в первую очередь, из-за своей однородности (в смысле единиц измерения) с различными характеристиками центра группирования.

П

ПАРНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

измеритель тесноты линейной зависимости между двумя величинами на фоне действия всех остальных из p величин ξ1, ξ2,…, ξp, описывающих изучаемое явление или процесс; через ковариации определяется по формуле:

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj} \cdot \sigma_{kk}}},$$

где σ_{jk} – ковариация между величинами ξi и ξк; σ_{jj} $u \cdot \sigma_{kk}$ – дисперсия этих величин I,k=1,2,…,p.

В статистическом анализе используют статистические оценки, или выборочные П.к.к., рассчитываемые по данным n наблюдений по формуле:

$$r_{jk} = \frac{\sum_{i=1}^{n} \left(x_i^{(j)} - \bar{x}^{(j)}\right)\left(x_i^{(k)} - \bar{x}^{(k)}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_i^{(j)} - \bar{x}^{(j)}\right)^2 \cdot \sum_{i=1}^{n} \left(x_i^{(k)} - \bar{x}^{(k)}\right)^2}},$$

где

$x_i^{(j)}$ – значение j-го показателя для i-го наблюдения,

$\bar{x}^{(j)}$ – среднее значение j-го показателя (j=1,2,…,p; i=1,2,…,n).

П.к.к. может принимать значения в диапазоне от -1 до +1. Причём для статистически независимых переменных он равен нулю, но из равенства нулю значения П.к.к., вообще говоря, не следует взаимной независимости случайных величин, а следует лишь тот факт, что если между случайными величинами связь и существует, то она нелинейна. С другой стороны из факта $|\rho_{jk}|=1$ следует, что анализируемые переменные связаны чисто функциональным линейным соотношением.

Для проверки факта статистически значимого отличия от нуля величины выборочного П.к.к., т.е. для проверки гипотезы H0:

$\rho_{jk} = 0,$

используется статистика

$$\gamma_n = \frac{r_{jk}}{\sqrt{1-r_{jk}^2}}\sqrt{n-2},$$

которая в предположении справедливости проверяемой гипотезы должна «вести себя» приблизительно как стьюдентовская случайная величина t с n – 2 степенями свободы соответственно, если окажется, что

$$\left|\frac{r_{jk}}{\sqrt{1-r_{jk}^2}}\sqrt{n-2}\right| > t_{\frac{\alpha}{2};(n-2)},$$

то гипотеза H0 отвергается с вероятностью ошибки α; tα/2; (n - 2) – это (α/2) 100%-ая точка t-распределения.

Оценка интервальная (при уровне доверия P0 = 1 - 2α)

для истинного значения ρ jk строится по формуле: th z1 < ρ jk < th z2, где th z – тангенс гиперболический от аргумента z, соответственно,

$z = arctg\, r_{jk} = \frac{1}{2}\ln\frac{1+r_{jk}}{1-r_{jk}}$ – т.н. z-преобразование Фишера. Так что

$$z_{1,2} = \frac{1}{2}\ln\frac{1+r_{jk}}{1-r_{jk}} \mp \frac{\omega_\alpha}{\sqrt{n-3}} - \frac{r_{jk}}{2(n-1)},$$

где ωα – 100α%-ая точка стандартного нормального распределения.

ПЛОТНОСТЬ ВЕРОЯТНОСТИ

см в ст. Функция плотности вероятности

ПЛОТНОСТЬ ПОТОКА СОБЫТИЙ

см. в ст. Поток событий

ПЛОТНОСТЬ РАСПРЕДЕЛЕНИЯ

см. в ст. Функция плотности вероятности

ПОЛНАЯ СИСТЕМА СОБЫТИЙ (ПОЛНАЯ ГРУППА СОБЫТИЙ)

такая система *случайных событий*, что в результате произведённого случайного эксперимента непременно произойдёт одно из них. Пусть (Ω, F, P) *вероятностное пространство*, тогда любое разбиение множества Ω будет П.с.с.

При рассмотрении формулы Байеса события полной группы называют гипотезами, а вероятности наступления гипотез – *априорными вероятностями*.

ПОТОК СОБЫТИЙ

последовательность событий, наступающих одно за другим в случайные моменты времени. Если рассматривать события, заданные на некотором вероятностном пространстве (Ω, F, P), и ввести случайные величины $X(\omega)$, то последовательность событий можно рассматривать как последовательность *случайных величин*, т.е. *случайный процесс*.

Определим некоторые свойства П.с.: поток называется потоком без последствия, если вероятность наступления некоторого числа событий в течение промежутка времени $(T, T+t)$ не зависит от того, какое число событий появлялось ранее. Это означает, что условная вероятность появления некоторого числа событий за промежуток $(T, T+t)$ при любом предположении о количестве наступлений событий до момента T совпадает с безусловной вероятностью; поток называется ординарным, если вероятность появления более одного раза за малый промежуток времени Δt некоторого элементарного события бесконечно мала по сравнению с вероятностью появления события один раз. Плотность П.с. – это среднее число событий в единицу времени. Ординарный П.с. без последействия называется пуассоновским потоком. Если события образуют пуассоновский поток, то число событий k, попадающих на любой участок времени (t_0, t_0+t) распределено по закону Пуассона:

$$P_k = \frac{a^k}{k!}e^{-a},\ k \geq 0,$$

где a – математическое ожидание числа точек, попадающих на участок: $a = \int_{t_0}^{t_0+t} \lambda(t)dt$, $\lambda(t)$ – функция плотности потока. Если $\lambda(t) = \text{const}$, то поток называется простейшим (стационарным пуассоновским) потоком.

Промежуток времени, протекший между появлениями двух последующих появлений интересующего нас события – случайная величина, которую мы обозначим через τ. Для простейшего потока справедливо распределение:

$$P(\tau > t) = P_0(t) = e^{-\lambda t}.$$

Функция распределения случайной величины τ есть $F(t) = 1 - e^{-\lambda t}$.

Данный закон распределения – показательный закон.

ПОТОК СОБЫТИЙ БЕЗ ПОСЛЕДЕЙСТВИЯ

см. в ст. Поток событий

ПОТОК СОБЫТИЙ ОРДИНАРНЫЙ

см. в ст. Поток событий

ПОТОК СОБЫТИЙ ПРОСТЕЙШИЙ

см. в ст. Поток событий

ПРАВИЛО ТРЁХ СИГМ

вероятность того, что *случайная величина* отклонится от своего *математического ожидания* на величину, большую, чем утроенное среднее квадратичное отклонение, практически равна нулю (см. рисунок).

При рассмотрении нормального закона распределения выделяется важный частный случай, известный как П.т.с.

Вероятность того, что отклонение нормально распределённой случайной величины от математического ожидания меньше заданной величины Δ можно рассчитать по зависимости:

$$P(|X - m| < \Delta) = \overline{\Phi}\left[\frac{m + \Delta - m}{\sigma}\right] - \overline{\Phi}\left[\frac{m - \Delta - m}{\sigma}\right] = \overline{\Phi}\left[\frac{\Delta}{\sigma}\right] - \overline{\Phi}\left[-\frac{\Delta}{\sigma}\right] = 2\overline{\Phi}\left[\frac{\Delta}{\sigma}\right]$$

Если принять Δ = 3σ, то получается с использованием табл. значений функции Лапласа:

$$P(|X - m| < 3\sigma) = 2\overline{\Phi}(3) = 2 \cdot 0{,}49865 = 0{,}9973$$

Т.е. вероятность того, что случайная величина отклонится от своего математического ожидания на величину, большую, чем утроенное среднее квадратичное отклонение, практически равна нулю; это и есть П.т.с.

На практике считается, что если для какой-либо случайной величины выполняется П.т.с., то эта случайная величина имеет нормальное распределение.

Правило справедливо только для случайных величин, распределённых по нормальному закону.

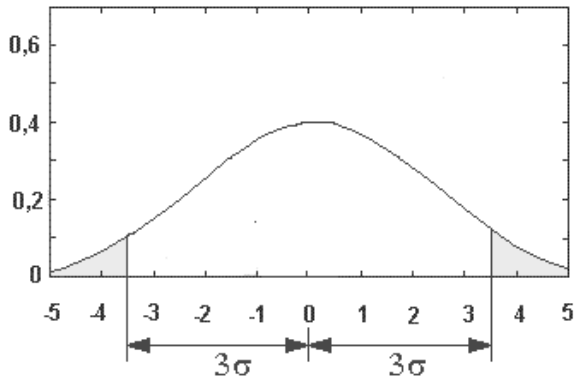


Рис. **Правило трёх сигм**

ПРИНЦИП ПРАКТИЧЕСКОЙ НЕВОЗМОЖНОСТИ МАЛОВЕРОЯТНЫХ СОБЫТИЙ

практически можно считать, что *случайное событие*, которое имеет очень малую *вероятность*, в единичном испытании не наступит.

При решении многих практических задач приходится иметь дело с событиями, вероятность которых весьма мала, т.е. близка к нулю. Можно ли считать, что маловероятное событие А в единичном испытании не произойдет? Такого заключения сделать нельзя, т.к. не исключено, хотя и мало вероятно, что событие А наступит.

Казалось бы, появление или непоявление маловероятного события в единичном испытании предсказать невозможно. Однако длительный опыт показывает, что маловероятное событие в единичном испытании в подавляющем большинстве случаев не наступает.

Возникает вопрос: насколько малой должна быть вероятность события, чтобы можно было считать невозможным его появление в одном испытании? На этот вопрос нельзя ответить однозначно. Для задач, различных по существу, ответы разные. Напр., если вероятность того, что парашют при прыжке не раскроется, равна 0,01, то было бы недопустимым применять такие парашюты. Если же вероятность того, что поезд дальнего следования прибудет с опозданием, равна 0,01, то можно практически быть уверенным, что поезд прибудет вовремя.

Достаточно малую вероятность, при которой (в данной определённой задаче) событие можно считать практически невозможным, называют уровнем значимости. На практике обычно принимают уровни значимости, заключенные между 0,01 и 0,05. Уровень значимости, равный 0,01, называют однопроцентным; уровень значимости, равный 0,02, называют двухпроцентным, и т. д.

Из П.п.н.м.с. вытекает следствие: если случайное событие имеет вероятность, близкую к единице, то практически можно считать, что в единичном испытании это событие наступит. Ответ на вопрос о том, какую вероятность считать близкой к единице, зависит от существа рассматриваемой задачи.

ПРОСТРАНСТВО ЭЛЕМЕНТАРНЫХ СОБЫТИЙ

множество всех взаимно или попарно исключающих друг друга исходов случайного эксперимента (*элементарных событий*), которые вместе образуют полную систему событий.

Понятие П.э.с. относится к основополагающим понятиям в *теории вероятностей*.

П.э.с. обозначается заглавной буквой греческого алфавита Ω. Входящие в него элементарные события обозначаются строчными буквами греческого алфавита, при необходимости – с индексами: ω, ω_1, ω_2, ω_3,... П.э.с. описывается, как и любое множество – фигурными скобками, в которых перечисляются входящие в него элементарные события, напр., $\Omega = \{\omega_1, \omega_2, \omega_3, ...\}$. Для наглядности Ω изображают в виде некоторой области на плоскости, а элементарные исходы ω_i – точками в этой области.

П.э.с. называется дискретным, если число его элементов конечно или счётно. Конечное пространство содержит конечное число элементарных событий, счётное – бесконечное число, однако такое, которое можно перенумеровать (говорят также – пересчитать). Любое пространство элементарных событий не являющееся дискретным, называется недискретным, и при этом, если наблюдаемыми результатами (нельзя произносить случайными событиями) являются точки того или иного числового арифметического или координатного пространства, то пространство называется непрерывным. П.э.с. Ω вместе с алгеброй событий F и вероятностью P образует т.н. *вероятностное пространство*.

ПРОТИВОПОЛОЖНОЕ (ДОПОЛНИТЕЛЬНОЕ) СОБЫТИЕ

см. в ст. Случайное событие

ПРОЦЕНТНАЯ ТОЧКА РАСПРЕДЕЛЕНИЯ

возможное значение W_q случайной величины ξ, для которого вероятность события $\xi > W_q$ равна q, т.е. $1\text{-}F(W_q) = P(\xi \geq W_q) = q$. Из определения *квантилей* U_q и процентных точек W_q следует соотношение между ними $U_q = W_{(1-q)}$.

Процентные точки, как и квантили, играют существенную роль в определении критических областей при *проверке гипотез*, а также в определении *интервальных оценок*.

Р

РАЗЛОЖЕНИЕ ЭДЖВОРТА

ряд, введённый Эджвортом, дающий асимптотическое приближение *функции плотности распределения* или для *функции распределения вероятностей* через соответствующие функции нормального распределения.

Для функции плотности распределения Р.Э. имеет вид:

$$f(x)=\varphi(x)- \\ -\frac{1}{3!}\frac{\mu_3}{\sigma^3}\varphi^{(3)}(x)+ \\ +\frac{1}{4!}\left(\frac{\mu_4}{\sigma^4}-3\right)\varphi^{(4)}(x)+\frac{10}{6!}\left(\frac{\mu_3}{\sigma^3}\right)^2\varphi^{(6)}(x)- \\ -\frac{1}{5!}\left(\frac{\mu_5}{\varsigma^5}-10\frac{\mu_3}{\sigma^3}\right)\varphi^{(5)}(x)-\frac{35}{7!}\frac{\mu_3}{\sigma^3}\left(\frac{\mu_4}{\sigma^4}-3\right)\varphi^{(7)}(x)-\frac{280}{9!}\left(\frac{\mu_3}{\sigma^3}\right)^3\varphi^{(9)}(x)+ \\ +\ldots \quad ,$$

где в каждой строке выписаны слагаемые одного и того же порядка, μ_ν – *центральные моменты* случайной величины, $\varphi(x)$ – функция плотности распределения нормированной нормальной величины, $\varphi^{(\nu)}(x)=(-1)^\nu H_\nu(x)\varphi(x)$, где $H_\nu(x)$ есть полиномы Эрмита степени ν. Т.о., представленное разложение есть разложение по ортогональным функциям. Для того, чтобы получить соответствующее разложение для функции распределения $F(x)$, нужно только заменить $\varphi(x)$ на $\Phi(x)$.

Вводя в рассмотрение коэффициенты ассиметрии и эксцесса γ_1 и γ_2, можно записать выражение для $f(x)$ до членов порядка n^{-1} в виде:

$$f(x)=\varphi(x)-\frac{\gamma_1}{3!}\varphi^{(3)}(x)+\frac{\gamma_2}{4!}\varphi^{(4)}(x)+\frac{10\gamma_1^2}{6!}\varphi^{(6)}(x$$

.

При больших x записанное выражение будет давать отрицательные значения, что противоречит тому, что $f(x)\geq 0$, но это согласуется с тем, что представленное разложение даёт не точное равенство, а приближённое.

РАСПРЕДЕЛЕНИЕ БИНОМИАЛЬНОЕ

распределение вероятностей появления m числа событий в n независимых испытаниях, в каждом из которых вероятность появления события постоянна и равна p. Вероятность возможного числа появления события вычисляется по формуле Бернулли:

$$P_{n,m}=P(X=m)=C_n^m p^m(1-p)^{n-m}=\frac{n!}{m!(n-m)!}p^m(1-p)^{n-m}$$

где m – возможное значение случайной величины *X*, указывающее, какое число раз при n испытаниях может наступить интересующее нас событие.

Полигон вероятностей Р.б. имеет вид (см. рис. 1):

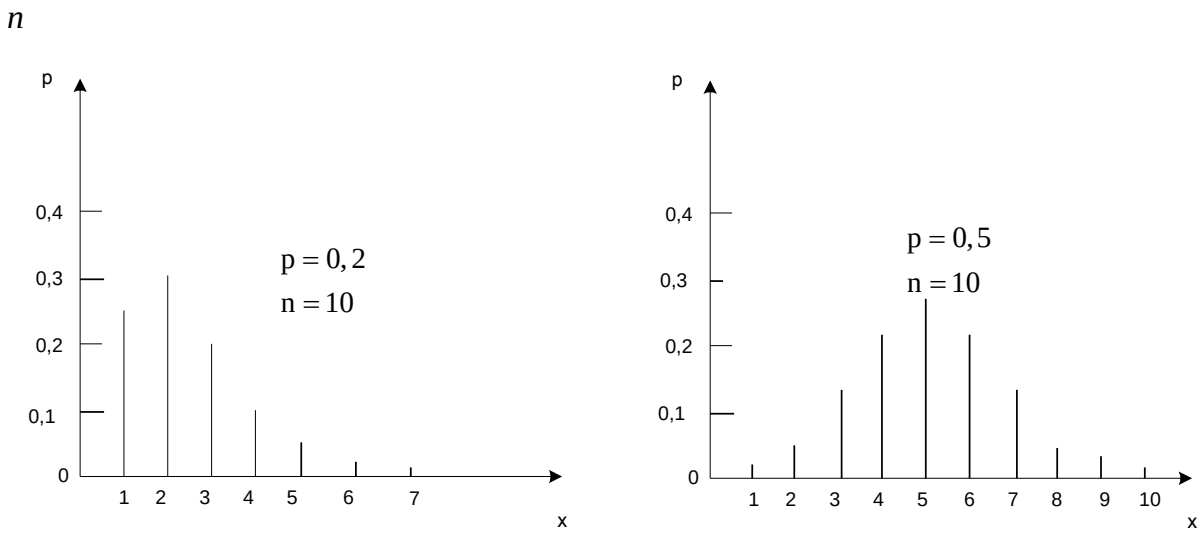


Рис. 1. **Полигон вероятностей Р.б. для различных значений p и n**

Р.б. может быть задано в виде ряда распределения, значения которого определяются по формуле бернулли и в виде функции распределения:

$$F(x)=\begin{cases}0, m\le 0;\\ \sum_{m=0}^{x} P(X=m), 0<m\le n;\\ 1, m>n.\end{cases}$$

Числовые характеристики:

Среднее: $M[X]=np$;

Дисперсия: $D[X]=np(1-p)$;

Асимметрия: $A_x=\dfrac{1-2p}{\sqrt{np(1-p)}}$;

Эксцесс: $\varepsilon_x=\dfrac{1-6p(1-p)}{np(1-p)}$.

Р.б. широко используется в теории и практике статистического контроля качества продукции, при описании функционирования систем массового обслуживания, в теории стрельбы и др. практических приложениях.

Биномиальная случайная величина может быть аппроксимирована: нормальной случайной величиной со средним np и дисперсией npq при $npq>5,\ 0{,}1\le p\le 0{,}9$. При $npq>25$ такую аппроксимацию можно применить независимо от значения p; пуассоновской случайной величиной со средним np при $p<0{,}1$.

Сумма k независимых биномиальных случайных величин с параметрами n, p имеет Р.б. с параметрами n', p, где $n'=\sum_{i=1}^{k} n_i$. Статистическая оценка параметра распределения $\tilde{p}=\dfrac{m}{n}$.

РАСПРЕДЕЛЕНИЕ ВЕЙБУЛЛА

распределение *случайной величины непрерывной* x с плотностью вероятности:

$$f(x)=\begin{cases}\dfrac{\alpha}{\beta^{\alpha}}x^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, & x\ge 0\\ 0, & x<0,\end{cases},$$

где α, β – параметры распределения.

Функция распределения определяется зависимостью: $F(x)=1-e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$.

Вид графика функции плотности и функции распределения сильно зависит от параметров распределения α и β (см. рис. 1).

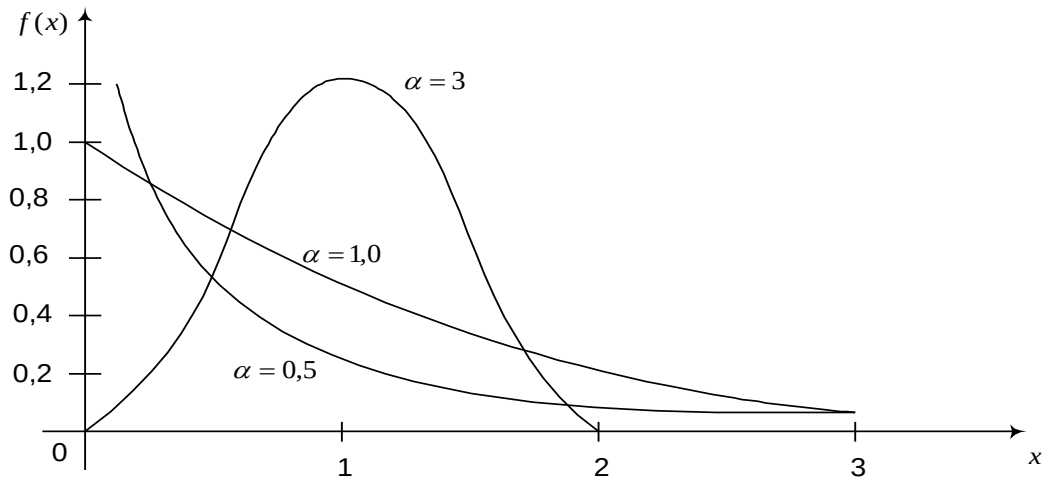


Рис. 1. **График функции плотности Р.В. ($\alpha, \beta = 1$).**

Числовые характеристики:

Среднее: $M[X] = \beta\,\Gamma\left(\dfrac{\alpha+1}{\alpha}\right)$;

Мода: $M_0 = \dfrac{\beta(\alpha-1)^{\frac{1}{\alpha}}}{\alpha^{\frac{1}{\alpha}}}$, $\alpha > 1$;

Дисперсия:

$$D[X] = \beta^2 \left\{\Gamma\left(\frac{\alpha+2}{\alpha}\right) - \left[\Gamma\left(\frac{\alpha+1}{\alpha}\right)\right]^2\right\};$$

Асимметрия:

$$A_x = \frac{\Gamma(1+\frac{3}{\alpha})\beta^3 - 3\nu\Gamma(1+\frac{2}{\alpha})\beta^2 - 2\nu^3}{\sigma^3},$$

где $\Gamma(z) = \int_0^\infty x^{x-1}e^{-x}dx$ – *гамма-функция Эйлера*;

$\nu^k = \int_{-\infty}^{\infty} x^k f(x)dx$ – начальный момент порядка k.

Данному распределению подчиняются пределы упругости стали, характеристики прочности некоторых материалов, время отказов разного типа в теории надежности. При $\alpha = 1$ Р.В. определяет экспоненциальное распределение с параметром

$$\lambda = \frac{1}{\beta}.$$

Статистические оценки параметров распределения определяются из решения системы уравнений:

$$\begin{cases} \tilde{\alpha} = \dfrac{n}{\dfrac{1}{\tilde{\beta}^{\tilde{\alpha}}}\sum_{i=1}^{n} x_i^{\tilde{\alpha}} \ln x_i - \sum_{i=1}^{n} \ln x_i} \\ \tilde{\beta} = \left[\dfrac{1}{n}\sum_{i=1}^{n} x_i^{\tilde{\alpha}}\right]^{\frac{1}{\tilde{\alpha}}} \end{cases}.$$

РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ

совокупность всех возможных значений *случайной величины* и соответствующих им *вероятностей*.

В зависимости от вида случайной величины её поведение может быть задано в виде ряда распределения, функции распределения или плотности вероятностей. Р.в. случайной величины *дискретной* определяется рядом распределения и функцией распределения.

Ряд распределения – совокупность значений случайной величины и вероятностей их наблюдения. Задается в виде табл., графика (*полигона*) распределения (см. рис. 1).

	x_1	x_2	x_3	⋯	x_{n-1}	x_n
p_i	p_1	p_2	p_3	⋯	p_{n-1}	p_n

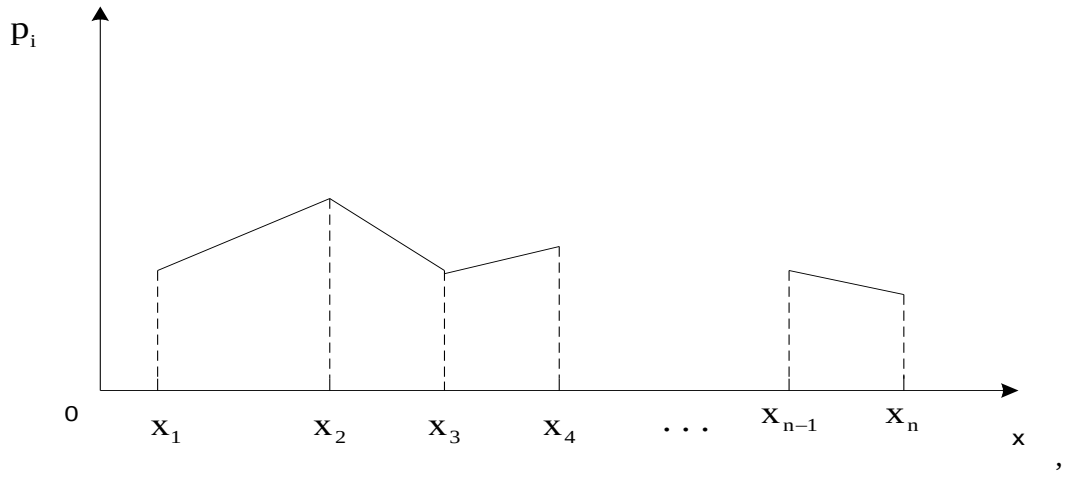


Рис. 1. **Полигон распределения**

где x_i – возможные значения случайной величины; p_i – вероятность появления x_i; $\sum_{i=1}^{n} p_i = 1$, или формулы, определяющей вероятности каждого значения случайной величины. Функция распределения – функция, значение которой в точке *x* выражает вероятность того, что случайная величина X примет значение меньшее, чем заданное *x:* $F(x) = P(X < x)$.

Для дискретной случайной величины X, которая может принимать значения $x_1, x_2, ..., x_n$, функция распределения имеет вид:

$$F(x) = \sum_{x_i < x} p(X = x_i),$$

где символ $x_i < x$ означает, что суммирование распространяется на все те возможные значения случайной величины, которые меньше аргумента *x*. Функция распределения дискретной случайной величины графически представляет собой ступенчатую функцию.

Р.в. случайной величины *непрерывной* определяется функцией распределения и плотностью распределения вероятностей. Функция распределения (интегральный закон распределения) определяется, как и для дискретной случайной величины: $F(x) = P(X < x)$.

Свойства функции распределения: 1. функция распределения монотонно возрастает в интервале от 0 до 1 $0 \le F(x) \le 1$; 2. вероятность появления случайной величины в интервале $[x_1; x_2]$ определяется зависимостью $p(x_1 \le X \le x_2) = F(x_2) - F(x_1).$; 3. функция распределения – неубывающая функция своего аргумента $F(x_2) \ge F(x_1)$, *если* $x_2 \ge x_1$; 4. $F(-\infty) = 0,\ F(+\infty) = 1.$; 5. вероятность появления любого частного значения непрерывной случайной величины равна нулю.

Плотность распределения (дифференциальный закон распределения) – предел отношения вероятности попадания непрерывной случайной величины *X* с функцией распределения $F(x)$ на элементарный участок $[x, x + \Delta x]$ к длине этого участка Δx, когда $\Delta x \to 0$, т.е. производная от функции распределения:

$$f(x) = \lim_{\Delta x \to 0} \frac{p(x < X < x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \frac{dF(x)}{dx}.$$

Кривая, изображающая плотность распределения $f(x)$, называется кривой u- распределения.

Свойства плотности распределения: 1. плотность распределения есть величина неотрицательная $f(x) \geq 0$; 2. интеграл в бесконечных пределах от плотности распределения равен единице $\int_{-\infty}^{\infty} f(x)dx = 1$; 3. вероятность появления случайной величины в интервале $[x_1; x_2]$ определяется зависимостью $p(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx$; 4. по плотности распределения можно определить функцию распределения $F(x) = \int_{-\infty}^{x} f(x)dx$. Любая интегрируемая функция $f(x)$, удовлетворяющая свойствам 1 и 2, является плотностью распределения непрерывной случайной величины *X*.

РАСПРЕДЕЛЕНИЕ ГРАММА – ШАРЛЬЕ

распределение, плотность которого задается рядом Грамма - Шарлье, определяемого выражением (типа А):

$$f_A(x) = f(x) + \sum_{k=3}^{n} a_k f^{(k)}(x),$$

где *x* – нормированное значение случайной величины *X*, или (типа В):

$$f_B(x) = \varphi(x) \sum_{m=0}^{n} b_m g_m(x).$$

Ряд $f_A(x)$ называется рядом Грамма - Шарлье типа А, если

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \text{ а } f^{(k)}(x)$$

– *k*-я производная от $f(x)$, которая может быть представлена как

$$f^{(k)}(x) = (-1)^k H_k(x) f(x),$$

где $H_k(x)$ – многочлен Чебышева - Эрмита.

Производные $f^{(k)}(x)$ и многочлены $H_k(x)$ обладают свойством ортогональности, благодаря чему коэффициенты a_k могут быть определены при помощи моментов данного ряда распределения.

Ряд $f_B(x)$ называется рядом Грамма - Шарлье типа В, если

$$\varphi(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0,1,2,...,$$

а $g_m(x)$ – многочлены, аналогичные $H_k(x)$.

Область эффективного использования Р.Г.-Ш. – оценка качества продукции и технологических процессов, в частности, при расчёте вероятности выхода значений признака качества за пределы поля допуска в случаях, когда поле рассеяния наблюдаемого признака оказывается в процессе работы сдвинутым относительно поля допуска (ряд Грамма - Шарлье типа А).

РАСПРЕДЕЛЕНИЕ ГИПЕРГЕОМЕТРИЧЕСКОЕ

случайная величина дискретная X имеет Р.г., если вероятность возможного значения случайной величины определяется по формуле:

$$P(X = \kappa) = \frac{C_M^{\kappa} C_{N-M}^{n-\kappa}}{C_N^n},$$

где *N* – размер ген. совокупности; *n* – размер выборки; *M* – число интересующих нас событий (элементов) в ген. совокупности; *к* – число интересующих нас событий (элементов) в выборке.

Числовые характеристики: Среднее: $M[X] = \frac{nM}{N}$; Дисперсия:

$$D[X] = \frac{nM(N-M)(N-n)}{N^2(N-1)};$$

Асимметрия:

$$A_x = \frac{(1-2\frac{M}{N})(N-2n)\sqrt{N-1}}{\sqrt{n\frac{M}{N}(1-\frac{M}{N})}(N-2)\sqrt{N-n}};$$

Эксцесс:

$$\varepsilon_x = \frac{C_1(N) - C_2(N)6\frac{M}{N}(1-\frac{M}{N})}{n\frac{M}{N}(1-\frac{M}{N})} + C_3(N) + C_4(N),$$

где $C_1(N) = \frac{(N-1)N(N+1)}{(N-2)(N-3)(N-n)};$

$$C_2(N) = \frac{(N-1)N^2}{(N-2)(N-3)(N-n);};$$

$$C_3(N) = 3\left[\frac{(N-1)N^2}{(N-2)(N-3)(N-n)} - 1\right];$$

$$C_4(N) = \frac{18(N-1)}{(N-2)(N-3)} - \frac{6(N-1)}{(N-2)(N-3)n\frac{M}{N}(1-\frac{M}{N})} - \frac{3(N-1)Nn}{(N-2)(N-3)(N-n)}.$$

Р.г. широко используется в практике *статистического контроля качества* продукции, в задачах, связанных с организацией выборочных обследований.

Типичная схема, в которой применяется Р.г.: проверяется партия готовой продукции, которая содержит M годных и N-M негодных изделий. Случайным образом выбирают *n* изделий. Число годных изделий к, среди выбранных, описывается Р.г.

При $\frac{n}{N} < 0{,}1$ Р.г. может быть приближённо заменено *распределением биномиальным* с вероятностью успеха $p = \frac{M}{N}$ и числом испытаний п.

РАСПРЕДЕЛЕНИЕ ДВУСТОРОННЕЕ ЭКСПОНЕНЦИАЛЬНОЕ ЛАПЛАСА

распределение *случайной величины непрерывной* X, с плотностью вероятности

$$f(x) = \frac{\lambda}{2} e^{-\lambda|x-\beta|}, \ -\infty < x < \infty,$$

где λ – параметр масштаба, $-\infty < \beta < \infty$ – параметр сдвига.

Функция распределения определяется зависимостью:

$$F(x) = \begin{cases} \frac{1}{2} e^{\lambda(x-\beta)}, & x \le \beta; \\ 1 - \frac{1}{2} e^{-\lambda(x-\beta)}, & x > \beta. \end{cases}$$

График функции плотности представляет собой как бы результат «склеивания» графика показательного распределения со своим зеркальным, относительно вертикальной оси, отражением (см. рис. 1).

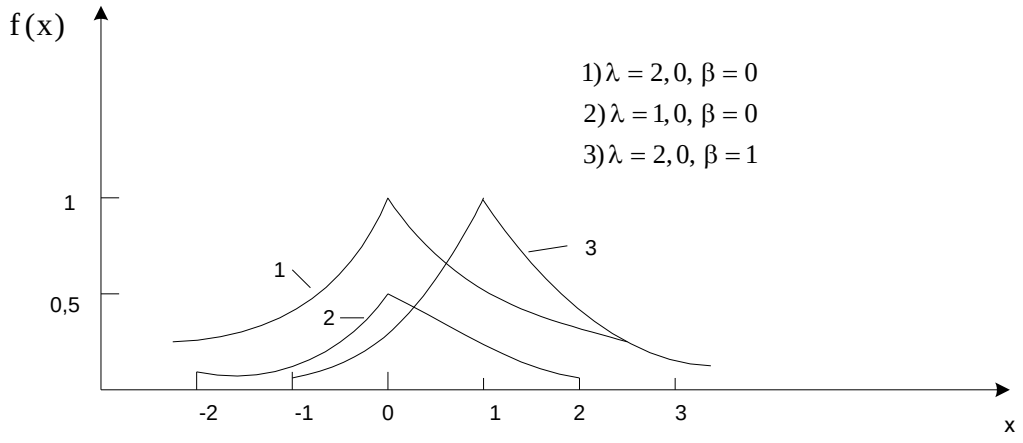


Рис. 1. **График функции плотности распределения Лапласа**

Числовые характеристики:

Среднее: $M[x] = \beta;$

Мода: $M_0 = \beta;$

Дисперсия: $D[x] = \frac{2}{\lambda^2};$

Асимметрия: $A_x = 0;$ Эксцесс: $\varepsilon_x = 3.$

Р.д.э.Л. используется для описания распределения остаточных случайных компонент (ошибок) в регрессионных моделях.

РАСПРЕДЕЛЕНИЕ КОШИ

распределение *случайной величины непрерывной* x с плотностью вероятности

$$f(x)=\frac{1}{\pi}\frac{c}{c^2+(x-a)^2},\quad -\infty<x<\infty,$$

где $c>0$ – параметр масштаба, a – параметр центра группирования, определяющий одновременно значения моды и медианы. Функция распределения определяется зависимостью:

$$F(x)=\frac{1}{2}+\frac{1}{\pi}arctg\frac{x-a}{c}.$$

Случайная величина, имеющая Р.К., является унимодальным, симметричным относительно своего модального значения и стандартным примером величины, не имеющей *математического ожидания* и *дисперсии*. График функции плотности имеет вид (см. рис. 1):

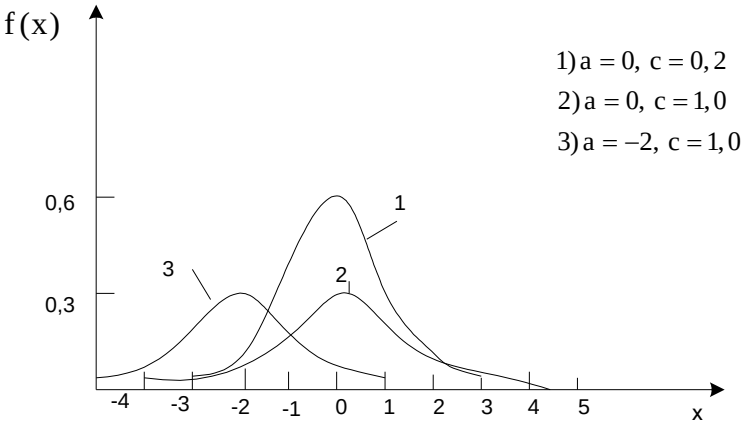


Рис. 1. **График функции плотности Р.К.**

Если случайная величина имеет Р.К. с параметрами c и a, то любая линейная функция b_0+b_1x имеет это же распределение с параметрами $c'=|b_1|\cdot c$ и $a'=b_0+b_1a$. Если случайные величины независимы и имеют Р.К., то их среднее арифметическое имеет тоже Р.К.

РАСПРЕДЕЛЕНИЕ ЛОГАРИФМИЧЕСКИ-НОРМАЛЬНОЕ

см. в ст. Логарифмически-нормальный закон распределения.

РАСПРЕДЕЛЕНИЕ МАКСВЕЛЛА

распределение *случайной величины непрерывной* X с плотностью вероятности

$$f(x)=\sqrt{\frac{2}{\pi}}\frac{x}{a^2}e^{\frac{-x^2}{2a^2}},$$

где $a>0$ – параметр распределения. Функция распределения определяется зависимостью

$$F(x)=2\Phi(\frac{x}{a})-\sqrt{\frac{2}{\pi}}\frac{x}{a}e^{\frac{-x^2}{2a^2}}-1,$$

где $\Phi(x)$ – функция стандартного нормального распределения.

Числовые характеристики:

Среднее: $M[x]=2\sqrt{\frac{2}{\pi}}a$;

Мода: $M_0=\sqrt{2}a$;

Дисперсия: $D[x]=\frac{3\pi-8}{\pi}a^2.$

Если x_1,x_2,x_3 – независимые случайные величины, имеющие нормальное распределение с параметрами $m_x=0,\sigma_x^2$, то случайная величина

$$\sqrt{x_1^2+x_2^2+x_3^2}$$

имеет Р.М. Т.о., Р.М. можно рассматривать как распределение длины случайного вектора, координаты которого в декартовой системе координат в трехмерном пространстве независимы и нормально распределены со средним 0 и дисперсией σ_x^2. Р.М. используется при изучении распределений существенно положительных случайных величин; часто применяется при статистическом анализе качества технологических процессов.

Параметр распределения a можно оценить, используя среднюю арифметическую выборочной совокупности

$\tilde{a} = 0{,}6267\bar{x}$, где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

РАСПРЕДЕЛЕНИЕ МАРГИНАЛЬНОЕ (ЧАСТНОЕ)

распределение *случайной величины* или множества случайных величин, рассматриваемых в качестве компонента или какой-то части компонентов некоторого случайного вектора с заданным законом распределения в ситуации, когда на значения оставшейся части компонентов вектора не накладывается никаких условий.

Выделим из системы величин $(x_1, x_2, ..., x_n)$ частную систему $(x_1, x_2, ..., x_m)$, тогда функция распределения этой системы определяется по формуле: $F_1(x_1, x_2, ..., x_m) = F(x_1, x_2, ..., x_m, \infty, ..., \infty)$ В частности, функция распределения каждой из величин, входящих в систему, получится, если в функции распределения системы положить все остальные аргументы равными $+\infty$, напр., $F_1(x_1) = F(x_1, \infty, ..., \infty)$.

Для описания закона распределения системы непрерывных случайных величин чаще используют плотность распределения системы. Зная функцию распределения системы, плотность распределения можно определить по формуле:

$$f(x_1, x_2, ..., x_n) = \frac{\partial^n F(x_1, x_2, ..., x_n)}{\partial x_1, \partial x_2, ..., \partial x_n}.$$

По плотности распределения можно определить функцию распределения:

$$F(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}...\int_{-\infty}^{x_n} f(x_1, x_2, ..., x_n)dx_1, dx_2, ..., dx_n$$

Чтобы найти плотность распределения любой подсистемы $(X_1, X_2, ..., X_m)$, входящей в систему $(X_1, X_2, ..., X_n)$, надо проинтегрировать совместную плотность $(n-m)$ раз по аргументам $(x_{m+1}, ..., x_n)$, относящимся к остальным случайным величинам:

$$f(x_1, x_2, ..., x_m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n)dx_{m+1}, ..., dx_n$$

$n-m$

В частности, плотность распределения одной случайной величины X_k, входящей в систему, равна:

$$f_k(x_k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty} f(x_1, ..., x_n)dx_1, ..., dx_{k-1}dx_{k+1}, ..., dx_n$$

$n-1$

.

РАСПРЕДЕЛЕНИЕ НОРМАЛЬНОЕ (ГАУССОВСКОЕ)

распределение *случайной величины непрерывной* X с параметрами m_x и σ_x, плотность распределения которой имеет вид:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{\frac{-(x-m_x)^2}{2\sigma_x^2}}.$$

Среди непрерывных случайных величин *распределение нормальное* (закон Гаусса) занимает центральное место. С ним приходится встречаться при анализе погрешностей измерений, контроле технологических процессов и режимов, при анализе и прогнозировании различных явлений в экономике, биологии, медицине и других областях знаний.

Нормальный закон очень широко применяется на практике. Он проявляется во всех тех случаях, когда случайная величина X является результатом действия большого числа различных факторов, причем каждый фактор в отдельности на случайную величину X влияет незначительно и не преобладает по своему влиянию над остальными, а характер воздействия – аддитивный.

Примерами случайных величин, имеющих нормальное распределение служат: отклонение действительных размеров деталей, обработанных на станке, от номинальных размеров; ошибки при различных измерениях; ошибки рассеивания снарядов (пуль) при стрельбе; курсы акций и другие.

Осн. особенностью, выделяющей нормальный закон среди других законов, состоит в том, что он является предельным законом, к которому, при определённых условиях, приближаются другие законы распределения.

Функция распределения нормального закона имеет вид:

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \int_{-\infty}^{x} e^{\frac{-(x-m_x)^2}{2\sigma_x^2}} dx,$$

где $M[X] = m_x$ – *математическое ожидание*;

$D[X] = \sigma_x^2$ – *дисперсия* (σ_x – стандартное отклонение).

График плотности Р.н. называют нормальной кривой ((кривой Гаусса) см. рис. 1).

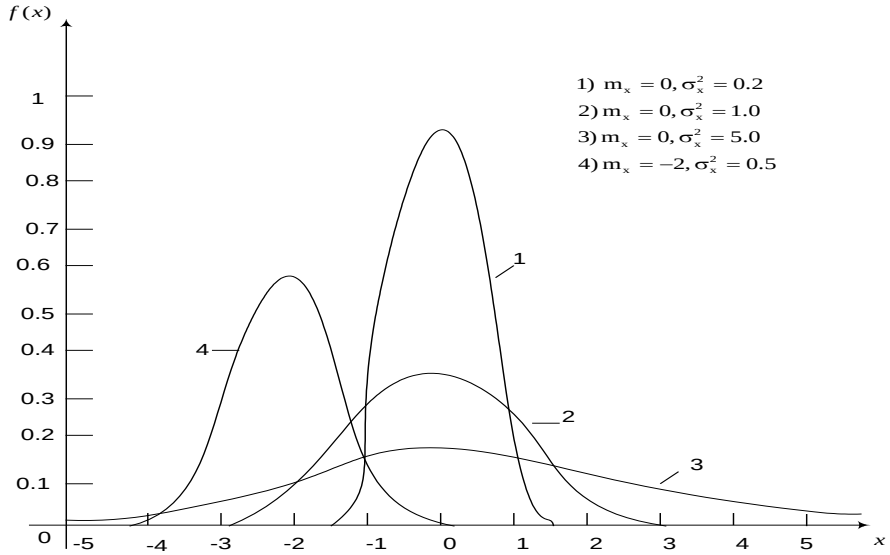


Рис. 1. **Функция плотности нормального распределения**

График интегральной функции нормального распределения (функции распределения) имеет вид (см. рис. 2):

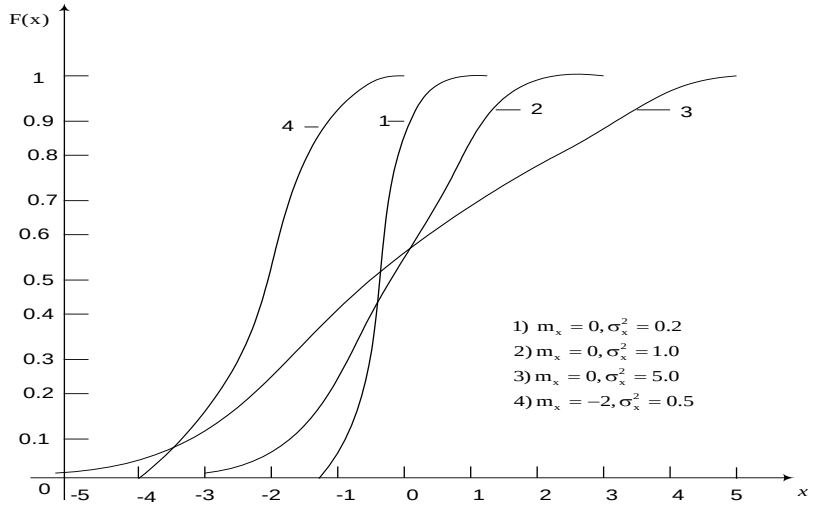


Рис. 2. **Функция распределения нормальной случайной величины**

Свойства Р.н.: 1. функция плотности Р.н. определена на всей оси ox, т.е. каждому значению x соответствует вполне определенное значение функции; 2. при всех значениях x функция плотности принимает положительные значения; 3. предел функции плотности при неограниченном возрастании (уменьшении) x равен нулю $\lim\limits_{x\to\pm\infty} f(x) = 0$; 4. функция плотности Р.н. в точке $x = m_x$ имеет макс., равный $f(x)_{\max} = \frac{1}{\sqrt{2\pi}\sigma_x}$; 5. график функции плотности $f(x)$ симметричен относительно прямой, проходящей через точку $x = m_x$. Отсюда следует равенство математического ожидания, *моды и медианы* для случайной величины, имеющей Р.н.; 6. кривая распределения имеет две точки перегиба с координатами $\left(m_x - \sigma_x; \frac{1}{\sqrt{2\pi e}\sigma_x}\right)$ и $\left(m_x + \sigma_x; \frac{1}{\sqrt{2\pi e}\sigma_x}\right)$; 7. нечётные центральные моменты Р.н. равны нулю. Используя определение центральных моментов, можно вывести следующее рекуррентное соотношение: $\mu_k = (k-1)\sigma_x^2\mu_{k-2}$. Можно убедиться, что центральные моменты нечётных порядков равны нулю, поскольку $\mu_1 = 0$, а для центральных моментов четного порядка получить следующие значения: $\mu_2 = \sigma_x^2$, $\mu_4 = 3\sigma_x^4$, $\mu_6 = 15\sigma_x^6$; 8. коэффициент асимметрии и эксцесс Р.н. равны нулю $A_x = \frac{\mu_3}{\sigma_x^3} = 0$; $\varepsilon_x = \frac{\mu_4}{\sigma_x^4} - 3 = 0$. Становится ясной важность вычисления этих коэффициентов для эмпирических рядов распределения, так как они характеризуют скошенность и крутость данного ряда по сравнению с нормальным; 9. форма нормальной кривой не изменяется при изменении математического ожидания. При уменьшении или увеличении m_x график кривой сдвигается влево или вправо; 10. при изменении среднеквадратического (стандартного) отклонения σ_x меняется форма кривой. С возрастанием σ_x макс. ордината кривой распределения уменьшается, а с уменьшением σ_x – возрастает.

Р.н. с параметрами $m_x = 0$ и $\sigma_x = 1$ называется стандартным Р.н., а плотность распределения называется нормированной плотностью, её график – нормированной нормальной кривой

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Нормированную нормальную кривую можно представить как кривую распределения нормированной случайной величины

$$T = \frac{(X - m_x)}{\sigma_x}.$$

На практике часто приходится вычислять вероятность того, что случайная величина, имею-

щая Р.н., находится в заданном интервале. Эту вероятность можно вычислить по формуле:

$$P(x_1 \le X \le x_2) = \frac{1}{2}\left[\Phi(z_2) - \Phi(z_1)\right],$$

где $\Phi(z) = \frac{2}{\sqrt{2\pi}} \int_0^z e^{-\frac{t^2}{2}} dt$ – функция Лапласа (интеграл вероятностей);

$z_1 = \frac{x_1 - m_x}{\sigma_x}$, $z_2 = \frac{x_2 - m_x}{\sigma_x}$ (см. рис. 1).

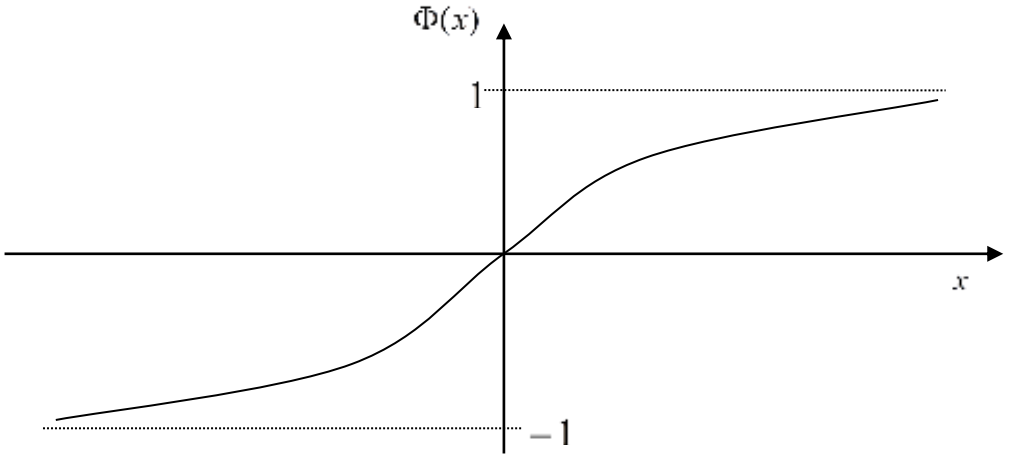


Рис. 1. **График функции Лапласа**

Интеграл $\int e^{-\frac{}{2}} dt$ не выражается через элементарные функции, поэтому для его вычисления используют специальные табл.

Функция Лапласа имеет свойства: функция нечётная, т.е. $\Phi(-z) = -\Phi(z)$; $\Phi(0) = 0$; $\Phi(\infty) = 1$. При симметричном интервале относительно математического ожидания:

$$P(|X - m_x| < \varepsilon) = \Phi\left(\frac{\varepsilon}{\sigma_x}\right).$$

Через функцию Лапласа выражается и функция распределения $F(x)$ нормальной случайной величины X.

РАСПРЕДЕЛЕНИЕ НОРМАЛЬНОЕ ДВУМЕРНОЕ

распределение системы двух *случайных величин непрерывных* X и Y с плотностью вероятности:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r_{xy}^2}} e^{-\frac{1}{2(1-r_{xy}^2)}\left[\frac{(x-m_x)^2}{\sigma_x^2} - 2r_{xy}\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right]},$$

где m_x, m_y – *математические ожидания* случайных величин X и Y;

σ_x, σ_y – стандартные отклонения случайных величин X и Y;

r_{xy} – *коэффициент корреляции* случайных величин X и Y.

Из всех законов распределения системы двух случайных величин наибольшее применение на практике имеет нормальное распределение. Для случая независимых случайных величин

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\frac{(x-m_x)^2}{\sigma_x^2} + \frac{(y-m_y)^2}{\sigma_y^2}\right]}.$$

График функции плотности нормального распределения системы 2-х случайных величин имеет вид (см. рис. 1):

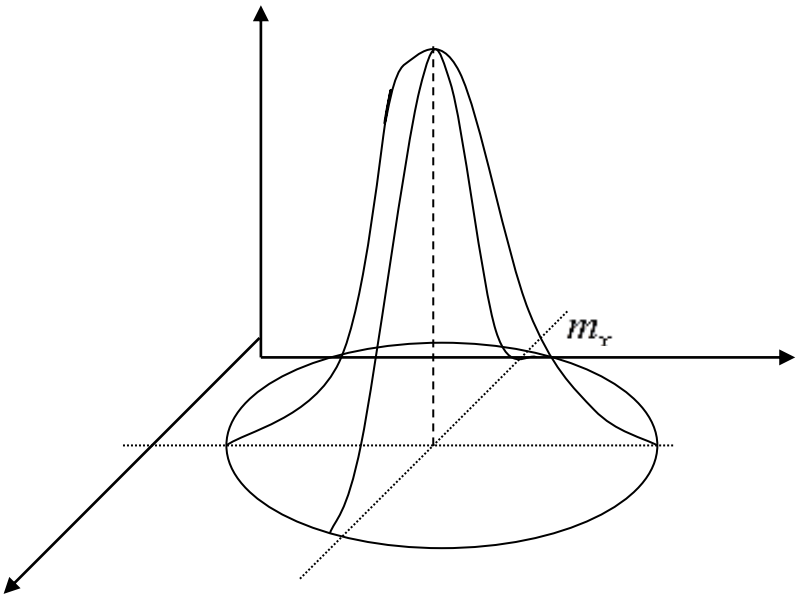


Рис. 1. **График нормального распределения системы 2-х случайных величин**

Р.н.д. используется при исследовании связи между двумя показателями.

РАСПРЕДЕЛЕНИЕ НОРМАЛЬНОЕ МНОГОМЕРНОЕ

совместное распределение нескольких *случайных величин* X1, X2,..., Xn называется многомерным нормальным, если соответствующая плотность вероятности имеет вид:

$$f(x_1, x_2, .., x_n) = Ke^{-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)},$$

где $K = (2\pi)^{-\frac{1}{2}n} \left|\sum\right|^{-\frac{1}{2}}$, $(x-\mu)$

– вектор, компонентами которого являются отклонения случайной величины от *математического ожидания*, $(x-\mu)^T$ – транспонированный вектор, μ – математическое ожидание случайного вектора X, $\sum$ – *ковариационная матрица* (невырожденная)

$$\sum = \begin{pmatrix} \sigma_{11}\sigma_{12} \dots \sigma_{1n} \\ \sigma_{21}\sigma_{22} \dots \sigma_{2n} \\ \dots\dots\dots\dots\dots \\ \sigma_{n1}\sigma_{n2} \dots \sigma_{nn} \end{pmatrix},$$

где $\left|\sum\right|$ – определитель ковариационной матрицы; $\sum^{-1}$ – матрица, обратная ковариационной; σ_{ij} – элемент ковариационной матрицы; $\sigma_{ii} - \sigma_i^2$ – *дисперсия* случайной величины Xi; σ_i – отклонение среднее квадратическое.

Частный случай Р.н.м. – одномерный нормальный закон распределения:

при $K = 1/\sigma\sqrt{2\pi}$; $\sum^{-1} = 1/\sigma^2$

поэтому $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Если случайные величины, являющиеся компонентами случайного вектора, имеющего Р.н.м. не коррелированны между собой, то они независимы. Важнейшие свойства Р.н.м.: частные условные распределения от Р.н.м. также являются нормальными распределениями; сумма X независимых случайных величин X_1 и X_2, распределённых нормально, имеет нормальное распределение; обратно, если $X = X_1 + X_2$ имеет нормальное распределение и X_1 и X_2 независимы, то X_1 и X_2 также распределены нормально (теорема Крамера).

Общее количество параметров k, задающих Р.н.м, равно $1/2(n+1)(n+2)-1$. С увеличением n оно быстро растёт. Так, k=2 при n=1, k=20 при n=5, k=65 при n=10.

РАСПРЕДЕЛЕНИЕ ОТРИЦАТЕЛЬНОЕ БИНОМИАЛЬНОЕ

случайная величина X имеет Р.о.б. (распределение Паскаля) с параметрами (r, p), если в последовательности *Бернулли испытания* с вероятностью успеха p и вероятностью неудачи $q = 1 - p$ вероятность числа неудач *к*, происшедших до r-го успеха определяется по формуле:

$$P(x = \kappa) = C^{\kappa}_{r+\kappa-1} p^r (1-p)^{\kappa},$$

где r – число успехов, целое положительное число; *к* – число неудач происшедших до числа успехов *r*.

Полигон вероятностей Р.о.б. имеет вид (см. рис. 1).

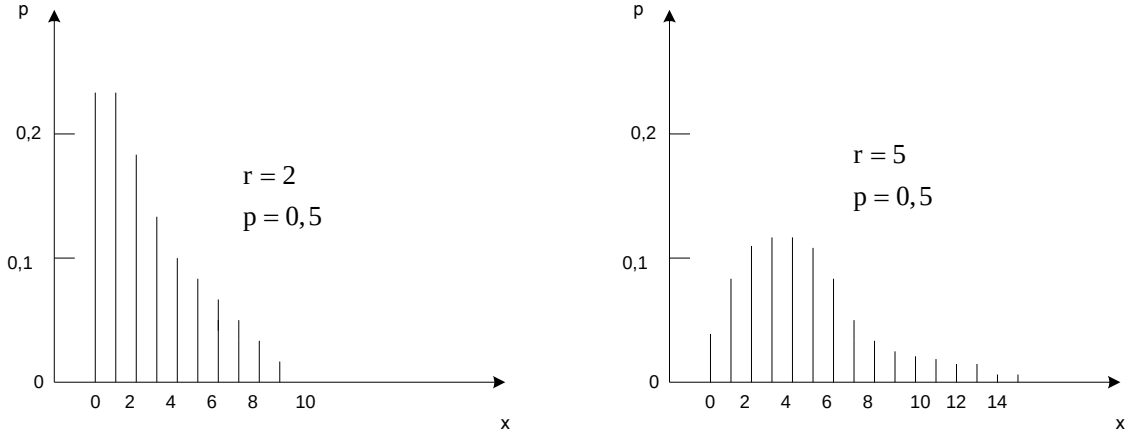


Рис. 1. **Полигон вероятностей для различных значений** r **при** $p = 0,5$

Числовые характеристики:

Среднее: $M[X] = \dfrac{r(1-p)}{p}$;

Дисперсия: $D[X] = \dfrac{r(1-p)}{p^2}$;

Асимметрия: $A_x = \dfrac{2-p}{\sqrt{r(1-p)}}$;

Эксцесс: $\varepsilon_x = \dfrac{6}{r} + \dfrac{p^2}{r(1-p)}$.

Р.о.б. используется в статистике несчастных случаев и заболеваний; в задачах, связанных с количеством особей данного вида в выборках из биологических популяций, в задачах оптимального резервирования элементов, в теории стрельбы. При целом $r > 0$ Р.о.б. интерпретируется как распределение времени ожидания r - го «успеха» в схеме испытаний Бернулли с вероятностью «успеха» p, напр., количество испытаний до второго выпадения герба, в этом случае оно иногда называется распределением Паскаля и является дискретным аналогом *гамма - распределения.*

При $r = 1$ Р.о.б. совпадает с геометрическим распределением.

При выборе между *распределениями биномиальным*, отрицательным биномиальным и Пуассона можно руководствоваться соотношениями: биномиальное – дисперсия<среднее; отрицательное биномиальное – дисперсия>среднее; Пуассона – дисперсия = среднее.

Статистическая оценка параметра распределения

$$\tilde{p} = \frac{r-1}{r+k-1}.$$

РАСПРЕДЕЛЕНИЕ ПАРЕТО

распределение *случайной величины непрерывной* X с плотностью вероятности

$$f(x) = \begin{cases} \dfrac{\alpha}{c_0}\left(\dfrac{c_0}{x}\right)^{\alpha+1}, & x_0 < x < \infty; \\ 0, & x \le x_0. \end{cases}$$

Функция распределения определяется зависимостью:

$$F(x) = 1 - \left(\frac{c_0}{x}\right)^{\alpha},$$

где $\alpha > 0,\ c_0 (x > c_0)$ – параметры распределения, т.е. область значений случайной величины X есть полупрямая $(c_0, +\infty)$.

Функция плотности имеет вид монотонно убывающей кривой, выходящей из точки $(c_0, \alpha / c_0)$ (см. рис. 1).

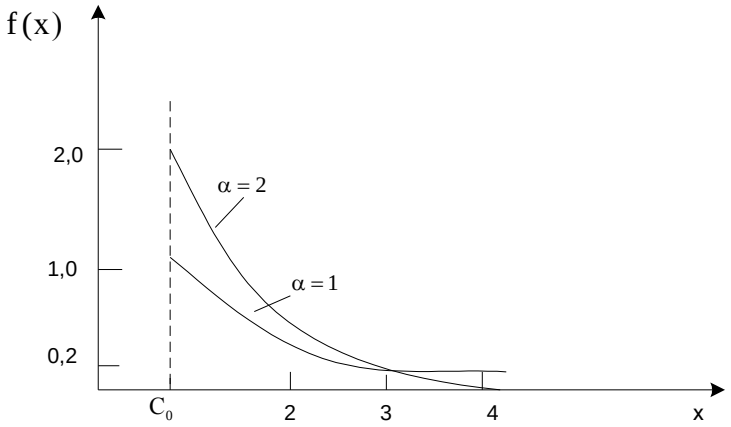


Рис.1. **График функции плотности Р.П.**

Осн. числовые характеристики существуют не всегда, а лишь при соблюдении определенных требований к параметру α.

Среднее: $M[x] = \frac{\alpha}{\alpha - 1} c_0,\ \alpha > 1;$

Мода: $M_0 = c_0;$

Дисперсия: $D[x] = \frac{\alpha}{(\alpha - 1)^2 (\alpha - 2)} c_0^2,\ \alpha > 2;$

Асимметрия: $A_x = \frac{2(1+\alpha)}{\alpha - 3} \sqrt{\frac{\alpha - 2}{\alpha}},\ \alpha > 3;$

Эксцесс: $\varepsilon_x = \frac{6(\alpha^3 + \alpha^2 - 6\alpha - 2)}{\alpha(\alpha - 3)(\alpha - 4)},\ \alpha > 4.$

Р.П. используется в страховании или налогообложении, когда интерес представляют ущербы и доходы, которые превосходят некоторую величину c_0. Статистическая оценка параметра распределения

$$\tilde{\alpha} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \ln x_i}.$$

РАСПРЕДЕЛЕНИЕ ПОЛИНОМИАЛЬНОЕ (МУЛЬТИНОМИАЛЬНОЕ)

обобщение *распределения биномиального* на случай k событий. Пусть имеется последовательность независимых испытаний, в каждом из которых может осуществиться одно из k попарно несовместных событий $A_1, A_2, ..., A_k$ с вероятностями $p_1, p_2, ..., p_k$ соответственно $(p_1 + p_2 + ... + p_k = 1)$.

Т.о., при каждом испытании события $A_1, A_2, ..., A_k$ образуют полную группу событий. Тогда распределение вероятностей $P(x_1, x_2, ..., x_k)$ того, что в результате *n* независимых испытаний события $A_1, A_2, ..., A_k$ появятся соответственно $x_1, x_2, ..., x_k$ раз, причем

$$\sum_{i=1}^{k} x_i = n,$$

называется полиномиальным и определяется формулой:

$$P(x_1, x_2, ..., x_k) = \frac{n!}{x_1! x_2! ... x_k!} \cdot p_1^{x_1} p_2^{x_2} ... p_k^{x_k},$$

где $n! = n(n-1)(n-2)...2 \cdot 1.$

В случае $k = 2$ Р.п. сводится к биномиальному. При достаточно большом *n* Р.п. можно аппрок-

симировать распределением непрерывной случайной величины χ^2 («хи» – квадрат).

Числовые характеристики:

Средние: $M[x_i] = np_i, \quad i = \overline{1,k};$

Дисперсии: $D[x_i] = np_i(1-p_i), \quad i = \overline{1,k};$

Ковариации: $\sigma_{ij} = -mp_i p_j, \quad i, j = \overline{1,k}, \quad i \neq j.$

П.р. применяется при статистической обработке выборок из больших совокупностей, элементы которых разделяются более чем на две категории (напр., в различных социологических, экономико-социологических, медицинских и др. выборочных обследованиях).

РАСПРЕДЕЛЕНИЕ ПУАССОНА

распределение *случайной величины дискретной*, при котором она может принять одно из возможных значений 0, 1, 2, … , *n* с вероятностью

$$P(X = m) = P_m = \frac{\lambda^m}{m!} e^{-\lambda},$$

где *m*=0, 1, 2, … , *n*; $\lambda = np$ – параметр распределения, характеризующий интенсивность появления событий в п испытаниях (см. рис. 1).

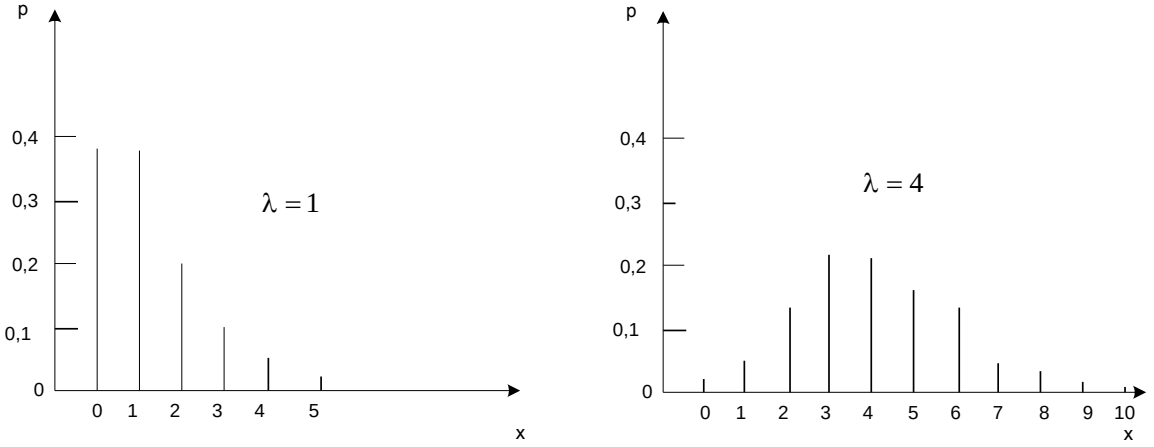


Рис. 1 **Полигон вероятностей Р.П. для различных значений** λ

Р. П. может быть задано в виде ряда распределения, значения которого определяются по приведённой выше формуле, и в виде функции распределения:

$$F(x) = \begin{cases} 0, m \le 0; \\ \sum_{m=0}^{x} P(X = m), 0 < m \le n; \\ 1, m > n. \end{cases}$$

Числовые характеристики:

Среднее: $M[X] = np = \lambda;$

Дисперсия: $D[X] = \lambda;$

Асимметрия: $A_x = \frac{1}{\sqrt{\lambda}};$

Эксцесс: $\varepsilon_x = \frac{1}{\lambda}.$

Р.П. – предельный случай *распределения биномиального* при $p \to 0, \quad n \to \infty$, т. о. что $np = \lambda$.

Отсюда следует, что Р.П. с параметром $\lambda = np$ можно применять вместо биномиального, когда число опытов *n* достаточно велико, а вероятность *p* – достаточно мала, т.е. в каждом отдельном опыте интересующее событие происходит крайне редко, при этом $np \approx npq$. Отсюда происходит название «закон редких явлений», применяющееся иногда для закона Пуассона,

Примеры случайных величин, имеющих Р.П.: число α-частиц, испускаемых радиоактивным источником за определённый промежуток времени; число требований на выплату страховых сумм за год; число «требований на обслуживание», поступивших в систему массового обслуживания (СМО) за единицу времени; число отказов элементов при испытании на надёж-

ность сложных устройств; число несчастных случаев и редких заболеваний и др.

Статистическая оценка параметра распределения:

$$\tilde{\lambda} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

РАСПРЕДЕЛЕНИЕ РАВНОМЕРНОЕ (ПРЯМОУГОЛЬНОЕ)

распределение *случайной величины непрерывной* X на отрезке $[a,b]$ с постоянной плотностью вероятности:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0, & x < a, x > b. \end{cases}$$

Функция распределения определяется зависимостью:

$$F(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & \text{a} \le x \le \text{b}, \\ 1, & x > \text{b}. \end{cases}$$

Графики функций распределения и плотности вероятности представлены на рис. 1.

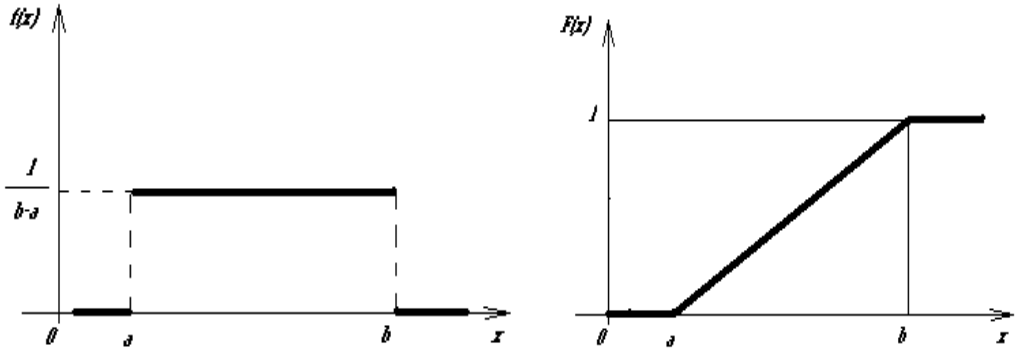


Рис. 1. **Графики функций распределения и плотности вероятности**

Числовые характеристики:

Среднее, медиана: $m_x = Me = \dfrac{a+b}{2}$;

Дисперсия: $\sigma_x^2 = \dfrac{(b-a)^2}{12}$;

Асимметрия: $A_x = 0$;

Эксцесс: $\varepsilon_x = -\dfrac{6}{5}$.

Р.р. применяется при анализе ошибок округления; оценке времени ожидания «обслуживания» при периодическом, через каждые T единиц времени, включения (прибытия) «обслуживающего устройства». Р.р. иногда используется в качестве «нулевого приближения» в описании априорного распределения анализируемых параметров в байесовском подходе в условиях полного отсутствия априорной информации об этом распределении. Широко Р.р. используется для генерации случайных чисел, имеющих различные распределения (нормальное, Пуассона, биномиальное, показательное и др.).

Статистическая оценка параметров распределения

$$\tilde{a} = \bar{x} - 3{,}5s; \quad \tilde{b} = \bar{x} + 3{,}5s,$$

где $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$; $\quad s = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

РАСПРЕДЕЛЕНИЕ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

функция, которая однозначно определяет вероятность того, что *случайная величина* принимает заданное значение или принадлежит к некоторому заданному интервалу.

Если случайная величина принимает конечное число значений, то распределение задается функцией P(X = x), ставящей каждому возможному значению x случайной величины X вероятность того, что X = x.

Случайная величина принимает бесконечно много значений. Это возможно лишь тогда, когда вероятностное пространство, на котором определена случайная величина, состоит из бесконечного числа элементарных событий. Тогда распределение задается набором вероятностей P(a ≤X <b) для всех пар чисел a, b таких, что a<b. Распределение может быть задано с помощью т.н. функции распределения F(x) = P(X<x), определяющей для всех действительных *x* вероятность того, что случайная величина X принимает значения, меньшие x: P(a ≤X <b) = F(b) – F(a). Это соотношение показывает, что как распределение может быть рассчитано по функции распределения, так и, наоборот, функция распределения – по распределению.

Используемые в вероятностно-статистических методах и других прикладных исследованиях функции распределения бывают либо дискретными, либо непрерывными, либо их комбинациями. Дискретные функции распределения соответствуют случайным величинам *дискретным,* принимающим конечное число значений или же значения из множества, элементы которого можно перенумеровать натуральными числами. Непрерывные функции распределения соответствуют случайным величинам *непрерывным*. Они монотонно возрастают при увеличении аргумента – от 0 при $x \to -\infty$ до 1 при $x \to +\infty$. Непрерывные функции распределения, используемые в вероятностно-статистических методах, имеют производные. Первая производная *f(x)* функции распределения *F(x)* называется плотностью вероятности:

$$f(x) = \frac{dF(x)}{dx}$$

По плотности вероятности можно определить функцию распределения:

$$f(x) = \int_{-\infty}^{x} f(y)dy$$

Для любой функции распределения: $\lim_{x\to-\infty} F(x) = 0,\ \lim_{x\to+\infty} F(x) = 1,$

а потому $\int_{-\infty}^{+\infty} f(x)dx = 1.$

Функция распределения полностью характеризует случайную величину с вероятностной точки зрения. Зная функцию распределения случайной величины, можно указать, где располагаются возможные значения случайной величины и какова вероятность появления её в том или ином интервале.

РАСПРЕДЕЛЕНИЕ СОВМЕСТНОЕ (МНОГОМЕРНОЕ)

распределение вероятностей, для которого многомерной функцией распределения *n случайных величин* $X_1, X_2, ..., X_n$ или n – мерной функцией распределения случайного вектора является вероятность совместного выполнения *n* неравенств вида $X_i < x_i$:

$$F(x_1, x_2, ..., x_n) = P((X_1 < x_1) \cdot (X_2 < x_2)...(X_n < x_n)),$$

где x_i – значение случайной величины X_i.

Плотностью распределения n – мерного случайного вектора, компоненты которого являются непрерывными случайными величинами, называется *n*-я смешанная частная производная функции $F(x_1, x_2, ..., x_n)$, взятая один раз по каждому аргументу

$$f(x_1, x_2, ..., x_n) = \frac{\partial^n F(x_1, x_2, ..., x_n)}{\partial x_1 \partial x_2 ... \partial x_n}.$$

При независимости компонент случайного вектора плотность распределения непрерывных случайных величин $X_1, X_2, ..., X_n$ равняется произведению плотностей распределения отдельных случайных величин $X_1, X_2, ..., X_n$

$$f = (x_1, x_2, ..., x_n) = f_1(x_1) f_2(x_2) ... f_n(x_n).$$

РАСПРЕДЕЛЕНИЕ СТЬЮДЕНТА (*t* - РАСПРЕДЕЛЕНИЕ)

распределение *случайной величины непрерывной:*

$$T = \sqrt{n}\,\frac{\bar{x} - x_0}{\hat{s}},$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\hat{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$;

x_i – независимые нормальные случайные величины с $M[x_i] = m_x$ и $D[x_i] = \sigma_x^2$.

Функция плотности Р.С. имеет вид:

$$f(t) = \frac{\Gamma\left(\frac{\kappa+1}{2}\right)}{\sqrt{\kappa\pi}\,\Gamma\left(\frac{\kappa}{2}\right)}\left(1+\frac{t^2}{\kappa}\right)^{-\frac{\kappa+1}{2}},$$

где $\kappa = n-1$ – число степеней свободы; n – число наблюдений (опытов);

$\Gamma(s) = \int e^{-x} x^{s-1} dx$ – *гамма-функция Эйлера.*

Интегральная функция Р.С. определяется зависимостью:

$$F(t) = \frac{\tilde{A}\left(\frac{\hat{e}+1}{2}\right)}{\sqrt{\hat{e}\pi}\,\tilde{A}\left(\frac{\hat{e}}{2}\right)} \int_{-\infty}^{t}\left(1+\frac{t^2}{\hat{e}}\right)^{-\frac{\hat{e}+1}{2}} dt .$$

График функции плотности Р.С. представлен на рис. 1:

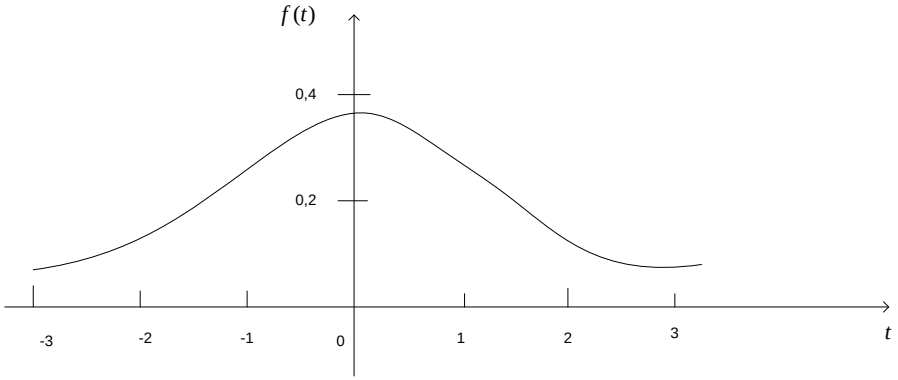


Рис. 1.

Числовые характеристики:

Среднее, мода: $M[T] = M_O = 0$;

Дисперсия: $D[T] = \frac{\kappa}{\kappa-2}, (\kappa > 2)$;

Асимметрия: $A_t = 0,\ k > 3;$

Эксцесс: $\varepsilon_t = \frac{6}{k-4},\ k > 4.$

Р.С. не зависит от *математического ожидания* и *дисперсии* случайной величины *X*, а зависит лишь от объёма выборки *n*. Функцию плотности часто называют законом распределения статистики - *t* или *t*- Р.С.

Р.С. в статистике используется для построения доверительных интервалов и проверки *гипотез статистических* при использовании малых выборок. При больших значениях $\kappa = n-1$ Р.С. асимптотически приближается к стандартному нормальному распределению.

РАСПРЕДЕЛЕНИЕ УСЕЧЁННОЕ

распределение, в котором крайние значения (с одного или обоих концов) были "отсечены". Усечение может быть сделано вследствие решения исключить эти значения из рассмотрения или просто из-за невозможности собрать данные о крайних значениях.

Если *случайная величина X* имеет функцию распределения $F(x)$, заданную по всей числовой прямой то распределение случайной величины *X* со значениями, принадлежащими только рассматриваемому отрезку, напр., $[a;b]$, называется усечённым. Функция распределения выражается через исходную:

$$F(x|a \le X < b) = \begin{cases} 0 & при\ \ x < a, \\ \dfrac{F(x)-F(a)}{F(b)-F(a)} & при\ \ a \le x \le b, \\ 1 & при\ \ x > b. \end{cases}$$

Если *X* имеет плотность $f(x)$, то плотность Р.у. выражается через плотность $f(x)$ по формуле:

$$f(x \mid a \le X < b) = \begin{cases} 0 & при \quad x < a, \\ \dfrac{f(x)}{\int_a^b f(x)dx} & при \quad a \le x \le b, \\ 0 & при \quad x > b. \end{cases}$$

В этой формуле интеграл можно заменить соответствующей суммой, если речь идет о дискретной случайной величине *X*.

Р.у. применяется в демографии, теории надежности. Классический пример строго усечённого распределения – плотность распределения значений IQ в университетах, в которых более низкая часть всей популяции была удалена в соответствии с требованиями для поступления.

РАСПРЕДЕЛЕНИЕ УСЛОВНОЕ

любой подсистемы $(X_1, X_2, ..., X_m)$, входящей в систему $(X_1, X_2, ..., X_n)$ – её закон распределения, вычисленный при условии, что остальные величины $X_{m+1}, ..., X_n$ приняли значения $x_{m+1}, ..., x_n$.

Условная функция распределения определяется зависимостью:

$$F(x_1, x_2, ..., x_m / x_{m+1}, ..., x_n) = P(x_1 < x_1, x_2 < x_1, ..., x_m < x_m, x_{m+1} = x_{m+1}, ..., x_n = x_3)$$

.

Условная плотность распределения определяется по формуле:

$$f(x_1, x_2, ..., x_m / x_{m+1}, ..., x_n) = \frac{f(x_1, x_2, ..., x_n)}{f_1(x_{m+1}, ..., x_n)}.$$

Для системы $(X_1, X_2, ..., X_n)$ независимых случайных величин плотность распределения равна произведению плотностей распределения отдельных величин, входящих в систему:

$$f(x_1, x_2, ..., x_n) = f_1(x_1) f_2(x_2) ... f_n(x_n).$$

Законы распределения системы n зависимых случайных величин, являющиеся функциями многих аргументов, чаще всего неудобны в практическом применении и к тому же для своего определения (хотя бы приближенного) требуют большого объёма экспериментальных данных. В большинстве экономических и инженерных приложений вместо законов распределения рассматриваются важнейшие числовые характеристики системы.

Условный закон распределения для системы двух случайных величин можно задаётся условной функцией распределения $F(x/y)$ и условной плотностью распределения $f(x/y)$. Условные плотности распределения имеют большее применение, т.к. системы непрерывных случайных величин имеют осн. практическое значение.

Для определения плотности распределения системы необходимо в общем случае знание плотности распределения одной случайной величины, входящей в систему, и условной плотности распределения другой случайной величины, входящей в эту систему. Поэтому условную плотность распределения системы двух случайных величин можно

$$f_1(x/y) = \frac{f(x,y)}{\int_{-\infty}^{\infty} f(x,y)dx},$$

определить по зависимостям:

$$f_2(x/y) = \frac{f(x,y)}{\int_{-\infty}^{\infty} f(x,y)dy}.$$

Условная плотность распределения обладает всеми свойствами безусловной плотности распределения

$$f_1(x/y) \ge 0, \quad \int_{-\infty}^{\infty} f_1(x/y)dx = 1,$$

$$f_2(y/x) \ge 0, \quad \int_{-\infty}^{\infty} f_2(y/x)dy = 1.$$

.

РАСПРЕДЕЛЕНИЕ «ХИ–КВАДРАТ»

(χ^2) – распределение *случайной величины непрерывной* $Y = X_1^2 + X_2^2 + ... + X_n^2$, где X_i – независимые нормальные случайные величины с $m_x = 0$ и $\sigma_x^2 = 1$.

Функция плотности распределения

$$\chi^2:\ f(y)=\frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}e^{-\frac{y}{2}}y^{\frac{n}{2}-1},\quad y\ge 0,$$

где $n=\kappa$ – число степеней свободы;

$\Gamma\left(\frac{n}{2}\right)$ – *гамма-функция Эйлера.*

График функции плотности распределения χ^2 имеет вид (см. рис. 1).

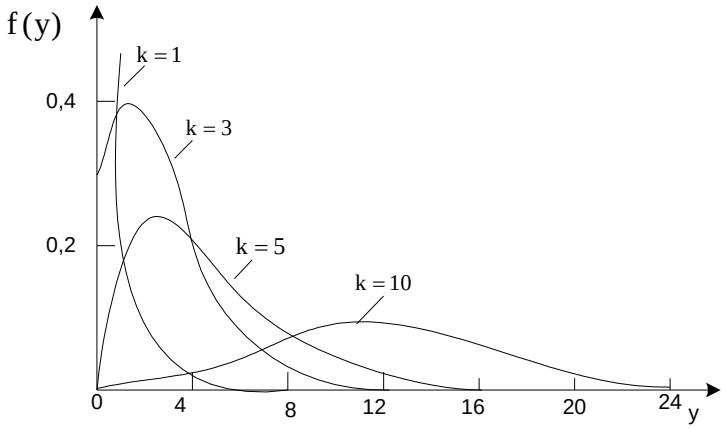


Рис. 1. **График функции плотности распределения** χ^2

Числовые характеристики:

Среднее: $M[Y]=k$;

Мода: $M_0=k-2,\ k\ge 2$;

Дисперсия: $D[Y]=2k$;

Асимметрия: $A_y=\frac{2^{3/2}}{\sqrt{k}}$;

Эксцесс: $\varepsilon_y=\frac{12}{k}$,

где k – число степеней свободы, равное числу независимых слагаемых.

Распределение χ^2 не зависит от числовых характеристик ($M[Y]$ и $D[Y]$), а зависит лишь от объёма выборки *n*.

Функция распределения χ^2 – распределения определяется зависимостью:

$$F(y)=\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\int_0^y e^{-\frac{u}{2}}u^{\frac{n}{2}-1}du.$$

Распределение χ^2 – частный случай *гамма-распределения* и обладает всеми его свойствами. С увеличением n распределение достаточно медленно приближается к нормальному. В статистике используется при построении *доверительных интервалов* для *дисперсии* (стандартного отклонения) и проверке *гипотез статистических*.

РАСПРЕДЕЛЕНИЕ ФИШЕРА (*F* – РАСПРЕДЕЛЕНИЕ)

распределение *случайной величины непрерывной*

$$X=\frac{X_1/\kappa_1}{X_2/\kappa_2},$$

где X_1 и X_2 – независимые случайные величины, имеющие χ^2 распределение, соответственно с k_1 и k_2 степенями свободы.

Если $x_1,x_2,...x_{k_1}$ – выборка из нормальной ($m_x,\sigma_x{}^2$) ген. совокупности, а $y_1,y_2,...,y_{k_2}$ – выборка из нормальной ($m_y,\sigma_y{}^2$) ген. совокупности,

то статистика

$$F=\frac{\frac{1}{k_1-1}\sum_{i=1}^{k_1}(x_i-\bar{x})^2}{\frac{1}{k_2-1}\sum_{i=1}^{k_2}(y_i-\bar{y})^2},$$

где $\bar{x}=\frac{1}{k_1}\sum_{i=1}^{k_1}x_i$, $\bar{y}=\frac{1}{k_2}\sum_{i=1}^{k_2}y_i$,

имеет F – распределение с (k_1-1,k_2-1) степенями свободы.

F – распределение имеет функцию плотности:

$$f(x)=\frac{\left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}}{B\left(\frac{k_1}{2},\frac{k_2}{2}\right)}x^{\frac{k_1}{2}-1}\left(1-\frac{k_1}{k_2}x\right)^{-\frac{k_1+k_2}{2}},\ x>0,$$

где $B(r,s)=\int x^{r-1}(1-x)^{s-1}dx$ – бета-функция.

График функции плотности F – распределения имеет вид (см. рис. 1).

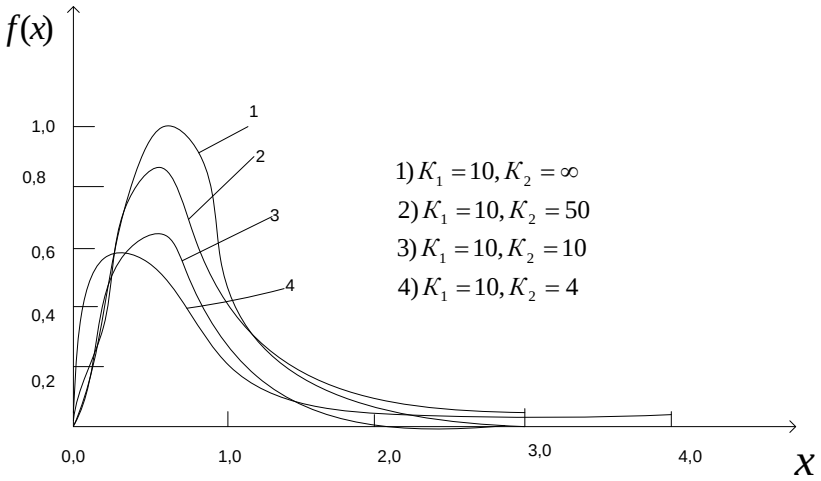


Рис. 1. **График функции плотности F – распределения**

Числовые характеристики:

Среднее: $M[X]=\frac{k_2}{k_2-2}$, $k_2>2$;

Мода: $M_0=\frac{(k_1-2)k_2}{k_1(k_2+2)}$, $k_1>1$;

Дисперсия: $D[X]=\frac{2k_2{}^2(k_1+k_2-2)}{k_1(k_2-2)(k_2-4)}$, $k_2>4$;

Асимметрия:

$$A_x=\frac{(2k_1+k_2-2)\sqrt{8(k_2-4)}}{k_1(k_2-2)^2(k_2-4)},\ k_2>6;$$

Эксцесс: $\varepsilon_x=\frac{3(k_2-6)(2+\frac{1}{2}A_x^2)}{k_2-8}$, $k_2>8.$

Функция распределения F - распределения определяется зависимостью:

$$F(x)=\frac{\left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}}{B\left(\frac{k_1}{2},\frac{k_2}{2}\right)}\int_0^x u^{\frac{k_1}{2}-1}\left(1-\frac{k_1}{k_2}u\right)^{-\frac{k_1+k_2}{2}}du$$

.

F – распределение в статистике используется для проверки *гипотез статистических* в *дискриминантном*, регрессионном и *дисперсионном* анализе; в *многомерном статистическом анализе*. При возрастании чисел k_1 и k_2 F – распределение приближается к нормальному.

РАСПРЕДЕЛЕНИЕ ХОТЕЛЛИНГА (T^2 - РАСПРЕДЕЛЕНИЕ)

распределение *случайной величины непрерывной* X с плотностью вероятности:

$$f(x)=\frac{\tilde{A}(\frac{n+1}{2})x^{\frac{k}{2}-1}(1+\frac{x}{n})^{-\frac{n+1}{2}}}{\tilde{A}(\frac{n-k+1}{2})\tilde{A}(\frac{k}{2})n^{\frac{k}{2}}},$$

где n,k – целочисленные параметры распределения $(n\geq k\geq 1)$, называются степенями свободы; $\Gamma(n)$ – гамма-функция Эйлера.

При $k=1$ Р.Х. сводится к *распределению Стьюдента*, а при любом $k>1$ может рассматриваться как обобщение распределения Стьюдента на многомерный случай. Если k – мерный случайный вектор Y имеет нормальное

распределение с ненулевым средним, то соответствующее распределение называется нецентральным р.Х. с n-степенями свободы и параметром нецентральности ν.

Р.Х. используется в статистике в той же ситуации, что и *t*-распределение Стьюдента, но только в многомерном случае. Если результаты наблюдений $X_1,...,X_n$ – независимые нормально распределенные случайные векторы с вектором средних ν и невырожденной *ковариационной матрицей* Σ, то статистика

$$T^2 - n(\overline{X} - \nu)^T S^{-1}(\overline{X} - \nu),$$

где $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

и $S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T$,

имеет T^2 – распределение с *n-1* степенями свободы. Этот факт положен в основу критерия Хотеллинга. Для численных расчётов используют табл. *бета-распределения* или *F-распределения Фишера*, поскольку случайная величина

$$\frac{n-k+1}{nk}T^2$$

имеет *F* – распределение с *k* и *n* − *k* + *1* степенями свободы.

РАСПРЕДЕЛЕНИЕ ЭКСПОНЕНЦИАЛЬНОЕ (ПОКАЗАТЕЛЬНОЕ)

распределение *случайной величины непрерывной* X с плотностью вероятности

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0; \\ 0 \quad , & x < 0, \end{cases}$$

где λ – параметр распределения.

Функция распределения определяется зависимостью:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0; \\ 0 \quad\quad , & x < 0. \end{cases}$$

Графики функции плотности и функции распределения имеют вид (см. рис. 1).

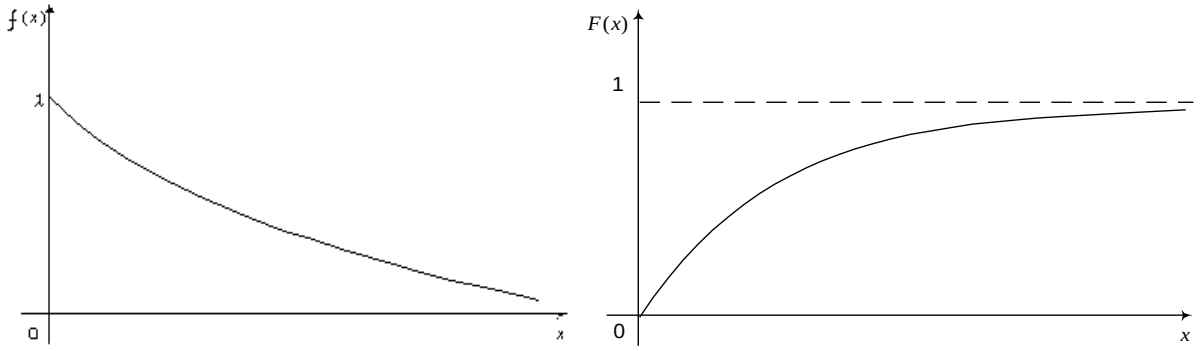


Рис.1. **Графики функции плотности и функции распределения**

Числовые характеристики:

Среднее: $M[x] = \frac{1}{\lambda}$;

Мода: $M_0 = 0$;

Дисперсия: $D[x] = \frac{1}{\lambda^2}$;

Асимметрия: $A_x = 2$;

Эксцесс: $\varepsilon_x = 6$.

Р.э. широко используется в теории массового обслуживания; теории надёжности, при изучении сроков службы различных устройств; описании времени безотказной работы отдельных элементов и системы в целом (случайных промежутков времени между появлениями двух последовательных редких событий простейшего потока). Р.э. обладает замечательным свойством, таким что вероятность безотказной работы элемента на временном интервале $(t, t+\Delta t)$ не зависит от времени предшествующей работы t, а зависит только от длины интервала Δt, т.е. имеет место свойство независимости процесса от его предыстории. Статистическая оценка параметра распределения:

$\tilde{\lambda} = \frac{1}{\bar{x}}$, где $\bar{x} = \frac{1}{n}\sum x_i$.

С

СЛОЖЕНИЯ ВЕРОЯТНОСТЕЙ ТЕОРЕМЫ

см. в ст. Теоремы сложения вероятностей

СЛУЦКОГО ТЕОРЕМА

см. в ст. Теорема Слуцкого

СЛУЧАЙНАЯ ВЕЛИЧИНА

величина, принимающая одно из возможных значений случайно, в зависимости от исхода случайного эксперимента или испытания. Событие, состоящее в том, что С.в. принимает какое-то конкретное значение или попадает в подмножество возможных значений, является *случайным событием*. Следовательно, С.в. можно задавать *законом распределения вероятностей*. С.в. по существу тесно связана с понятием ген. совокупности признаков и с вероятностным пространством (пространством элементарных событий, сигма алгеброй событий и вероятностной мерой).

С.в. (или признак) может иметь в качестве возможных значений числовые значения, ранговые значения или названия классов. С.в., являясь функцией, определённой на пространстве элементарных событий, может быть *дискретной* или *непрерывной* в соответствии с областью определения.

Дискретную случайную величину можно задать рядом распределения вероятностей, а непрерывную – плотностью распределения. На практике иногда достаточно иметь информацию о некоторых числовых характеристиках С.в.

С.в. могут быть одномерными, двумерными или многомерными. В качестве примера приведём дискретную случайную величину – количество очков при бросании двух игральных костей, заданную в виде ряда распределения вероятностей:

X=	x	2	3	4	5	6	7	8	9	10	11	12
	p	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Полученный ряд распределения вероятностей является симметричным относительно срединного значения – наиболее вероятного числа очков, равного семи. Отметим, что сумма вероятностей всех значений С.в. равна единице.

Наиболее общий способ задания любой числовой С.в. – построение функции распределения. Для этого все действительные С.в. считаются заданными в интервале (- ∞, + ∞); значения С.в. X не соответствующие элементарным событиям случайного эксперимента, трактуются как невозможные события и им приписывается вероятность 0. Функция распределения (кумулятивная, интегральная функция) задается формулой:

$$\Phi(x) = P(-\infty < X < x) = \begin{cases} \sum_{-\infty}^{x} p_i & (X - \text{дискретная}), \\ \int_{-\infty}^{x} p(y)dy & (X - \text{непрерывная}), \end{cases}$$

где суммирование ведется по всем i, для которых $x_i < x$.

Часто характер изменения С.в. определяется не путём задания её функции распределения вероятности, а каким-либо другим способом. Любая характеристика, получаемая таким способом, называется законом распределения С.в. При этом из задания этого закона распределения С.в. конкретными указаниями можно получить общую характеристику поведения С.в., т.е. её функцию распределения. Для построения функции распределения неколичественных признаков применяются числовые метки. Математически строгое, формально-абстрактное понятие С.в. формируется на основе тройки (пространство элементарных событий, сигма-алгебра, борелевских множеств, вероятностная мера).

СЛУЧАЙНАЯ ВЕЛИЧИНА ДИСКРЕТНАЯ

случайная величина, множество возможных значений которой является дискретным, т.е. конечным или счётным. С.в.д., кроме функции распределения может быть задана парой (x_i, p_i), в которой первый символ x_i обозначает принимаемое случайной величиной X значение, являющееся случайным событием $X = x_i$; второй символ p_i обозначает вероятность этого события, то есть $p_i = P(X = x_i)$. При этом I = 1, 2, …, n, если n конечное число, или I = 1, 2, … – бесконечное число. Последовательность этих пар (или табл., состоящая из двух строк) образует ряд распределения. Графическое изображение С.в.д. – *полигон* (многоугольник) распределения, представляющий собой перпендикуляры длиной p_i восстановленные в соответствующих точках x_i числовой оси *Ox*. При этом соседние вершины перпендикуляров соединяются отрезками. Сумма длин перпендикуляров (конечное или бесконечное) равна единице.

СЛУЧАЙНАЯ ВЕЛИЧИНА КОЛИЧЕСТВЕННАЯ

случайная величина, измеряемая в интервальной шкале, которая позволяет отличать, на сколько степень свойства, выражаемого числом на одном объекте больше или меньше соответствующего числа на другом. Осн. характеристики С.в.к. – средняя, *медиана*, *мода* и *дисперсия.*

Пример С.в.к., измеряемой в шкале интервалов – температура. Другим типом С.в.к. является переменная, измеряющая во сколько раз одно числовое значение степени свойства больше или меньше другого. Такой тип С.в.к. измеряется в шкале отношений. Пример С.в.к., измеряемой в шкале отношений – площадь, масса.

Следовательно, множество значений С.в.к. в зависимости от единиц измерения может быть множеством действительных чисел либо любым его подмножеством.

Шкала интервалов допускает преобразования $y = ax+в$, $a \neq 0$, $в \neq 0$. Шкала отношений допускает преобразования $y = ax$, $a > 0$.

СЛУЧАЙНАЯ ВЕЛИЧИНА МНОГОМЕРНАЯ (ВЕКТОРНАЯ)

p-мерный вектор, имеющий p случайных компонент: X_1, X_2, …, X_p. Возможными значениями С.в.м. являются точки p-мерного *евклидова пространства* (плоскости, физического пространства или гиперпространства). Для наглядности изучаемого понятия рассмотрим двумерную С.в.м., положив p=2, обозначив компоненты точки (вектора) на плоскости через X и Y. В случае *случайной величины дискретной* отметим точки на плоскости Oxy, для которых указана положительная вероятность попадания в них двумерной С.в.м.: $P(X = x_i, Y = y_j) = p_{ij}$ – здесь случайным событием является событие изображаемое двумя равенствами в скобках. Остальные точки, непопадающие в указанную область имеют вероятность, равную нулю. Из двумерного распределения можно получить одномерное распределение компоненты X:

$$(X = x_i, p_i = \sum_j p_{ij} = p_{i*}) \text{ и компоненты Y } (y_j, p_j = p_{*j}).$$

$$\text{При этом } \sum_i \sum_j p_{ij} = \sum_i p_{i*} = \sum_j p_{*j} = 1.$$

Рассмотрим двумерный случайный вектор, который называется случайной величиной непрерывной, т.к. вероятность попадания её значения в любую область, имеющую пл. D на плоскости (x, y) записывается в виде двойного интеграла:

$$P\big((X,Y) \in D\big) = \iint_D P\,(x,y)dxdy,$$

где p (x, y) – заданная неотрицательная функция, нормированная условием:

$$\iint_{-\infty}^{\infty} p(x,y)dxdy = 1.$$

Интегрирование распространяется на всю плоскость, при этом считается, что для точек, непринадлежащих заданной области возможных значений (X, У), плотность вероятности p (x, y) равна нулю. Часто для случайного вектора достаточно знать его *математическое ожидание* и *ковариационную матрицу*. Математическое ожидание случайного вектора равно вектору математических ожиданий соответствующих компонент. Далее определяется ковариационная матрица – квадратная матрица p-го порядка, являющаяся симметричной матрицей, элементы которой определяются как математические ожидания произведений отклонений случайных компонент от своих математических ожиданий.

Если X и Y независимые случайные величины,

то $P(X < x, Y < y) = P(X < x) \cdot P(Y < y)$.

Коэффициент ковариации между X и Y равен 0.

СЛУЧАЙНАЯ ВЕЛИЧИНА НЕПРЕРЫВНАЯ

случайная величина, множество возможных значений которой совпадает со множеством всех точек отрезка действительной числовой прямой и может быть задана плотностью распределения вероятностей, т.е. функцией y = p(x).

Плотность распределения есть функция действительной переменной *x*, удовлетворяющей свойствам:

$$p(x) \geq 0; \quad \int_{-\infty}^{\infty} p(x)dx = 1;$$

$$P(x_1 \leq X < x_2) = \int_{x_1}^{x_2} p(x)dx;$$

т.е. при графическом изображении *плотности распределения вероятностей* случайной величины X дифференциальная кривая y = p(x) лежит не ниже оси Ox. Пл. под всей кривой равна 1, а вероятность случайной величины X попасть на отрезок от x_1 до x_2 равна площади криволинейной трапеции, ограниченной отрезком, двумя перпендикулярами к оси Ox, восстановленными в точках x_1 и x_2, и частью кривой y = p(x), заключенной между вершинами перпендикуляров. Кроме того, $p(x) = F^1(x)$ (плотность вероятностей равна производной функции распределения F(x)). Плотность распределения и соответственно случайную величину можно доопределить положив p(x) = 0 для любого x не являющегося возможным значением X.

СЛУЧАЙНАЯ ВЕЛИЧИНА НОМИНАЛЬНАЯ

случайная величина, которая имеет конечное множество возможных значений. Номинальный признак подразделяется на категории (категоризованная С.в.н.), изучается с помощью *табл. сопряжённости* (одномерных, двумерных, трёхмерных и т.д.). Осн. задача – изучение связи, взаимозависимости различных типов. С.в.н. одномерная задается в виде ряда распределения, в котором в качестве значений (вариантов, уровней, градаций, категорий) выступают названия классов. Напр., ряд распределения (одномерная табл. сопряженности) имеет вид:

X =

X	за	против	воздержавшиеся
p	0,70	0,25	0,05

Здесь X – гипотетические результаты голосования выборов кандидата на должность.

В качестве характеристики С.в.н. можно указать *моду*. Кроме категоризованной С.в.н. существуют некатегоризованные С.в.н., для которых категории или правила отнесения к категории не существуют. Примеры С.в.н. некатегоризованных – имя, отчество, фамилия, место рождения.

СЛУЧАЙНАЯ ВЕЛИЧИНА ОДНОМЕРНАЯ (СКАЛЯРНАЯ)

случайная величина, содержащая один единственный признак; может быть *дискретной, количественной, непрерывной, номинальной, ординальной*. С.в.о. имеет одномерную функцию распределения, ряд распределения или плотность вероятности. В результате испытания она принимает одно значение, реализацию. Осн. харак-

теристиками С.в.о. являются *математическое ожидание* и *дисперсия.*

СЛУЧАЙНАЯ ВЕЛИЧИНА ОРДИНАЛЬНАЯ (ПОРЯДКОВАЯ)

случайная величина, измеряемая в ординальной (ранговой, порядковой) шкале. В отличие от случайной величины *номинальной* значения С.в.о. – классы, категории, градации упорядочены (расположены в порядке убывания или возрастания, рассматриваемого свойства). Упорядоченные классы записываются в виде рангов, обозначаемых натуральными числами: 1, 2, п. При этом случайная величина конечна с общим числом рангов, равном п. Для этой случайной величины можно указать в качестве характеристик моду и медиану. Примером двумерной С.в.о. служит случайная величина, заданная гипотетической *табл. сопряжённости* (см. табл.):

Таблица

A \ B	1	2	3	Итого:
1	0,20	0,20	0,10	0,50
2	0,10	0,10	0,05	0,30
3	0,05	0,00	0,15	0,20
Итого:	0,35	0,35	0,30	1,00

Случайная величина А принимает значения (оплата труда): высокая, средняя, низкая.

Случайная величина В принимает значения (учебная нагрузка): низкая, средняя, высокая.

Внутри табл. указаны вероятности преподавателя, случайно взятого из ген. совокупности, иметь соответствующую категорию по признаку А и признаку В, обозначаемую через Pij, напр., P23 означает, что преподаватель имеет среднюю оплату труда при высокой учебной нагрузке; в итоговом столбце вероятности значения признака А обозначаются через Pi*, напр., P1* = 0,50 – вероятность того, что случайно взятый преподаватель обладает высокой оплатой труда; аналогично

P_{2*} = 0,35 –

вероятность того, что случайно взятый преподаватель имеет среднюю учебную нагрузку.

СЛУЧАЙНОЕ БЛУЖДАНИЕ (ПРОЦЕСС СЛУЧАЙНОГО БЛУЖДАНИЯ)

(от англ. – Random walk) – часто встречающийся нестационарный тип поведения временного ряда. Напр., часто находят, что цены таких активов, как курсы акций, обменные курсы, цена на золото подвержены случайному блужданию. Различают обычно два типа моделей: С.б. без дрейфа (т.е. модель без свободного члена) и С.б. с дрейфом (т.е. при наличии в модели свободного члена).

Говорят, что временной ряд Y_t описывается моделью С.б. без дрейфа, если

$$Y_t = Y_{t-1} + u_t \quad (1),$$

где u_t – *белый шум*, т.е. *случайная величина* с *математическим ожиданием* ноль и постоянной *дисперсией* σ^2. Т.о., величина Y в момент t равна её значению в момент t-1 плюс случайный скачок u_t, следовательно, это модель авторегрессии первого порядка АР(1). Сторонники гипотезы эффективного рынка капиталов считают, что курсы акций имеют С.б. и, поэтому, невозможно зарабатывать большие деньги спекуляциями на фондовом рынке. Если бы можно было сегодня прогнозировать завтрашние курсы, то могли бы выигрывать все участники рынка. Последовательно раскрывая по формуле (1) Y_{t-1}, Y_{t-2} и т.д., нетрудно получить

$$Y_t = Y_0 + \sum u_t \quad (2),$$

где Y_0 – начальное значение Y, с которого начинается движение этого показателя. Отсюда математическое ожидание

$$E(Y_t) = E(Y_0 + \sum u_t) = Y_0 \quad (3),$$

а дисперсия

$D(Y_t) = t\sigma^2$ (4).

Другими словами, математическое ожидание Y_t равно начальному значению, а дисперсия возрастает от времени бесконечно, что нарушает условие стационарности. Т.о., С.б. без дрейфа – нестационарный стохастический процесс. Из (2) видно, что эффект каждого скачка на движение Y_t сохраняется навсегда, что дает основание утверждать (Kerry Patterson), что случайное блуждание обладает бесконечной памятью. Целесообразно отметить, что если (1) переписать как

$(Y_t - Y_{t-1}) = \Delta Y_t = u_t$(4),

где Δ – разностный оператор, то можно сделать вывод о том, что первые разности нестационарного процесса Y_t являются стационарными.

Говорят, что временной ряд Y_t описывается моделью С.б. с дрейфом, если

$Y_t = a + Y_{t-1} + u_t$ (5),

где a – константа, называемая параметром дрейфа. Употребление термина «дрейф» здесь объясняется тем, что (5) можно переписать как

$(Y_t - Y_{t-1}) = \Delta Y_t = a + u_t$(6).

Выражение (6) показывает, что Y_t дрейфует вверх или вниз в зависимости от знака параметра a. Модель (5) также является *авторегрессией* первого порядка АР(1). Нетрудно показать, что

$E(Y_t) = Y_0 + ta$ (7),

$D(Y_t) = t\sigma^2$(8).

Т.о., у этого процесса и математическое ожидание и дисперсия переменны, т.е. он также – нестационарный стохастический процесс.

Модель С.б. – пример процесса с единичным корнем.

См. также *Модель авторегрессии со скользящим средним в остатках*, *Модель авторегрессии - проинтегрированного скользящего среднего в остатках* (модель Бокса-Дженкинса).

СЛУЧАЙНОЕ СОБЫТИЕ

событие, которое в результате опыта может произойти, а может и не произойти.

Теория вероятностей, как и всякая другая наука, базируется на ряде осн. понятий. При помощи этих понятий дается логическое определение последующих более сложных понятий. Одно из осн. понятий, которым оперирует теория вероятностей – событие**.** Событием (иначе составным событием) называется некоторое подмножество пространства элементарных событий. Говорят, что событие наступило в результате испытания, если наступило одно из элементарных событий, входящих в данное событие.

Примерами С.с. могут служить: выпадение определённого количества очков при бросании игральной кости (опыт – бросание игральной кости, событие – выпадение определённого числа очков); выпадение двух гербов при троекратном бросании монеты (опыт – троекратное бросание монеты, событие – выпадение двух гербов) и т.д.

Различные события отличаются между собой по степени возможности их появления и по характеру взаимосвязи. Для правильного изучения закономерностей, присущих событиям, эти события принято классифицировать.

События называются эквивалентными (символ эквивалентности =), если они состоят из одних и тех же элементарных событий. Эквивалентные события наступают или не наступают одновременно. Событие называется невозможным, если оно не содержит ни одного элементарного события, иначе – это пустое подмножество пространства элементарных событий. Невозможное событие никогда не происходит. Событие называется достоверным, если оно содержит все элементарные события пространства Ω, иначе – если оно совпадает с самим пространством элементарных событий.

С.с. занимает промежуточное положение между достоверным и невозможным событием.

Два или несколько С.с. называются равновозможными, если условия их появления одинаковы и нет оснований утверждать, что какое-либо из них в результате опыта имеет больше шансов появиться, чем другое. Напр., выпадение любого количества очков от единицы до шести при бросании игральной кости; выпадение герба или цифры при подбрасывании монеты и т.д.

С.с. A и B называются совместными, если появление одного из них не исключает появление другого. Напр., бросаются две игральные кости. Событие A – выпадение 3 очков на первой игральной кости, событие B – выпадение 2 очков на второй игральной кости. A и B – совместные события.

С.с. $A_1, A_2, \ldots, A_n$

называются совместными, если совместны хотя бы два события из одной группы. Напр., производятся три выстрела по мишени.

A_1 – попадание в мишень при первом выстреле,

A_2 – попадание при втором выстреле,

A_3 – попадание при третьем выстреле.

A_1, A_2, A_3 образуют группу С.с.

С.с. A и B называются несовместными, если появление одного из них исключает появление другого. Напр., в магазин поступила партия товара одной номенклатуры, но разного цвета. Событие A – наудачу взятая коробка с товаром чёрного цвета, событие B – коробка с товаром коричневого цвета. A и B – несовместные события.

С.с. $A_1, A_2, \ldots, A_n$

называются несовместными, если события, входящие в группу попарно, несовместны. Напр., производится выстрел по мишени.

A_1 – попадание в десятку,

A_2 – попадание в восьмерку,

A_3 – попадание в шестерку,

A_4 – попадание в четверку,

A_5 – попадание в двойку,

A_6 – промах. A_1, A_2, A_3, A_4, A_5, A_6

образуют группу несовместных событий.

С.с. образуют полную группу, если в результате испытания обязательно наступит одно из них (единственно возможные события). Напр., вынимание одного шара из урны, в которой 2 белых и 3 черных шара. A – появление белого шара, B – появление чёрного шара. События A, B образуют полную группу событий. На практике часто рассматриваются два несовместных события, образующих полную группу. Такие события называются противоположными. Событие, противоположное событию A, принято обозначать $\overline{A}$. Напр., искажение A и $\overline{A}$ – не искажение к.-л. знака при телеграфной передаче, попадание B и промах $\overline{B}$ – при одном выстреле по цели.

События, которые одновременно являются несовместными, единственно возможными и равновозможными, называются случаями или шансами.

СЛУЧАЙНЫЕ ЧИСЛА (ТАБЛИЦА)

ряды чисел, являющихся реализациями последовательности взаимно независимых и одинаково распределённых *случайных величин.* Если в основе лежит любое распределение F, то тогда имеются F-распределенные С.ч. Если случайные величины имеют дискретное равномерное распределение на множестве чисел 0, 1, 2..... 9, то говорят о равномерно распределённых случайных цифрах.

Последовательности равномерно распределённых С.ч. получают либо с помощью физических генераторов (подбрасывание кубиков с цифрами; вытягивание из урны карточек с цифрами; преобразование случайных сигналов и др. физико-технические процессы), либо с помощью программных генераторов (аналитическим методом с помощью алгоритмов, обычно созданных как соответствующие программы для ЭВМ).

В действительности пользуются не равномерно распределённой случайной величиной R, возможные значения которой, вообще говоря, имеют бесконечное число десятичных знаков, а квазиравномерной случайной величиной R*, возможные значения которой имеют конечное число знаков. В результате замены R на R* разыгрываемая величина имеет не точно, а приближённо заданное распределение. Т.о. случайные величины, полученные на ЭВМ, называют псевдослучайными.

Последовательность псевдослучайных чисел носит детерминированный характер, но в определённых границах она удовлетворяет свойствам равномерного распределения и свойству случайности. Таблицы случайных чисел со-

ставляются одним из указанных выше способов. Достоинство программного способа получения – воспроизводимость последовательности, что даёт большие преимущества при расчётах на ЭВМ, т.к. не приходится загружать оперативную память машины таблицей случайных чисел.

Путём преобразования равномерно распределённых С.ч., выработанных с помощью генераторов, получают последовательности С.ч. с другими законами распределений. Способы преобразования: метод обратных функций, отбора, суперпозиций и специальные моделирующие алгоритмы.

С.ч. используются при выборочных обследованиях, имитационном моделировании. На применении случайных чисел основан метод статистических испытаний – *метод Монте - Карло.*

СЛУЧАЙНЫЙ ВЕКТОР

упорядоченный набор *случайных величин*; другое название – многомерная случайная величина.

Во многих приложениях статистики исследователь измеряет К > 1 признаков единиц ген. совокупности, каждый из которых является случайной величиной. Полезно рассматривать упорядоченный набор этих случайных величин как С.в. Элементы С.в. наз. компонентами. Распределение С.в. задается многомерной функцией распределения, которая в случае независимых компонент С.в. равна произведению одномерных функций распределения компонент, называемых частными, или маргинальными. Осн. параметры распределения С.в. – вектор средних и матрица коэффициентов ковариаций компонент. Выборка из С.в. образует случайную (стохастическую) матрицу.

Примеры многомерных случайных величин.

1. Успеваемость выпускника вуза характеризуется системой п случайных величин

$\xi_1, \xi_2, \ldots, \xi_n$

– оценками по различным дисциплинам, проставленными в приложении к диплому.

2. Погода в данном месте в определённое время суток характеризуется системой случайных величин: ξ_1 – температура; ξ_2 – влажность; ξ_3 – давление; ξ_4 – скорость ветра и т.п.

Случайные величины $\xi_1, \ldots, \xi_n$, входящие в систему, могут быть как дискретными, так и непрерывными.

Геометрически двумерную (ξ_1, ξ_2) и трёхмерную (ξ_1, ξ_2, ξ_3) случайные величины можно изобразить случайной точкой или С.в. плоскости Oxy или трёхмерного пространства Oxyz; при этом случайные величины ξ_1, ξ_2 или ξ_1, ξ_2, ξ_3 – составляющие этих векторов. В случае п-мерного пространства (п > 3) также говорят о случайной точке или С.в. этого пространства, хотя геометрическая интерпретация в этом случае теряет свою наглядность.

СЛУЧАЙНЫЙ ПРОЦЕСС (СЛУЧАЙНАЯ ФУНКЦИЯ)

семейство *случайных величин*, индексированных некоторым параметром, чаще всего играющим роль времени или пространства. В терминах *теории вероятностей* случайной называется функция, которая в результате опыта может принять тот или иной конкретный вид, причём неизвестно заранее – какой именно.

Конкретный вид, принимаемый случайной функцией в результате опыта, называется реализацией случайной функции. Если произвести серию опытов, то мы получим группу ("семейство") реализаций этой функции. Рассмотрим некоторую случайную функцию $X(t)$. Предположим, что над ней произведено n независимых опытов, в результате которых получено n реализаций случайной функции. Обозначим их соответственно номеру опыта

$x_1(t), \ldots, x_n(t)$.

Каждая реализация – обычная (неслучайная) функция. Т.о., в результате каждого опыта случайная функция $X(t)$ превращается в обычную, неслучайную функцию. Зафиксируем теперь некоторое значение аргумента t и посмотрим, во что превратится при этом случайная функция $X(t)$. Очевидно, она превратится в случайную величину в обычном смысле слова. Называют эту случайную величину сечением случайной функции, соответствующим

данному t. Т.о., если Ω – множество элементарных событий и t – непрерывный параметр, то С.п. называется функция двух аргументов

$$X(t)=\varphi(\omega,t),\ \omega\in\Omega.$$

Для каждого значения параметра t функция $\varphi(\omega,t)$ является функцией только ω и, следовательно, представляет собой случайную величину. Для каждого фиксированного значения аргумента ω (т.е. для каждого элементарного события) $\varphi(\omega,t)$ зависит только от t и является попросту функцией одного аргумента. На случайный процесс можно смотреть либо как на совокупность случайных величин $X(t)$, зависящих от параметра t, либо как на совокупность реализаций функции $X(t)$.

Отметим, что функция $\varphi(\omega,t)$ не является полной, исчерпывающей характеристикой случайной функции $X(t)$. Действительно, эта функция характеризует только закон распределения $X(t)$ для данного, хотя и произвольного t; она не отвечает на вопрос о зависимости случайных величин $X(t)$ при различных t. С этой точки зрения более полной характеристикой случайной функции $X(t)$ является двумерный закон распределения:

$p(t,x;\tau,y)$ – закон распределения системы двух случайных величин

$X(t),X(\tau)$, т.е. двух произвольных сечений случайной функции $X(t)$.

Числовые характеристики случайной функции.

Математическим ожиданием случайной функции $X(t)$ называется неслучайная функция $m_X(t)$, которая при каждом значении аргумента t равна математическому ожиданию соответствующего сечения случайной функции:

$$m_X(t)=M\left[X(t)\right].$$

Дисперсией $D_X(t)$ случайной функции $X(t)$ называется неслучайная функция, значение которой для каждого t равно дисперсии соответствующего сечения случайной функции:

$$D_X(t)=D\left[X(t)\right].$$

Отметим, что случайные функции с различной внутренней структурой могут иметь одинаковые математические ожидания и дисперсии. Для описания различий вводят специальную характеристику – ковариационную функцию. Ковариационной функцией случайной функции $X(t)$ называется неслучайная функция двух аргументов $K_X(t,t')$, которая при каждой паре значений (t,t') равна ковариационному моменту соответствующих сечений случайной функции:

$$K_X(t,t')=M\left[\left(X(t)-m_X(t)\right)\left(X(t')-m_X(t')\right)\right].$$

При $t=t'$ ковариационная функция совпадает с дисперсией случайной функции:

$$K_X(t,t)=D_X(t).$$

Ковариационная функция симметрична относительно своих аргументов:

$$K_X(t,t')=K_X(t',t).$$

Вместо ковариационной функции

$K_X(t,t')$

можно пользоваться корреляционной функцией:

$$\rho_X(t,t')=\frac{K_X(t,t')}{\sigma_X(t)\sigma_X(t')},$$

которая является *коэффициентом корреляции* величин

$X(t),\ X(t')$.

Для характеристик случайных функций справедливы все теоремы, аналогичные теоремам для характеристик случайных величин, при этом константы заменяются неслучайными функциями. Определим некоторые свойства С.п.: а) процесс называется стационарным, если для любой группы из конечного числа непересекающихся промежутков времени вероятность наступления определённого события на протяжении каждого из них зависит только от этих событий и от длительности промежутков времени, но не изменяется от сдвига всех временных отрезков на одну и ту же величину; б) процесс называется процессом без последействия (или процессом Марковского типа), если вероятность наступления некоторого события в течение промежутка времени $(T,T+t)$ не зависит от того, какое событие и как оно появлялось ранее. Это означает, что условная вероятность появления некоторого события за промежуток $(T,T+t)$ при любом предположении о наступлении событий до момента T совпадает

с безусловной вероятностью; в) процесс называется ординарным, если вероятность появления более одного раза за малый промежуток времени Δt некоторого элементарного события бесконечно мала по сравнению с Δt, т.е.

$$P_{>1}(\Delta t) = o(t).$$

Предположим, что в случайные моменты времени происходит некоторое событие. Т.о., мы имеем некоторый процесс появления события. Будем считать, что процесс обладает свойством стационарности, ординарности и не имеет последействия. Такой процесс называется простейшим (стационарным пуассоновским) потоком.

Отметим, что процесс без последействия полностью описывается функцией $p(t, x; \tau, y)$, при этом если рассматривать интегральную функцию распределения $F(t, x; \tau, y)$, то она должна удовлетворять условиям:

1. $\lim\limits_{y \to -\infty} F(t, x; \tau, y) = 0$,
   $\lim\limits_{y \to +\infty} F(t, x; \tau, y) = 1$;
2. Функция $F(t, x; \tau, y)$ непрерывна слева относительно аргумента y.

Кроме того, для функции $F(t, x; \tau, y)$ справедливо следующее равенство:

$$F(t, x; \tau, y) = \int_{-\infty}^{+\infty} F(s, z; \tau, y) dz F(t, x; s, z),$$

которое называют обобщённым уравнением Маркова, так как оно представляет собой распространение равенства перехода теории *Марковских цепей* на теорию С.п.

СРЕДНЕКВАДРАТИЧЕСКОЕ ОТКЛОНЕНИЕ

см в ст. Отклонение среднеквадратическое.

СТАЦИОНАРНОСТЬ В УЗКОМ СМЫСЛЕ

случайный процесс X(t) называется стационарным (случайный процесс, вероятностные характеристики которого не меняются с течением времени) в узком смысле, если его *n*-мерная плотность распределения не изменяется при сдвиге всех его аргументов на одинаковую произвольную величину τ. Формально условие С. в у.с. можно записать:

$$f_n(t_1, t_2, \dots, t_n; x_1, x_2, \dots, x_n) = f_n(t_1 + \tau, t_2 + \tau, \dots, t_n + \tau; x_1, x_2, \dots, x_n),$$

где f_n – плотность распределения стационарного случайного процесса.

Из С. в у.с. следует *стационарность в широком смысле*. Обратное верно только для нормальных процессов.

СТАЦИОНАРНОСТЬ В ШИРОКОМ СМЫСЛЕ

случайные процессы $X(t)$, имеющие постоянное среднее значение и корреляционную функцию, зависящую только от разности моментов (t_2-t_1), называют стационарным случайным процессом в широком смысле (а более частные случайные процессы, все характеристики которых не меняются с течением времени, в таком случае называются стационарным случайным процессом в узком смысле). Это понятие впервые ввёл Хинчин А.Я.

Обозначим

$\tau=t_2-t_1$, тогда $f_2(t_1, t_2, x_1, x_2) = f_2(\tau, x_1, x_2)$.

С. в ш.с. означает: $m_x(t) = \text{const}$ для всех t;

$Kx(t_1, t_2) = Kx(t_2-t_1)$,

т.е. автокорреляционная функция случайного процесса зависит только от разности t_2 и t_1.

Свойства корреляционной функции стационарного процесса:

$Kx(\tau) = Kx(-\tau); Kx(0) = Dx; |Kx(\tau)| \le Dx.$

Из *стационарности в узком смысле* следует С. в ш.с. Обратное верно только для нормальных процессов.

СТАЦИОНАРНЫЙ СЛУЧАЙНЫЙ ПРОЦЕСС

случайный процесс, вероятностные характеристики которого не меняются с течением времени.

В математической теории С.с.п. осн. роль играют моменты распределении вероятностей значений процесса $X(t)$, являющиеся простейшими числовыми характеристиками этих распределений. Особенно важны моменты первых двух порядков: среднее значение стационарных случайных процессов $MX(t) = m$ – математическое ожидание случайной величины $X(t)$ и корреляционная функция стационарных случайных процессов

$M(X(t_1)\, X(t_2)) = B(t_2-t_1)$

– математическое ожидание произведения $X(t_1)X(t_2)$ (выражающееся через дисперсию величин $X(t)$ и коэффициент корреляции между $X(t_1)$ и $X(t_2)$).

У С.с.п. $X(t)$ все вероятностные характеристики не должны зависеть от времени. Пусть $f(t,x)$ – одномерная плотность распределения С.с.п. Так как эта плотность не зависит от того, где взято сечение t, то имеет место равенство

$f(t_1,x) = f(t_2,x) = \ldots = f(x)$.

Зная одномерную плотность С.с.п. $X(t)$, можно найти его математическое ожидание и дисперсию:

$$M[X(t)] = \int_{-\infty}^{\infty} x \cdot f(t,x)dx = \int_{-\infty}^{\infty} x \cdot f(x)dx = m_x = const$$

$$D[X(t)] = \int_{-\infty}^{\infty} (x-m_x)^2 \cdot f(t,x)dx = \int_{-\infty}^{\infty} (x-m_x)^2 \cdot f(x)dx = D_x = const$$

У С.с.п. математическое ожидание и дисперсия являются постоянными величинами, не зависящими от времени. Выделение понятия С.с.п. и получение первых относящихся к нему математических результатов являются заслугой Слуцкого Е.Е. и относятся к кон. 20-х и нач. 30-х гг. 20 в. В дальнейшем важные работы по теории С.с.п. были выполнены Хинчиным А.Я., Колмогоровым А.Н., Крамером Г., Винером Н. и др.

Примеры С.с.п. встречаются в физике (в частности, гео- и астрофизике), механике и технике.

СХОДИМОСТЬ ПО ВЕРОЯТНОСТИ

вид сходимости измеримых функций (*случайных величин*), заданных на *вероятностном пространстве*.

Последовательность случайных величин $\{\xi_n\}$ сходится по *вероятности* к случайной вели-

чине ξ при $n \to \infty$, и пишут: $\xi_n \xrightarrow{P} \xi$, если для любого $\varepsilon > 0$ $P\left(\left|\xi_n - \xi\right| \geq \varepsilon\right) \to 0$ при $n \to \infty$ (или $P\left(\left|\xi_n - \xi\right| < \varepsilon\right) \to 1$ при $n \to \infty$).

С. по. в. не обязательно сопровождается сходимостью математических ожиданий или моментов других порядков: из $\xi_n \xrightarrow{P} \xi$ не следует, что

$E\xi_n \to E\xi$.

С. по. в. обладает обычными свойствами пределов числовых последовательностей:

если $\xi_n \xrightarrow{P} \xi$ и $\eta_n \xrightarrow{P} \eta$,

то: 1. $\xi_n + \eta_n \xrightarrow{P} \xi + \eta$; 2. $\xi_n \cdot \eta_n \xrightarrow{P} \xi \cdot \eta$;

если $\xi_n \xrightarrow{P} \xi$

и $g(x)$ – непрерывная функция,

то $g(\xi_n) \xrightarrow{P} g(\xi)$;

если $\xi_n \xrightarrow{P} c$ и $g(x)$ непрерывна в точке с, то $g(\xi_n) \to g(c)$

если $\xi_n \to \xi$ «почти наверное», то $\xi_n \xrightarrow{P} \xi$.

Т

ТЕОРЕМА (ФОРМУЛА) БАЙЕСА

одна из осн. теорем *теории вероятностей*, которая определяет вероятность наступления события в условиях, когда на основе наблюдений известна лишь некоторая частичная информация о событиях. По формуле Б. можно более точно пересчитывать вероятность, беря в учёт как ранее известную информацию, так и данные новых наблюдений.

Предположим, что событие B может произойти с одним и только с одним из n попарно несовместных событий (гипотез) $H_1, \ldots, H_n$ образующих полную группу. Будем эти события называть гипотезами. Справедлива формула полной вероятности

$$p(B) = \sum_{i=1}^{n} p(B/H_i)\,p(H_i),$$

где $p(B/H_i)$ – вероятность появления события В при условии, что произошло событие H_i, а $p(H_i)$ – априорная вероятность H_i.

В тесной связи с формулой полной вероятности находится формула Б. Пусть произведён опыт, и в результате него наступило событие B

$p(B) > 0$,

при этом нам известны вероятности (априорные вероятности) гипотез H_i

($\{H_1, \ldots, H_n\}$

некоторое разбиение пространства Ω с

$p(H_i) > 0$).

Тогда по формуле Б. определяется

$p(H_i / B)$

– апостериорная вероятность гипотезы H_i:

$$p(H_i / B) = \frac{p(B/H_i)\,p(H_i)}{p(B)}.$$

Используя формулу полной вероятности записываем:

$$p(H_i / B) = \frac{p(B/H_i)\,p(H_i)}{\sum_{i=1}^{n} p(B/H_i)\,p(H_i)}$$

Пример. На некоторой фабрике 30% продукции производится машиной 1, 25% – машиной 2, остальное – машиной 3. У машины 1 в брак идёт 1% всей производимой продукции, у машины 2 – 1,5%, у машины 3 – 2%. Наугад выбранная единица продукции оказалась браком. Какова вероятность того, что она произведена машиной 1?

Введём обозначения: B – выбранное изделие брак;

H_i – изделие произведено машиной i,

$i = 1, 2, 3$.

Имеем:

$p(H_1) = 0,3$, $p(H_2) = 0,25$,

$p(H_3) = 0,45$, $p(B/H_1) = 0,01$,

$p(B/H_2) = 0,015$, $p(B/H_3) = 0,02$.

По т.(ф.)Б. находим:

$$p(H_1/B) = \frac{0,01 \cdot 0,3}{0,01 \cdot 0,3 + 0,015 \cdot 0,25 + 0,02 \cdot 0,45} = 0,2 \cdot$$

Этот пример показывает возможность пересмотра мнения о распределении вероятностей на некотором пространстве в случае получения новой информации, что нашло свое применение в *байесовском подходе к оцениванию* в *математической статистике* и теории статистических решений.

Т.(ф.)Б. можно обобщить на вероятностные меры, привлекая понятие *функции плотности распределения*:

$$h(\theta / X) = \frac{f(X/\theta)}{g(X)} h_0(\theta),$$

где $f(X/\theta)$ – плотность условного распределения случайной величины X при данном значении $\theta \in \Theta$ (в математической статистике функция правдоподобия), $h_0(\theta)$ – априорная плотность распределения случайной величины θ, $g(X) = \int f(X/\theta) h_0(\theta) d\theta$.

ТЕОРЕМА БЕРНУЛЛИ

см. в ст. Бернулли теорема

ТЕОРЕМА МУАВРА-ЛАПЛАСА

частный случай *центральной предельной теоремы*, справедливый для *Бернулли испытания*.

При рассмотрении схемы Бернулли очень часто приходится решать задачи, в которых числа n и m настолько велики, что использование формулы Бернулли затруднительно. Поэтому возникает необходимость в использовании простых приближённых формул для вероятностей в схеме Бернулли при больших n и m.

Т.М.-Л. локальная: пусть p – фиксированное число,

$0 < p < 1$.

Если в выражении для

$$P_n(m) = C_n^m p^m (1-p)^{n-m}$$

устремить n к бесконечности и при этом изменять m так, чтобы величина

$$x = \frac{m - np}{\sqrt{np(1-p)}}$$

оставалась ограниченной:

$a \le x \le b$ ($a, b = \text{const}$),

то будем иметь:

$$P_n(m) \to \frac{1}{\sqrt{np(1-p)}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Т.о., при больших значениях n имеет место приближенное равенство:

$$P_n(m) \approx \frac{1}{\sqrt{np(1-p)}} \varphi(x), \text{ где } \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}$$

– функция плотности распределения для стандартного нормального закона распределения.

Надо заметить, что точность приближённой формулы существенно зависит от n и p. Анализ показывает, что точность повышается с ростом произведения $np(1-p)$. Т.М.-Л. пользуются, когда $np(1-p) \ge 10$.

Т.М.-Л., интегральная: вероятность события

$$a \le x \le b, \; x = \frac{m - np}{\sqrt{np(1-p)}}$$

в схеме Бернулли при $n \to \infty$ имеет своим пределом выражение

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx.$$

Содержание теоремы можно записать в виде

$$\lim_{n\to\infty} P_n\left(m_a \le m \le m_b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx,$$

где $a = \frac{m_a - np}{\sqrt{np(1-p)}}$ и $b = \frac{m_b - np}{\sqrt{np(1-p)}}$.

Чтобы придать этому равенству более простую форму, введём в рассмотрение функцию

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt,$$

называемую функцией Лапласа. Т.о., при больших значениях n

$$P_n(m_a \le m \le m_b) \approx \Phi\left(\frac{m_a - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{m_b - np}{\sqrt{np(1-p)}}\right).$$

ТЕОРЕМА ПУАССОНА

одна из предельных теорем для схемы Бернулли. В случае, когда p мало, применение локальной *теоремы Муавра-Лапласа* даёт большие погрешности, т.к. произведение $np(1-p)$ мало. В этом случае лучший результат для вероятности $P_n(m)$ даёт Т.П. Предположим, что p в схеме Бернулли является функцией от n,

$p = p(n)$.

Если m фиксировано, $n \to \infty$, а $p(n) \to 0$, притом так, что величина

$\lambda = \lim_{n\to\infty} np(n)$, $\lambda > 0$,

остаётся ограниченной, то справедливо соотношение

$$P_n(m) \to \frac{\lambda^m}{m!} e^{-\lambda}.$$

Т.о., имеет место приближенная формула Пуассона

$$P_n(m) \approx \frac{\lambda^m}{m!} e^{-\lambda},$$

которая используется в тех случаях, когда p мало, n велико, а каждое из чисел m и λ не слишком велико.

ТЕОРЕМЫ СЛОЖЕНИЯ ВЕРОЯТНОСТЕЙ

теоремы, позволяющие определять вероятность суммы совместных и *несовместных событий.Теорема 1. Вероятность суммы n совместных событий равна*

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_i P(A_i) - \sum_{i,j} P(A_i \cdot A_j) + \sum_{i,j,k} P(A_i \cdot A_j \cdot A_k) - \ldots + (-1)^{n-1} P(A_1 \cdot A_2 \cdot A_3 \cdot \ldots \cdot A_n).$$

Так, напр., вероятность суммы двух совместных событий равна сумме вероятностей этих событий без вероятности их совместного появления:

$$P(A+B) = P(A) + P(B) - P(AB).$$

Для трёх совместных событий теорема записывается в виде:

$$P(A_1 + A_2 + A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_1A_3) - P(A_2A_3) + P(A_1A_2A_3).$$

Теорема 2. Вероятность суммы несовместных событий равна сумме вероятностей этих событий:

$$P(A_1 + A_2 + \ldots + A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

Следствие 1. Если события $A_1, A_2, \ldots A_n$ образуют полную группу несовместных событий, то сумма их вероятностей равна единице:

$$P(A_1) + P(A_2) + \ldots + P(A_n) = 1.$$

Следствие 2. Сумма вероятностей противоположных событий равна единице

$$P(A) + P(\overline{A}) = 1.$$

Противоположные события – частный случай полной группы несовместных событий, в которой число событий равно двум.

При решении практических задач часто оказывается проще вычислить вероятность противоположного события. В этом случае вероятность события А вычисляют, используя следствие 2:

$$P(A) = 1 - P(\overline{A}).$$

Для нескольких событий:

$$P(A_1 + A_2 + \ldots + A_n) = 1 - P(\overline{A}_1 + \overline{A}_2 + \ldots + \overline{A}_n)$$

.

ТЕОРЕМА СЛУЦКОГО

теорема, существенно расширяющая возможности применения *центральной предельной теоремы* за счёт рассмотрения сходимости сумм и произведений случайных последовательностей.

Пусть $\{X_n\}$, $\{Y_n\}$ и $\{Z_n\}$, $n = 1, 2, \ldots$

– последовательности (быть может зависимых) *случайных величин* такие, что

$X_n \xrightarrow[n\to\infty]{F} X \in F_X(x)$

(сходимость по распределению),

где $F_X(x)$ — непрерывная функция распределения,

$Y_n \xrightarrow[n\to\infty]{P} 1$ (сходимость по вероятности),

$Z_n \xrightarrow[n\to\infty]{P} 0$ (сходимость по вероятности),

тогда $X_n + Z_n \xrightarrow[n\to\infty]{F} X \in F_X(x)$;

$X_n \cdot Z_n \xrightarrow[n\to\infty]{P} 0$;

$Y_n \cdot (X_n + Z_n) \xrightarrow[n\to\infty]{F} X \in F_X(x)$.

Т.С. обосновывает многие асимптотические распределения статистик в *математической статистике*.

ТЕОРЕМЫ УМНОЖЕНИЯ ВЕРОЯТНОСТЕЙ

теоремы, позволяющие определять вероятность произведения зависимых и *независимых событий*.

Теорема 1. Вероятность произведения (совместного наступления) нескольких зависимых событий равна произведению вероятности одного из них на условные вероятности остальных событий, вычисленных в предположении, что все предыдущие события имели место:

$$P(A_1 A_2 \dots A_n) = P(A_1) P(A_2 / A_1) P(A_3 / A_1 A_2) \dots P(A_n / A_1 A_2 \dots A_{n-1}).$$

Теорема 2. Вероятность произведения независимых событий равна произведению вероятностей этих событий:

$$P(A_1 A_2 \dots A_n) = P(A_1) P(A_2) \dots P(A_n)$$

.

Справедливость этой теоремы вытекает непосредственно из определения независимых событий.

ТЕОРЕМА ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ

см. в ст. Центральная предельная теорема

ТЕОРИЯ ВЕРОЯТНОСТЕЙ

математическая наука для описания случайных явлений. Случайными называются те явления, состояния которых невозможно предсказать. В качестве примера обычно приводят реальные опыты вроде бросания игральных костей (нельзя предсказать, какой гранью вверх ляжет кубик) или эксплуатации технических изделий (нельзя предсказать, сколько часов прослужит электрическая лампочка).

В этих неопределённых ситуациях нас может интересовать *вероятность* тех или иных событий. При игре в кости нам может быть важна вероятность выпадения шести очков (шестёрки); при бросании двух костей – вероятность выпадения двух шестёрок, вероятность получить в сумме 10 очков или больше, и т.п. Для технических изделий важна, в частности, вероятность безотказной работы в течение заданного срока, и т.д.

Случайные явления – часть природы и часть нашей жизни. В естественном языке есть средства для их обсуждения. Но смысл слов естественного языка подвижен, изменчив и не имеет чётких границ. Для точной науки это – недостаток. Поэтому Т.в. начинает с установления терминологии. Начальное понятие Т.в. – случайный эксперимент. Этот термин заменяет многие выражения естественного языка. Случайный эксперимент может быть чьим-то действием, вроде бросания монеты или испытания прибора, может быть явлением природы (погода) или человеческого общества (производство товаров, их цены). В случайном эксперименте могут происходить (или не происходить) различные события. События случайного эксперимента называют *случайными событиями*. Какие именно события произошли, мы узнаём только по окончании эксперимента. Те случайные события, которые нельзя расщепить на более простые события, называют элементарными событиями (в смысле неделимыми). Каждый случайный эксперимент оканчивается к.-л. элементарным событием, и притом только одним. Напр., при бросании «математической» игральной кости может выпасть чётное число очков – это случайное событие. Оно является составным и состоит из трёх элементарных событий: выпало 2 очка, выпало 4 очка, выпало 6 очков. Электрическая лампочка может прослужить, скажем, не менее 500 часов. Элементарные события в этом случайном эксперименте образуют неотрицательную часть числовой

прямой. Это случайное событие, но не элементарное случайное событие. Упомянутые события относятся к разным случайным экспериментам.

Совокупность элементарных событий случайного эксперимента называют его пространством элементарных событий. Среди разнообразия пространств элементарных событий выделяют два практически наиболее важных типа: пространства дискретные и пространства непрерывные. Последний термин не единствен и не общепринят. В дискретном случае множество элементарных событий конечно либо счётно. Непрерывные пространства элементарных событий – прямая или её интервалы; плоскость, её области; кривые на плоскости и т.д. или схожие с этими (или ещё более сложные) математические структуры.

Случайные события – подмножества пространства элементарных событий. Событие происходит, если происходит к.-л. элементарное событие из подмножества, представляющего случайное событие.

Следует отметить, что в пространствах элементарных событий, не являющихся дискретными, не все существующие в этих случаях подмножества считаются событиями. Однако при практических применениях Т.в. эта математическая тонкость обычно не ощущается.

Каждое случайное событие имеет свою вероятность. Вероятность случайного события – неотрицательное число, не превосходящее единицы. Мы воспринимаем вероятность случайного события как его правдоподобие, как меру возможности для этого события осуществиться в случайном эксперименте.

Если вероятность события велика (близка к 1), мы ожидаем его осуществления в случайном эксперименте. Если вероятность события мала (близка к 0), то мы полагаем, что в опыте этого события не будет. Более того, на практике мы руководствуемся правилом, что если вероятность интересующего нас события мала, то в однократном эксперименте оно не произойдёт. Вопрос о том, какую вероятность считать малой, решается в каждом случае отдельно, в зависимости от последствий ошибки.

При независимых повторениях опыта частота события приближается к его вероятности. В Т.в. это один из важных результатов, известный как *закон больших чисел*. Благодаря этой связи между вероятностью и частотой можно сказать, что в ряду независимых повторений опыта маловероятное событие будет появляться редко.

Случайные события часто обозначают прописными латинскими буквами A, B, … Символ вероятности – традиционно буква P. Так что вероятности событий A, B, … – P(A), P(B), и т.д. С событиями, относящимися к одному эксперименту, можно совершать логические действия. По отношению к представляющим их подмножествам в пространстве элементарных событий это операции объединения, пересечения, дополнения и другие. При этих действиях с событиями вероятности событий – результатов определённым образом вычисляются по вероятностям участвующих событий. Так, если события A и B одновременно произойти не могут (подмножества A и B не пересекаются), то

P($A \cup B$) = P(A) + P(B).

Это одно из важнейших свойств вероятности. Событие $A \cup B$, объединение подмножеств A и B, происходит тогда, когда происходит либо событие A, либо событие B. Для пересекающихся A и B надо добавить: либо и A, и B. Приведённая выше формула называется правилом сложения вероятностей. Из этого правила вытекают все прочие правила вычисления вероятностей при операциях с событиями.

Одно из важнейших понятий Т.в. – понятие независимости событий. События A и B (относящиеся к одному случайному эксперименту) называются независимыми,

если P(A∩B) = P(A)P(B).

Событие A∩B (пересечение событий) происходит тогда, когда происходят и A, и B. Понятие независимости событий порождает в Т.в. и многие другие формы независимости (стохастической независимости): независимость случайных величин, независимость случайных экспериментов и др.

Другое важное понятие Т.в. – условная вероятность события A при условии, что произошло событие B, которую обозначают P(A|B). По определению,

$P(A|B) = P(A \cap B) / P(B)$

для событий B такого рода, что $P(B) > 0$. Из определения условных вероятностей следует, что

$P(A \cap B) = P(A|B)P(B)$.

Эту формулу называют правилом умножения вероятностей. В некоторых случаях P(A|B) имеет смысл и для событий B нулевой вероятности. Но тогда P(A|B) приходится определять более сложным образом.

Если события A и B независимы, причём

$P(A) > 0, P(B) > 0,$

то $P(A|B) = P(A), P(B|A) = P(B)$.

Эти свойства соответствуют нашему интуитивному представлению об отсутствии взаимосвязи между двумя событиями. Само понятие независимости случайных событий отражает представление об отсутствии причинной связи. Любое из равенств

$P(A|B) = P(A)$

или $P(B|A) = P(B)$

влечёт за собой взаимную независимость событий A и B.

В реальных задачах, к которым мы применяем Т.в., часто ясна взаимная независимость тех или иных событий. Это помогает правильному построению вероятностной модели.

В дискретных пространствах элементарных событий вероятность любого события складывается из вероятностей элементарных событий. А именно, для любого события его вероятность есть сумма вероятностей тех элементарных событий, которые его составляют. Поэтому в дискретных пространствах лишь вероятности элементарных событий. Достаточно знать, как полная вероятность (равная 1) распределена между элементарными событиями.

Т.в. для конечных пространств элементарных событий устроена особенно просто. Для более сложных пространств этот случай служит источником наглядных представлений и ассоциаций. Особенно простым является вариант равномерного распределения вероятностей. Если общее число элементарных событий обозначить N, то при равномерном распределении вероятность каждого элементарного события лишь $\frac{1}{N}$.

Вероятность произвольного события A при равномерном распределении пропорциональна числу N(A) элементарных событий, составляющих A. Более точно:

$P(A)=N(A)/N$.

Эта формула – т.н. классическое определение вероятности. Азартные игры, с изучения которых и начиналась Т.в., все связны именно с конечными пространствами элементарных событий и равномерными распределениями вероятностей.

Равномерное распределение вероятностей сейчас играет важную роль в приложениях Т.в. Оно служит основой выборочного метода. Выборочный метод широко применятся в статистических исследованиях в экономике, социологии и пр. и в статистическом контроле качества массовой продукции. Метод состоит в случайном выборе объектов из обследуемой совокупности. Выбор называется случайным, если каждый объект совокупности может быть выбран, и все объекты имеют равные вероятности быть выбранными. (Это означает, что вероятность равномерно распределена между объектами совокупности). Выборочный метод применяют для формирования представительных (репрезентативных) выборок.

Распределение вероятностей в непрерывных пространствах элементарных событий часто можно описать с помощью функции плотности. Пусть, для определённости, пространство элементарных событий – числовая прямая (т.е. прямая, на которой введены координаты). Наиболее важные для практики события – отрезки и интервалы прямой и их комбинации; элементарные события – точки прямой. Каждую точку можно представлять себе как некоторое число – координату этой точки. Пусть далее это x. Функцией плотности вероятности

аргумента x называют такую функцию (с числовыми значениями), скажем, f(x), что для любого события A (отрезка [a, b]) его вероятность

$$P(A) = \int_a^b f(x)dx.$$

(Из этого определения следует, что f(x) ≥ 0 и что $\int_{-\infty}^{+\infty} f(x)dx = 1.$

В прикладных задачах встречаются разнообразные плотности вероятности. Напр., при описании случайного «времени ожидания» (а также времени безотказной работы, длительности телефонного разговора и т.п.) используется показательная плотность вероятности:

$$f(x) = \frac{1}{a} e^{-\frac{x}{a}}$$

для $x \geq 0$, $f(x) = 0$ для $x < 0$. Число $a > 0$ называют параметром распределения; в конкретных условиях параметр может принимать разные значения. Наиболее известным и важным для приложений (и для математической теории) является т.н. нормальное распределение. Нормальная плотность вероятности зависит от двух параметров.

Если пространство элементарных событий представляет собой плоскость и элементарными событиями являются пары чисел (x,y), т.е. координаты точек этой плоскости, то плотность вероятности – это функция переменных (x,y), скажем, f(x,y). Типичное событие в этом случае – область на упомянутой плоскости. Вероятность события – интеграл от плотности вероятности по этой области; интеграл в данном случае двумерный. Схожим образом определяются функции плотности и в более сложных случаях.

Осн. объекты изучения Т.в. – *случайные величины* и *случайные процессы*. Важные характеристики случайных величин – их *математические ожидания* и *дисперсии*. Связь случайных явлений с законами природы устанавливают т.н. предельные теоремы Т.в., в первую очередь закон больших чисел. Вторая важная предельная закономерность Т.в. *–центральная предельная теорема*.

Математическая статистика (первоначально называвшаяся математическими методами статистики) устанавливает свойства вероятностных распределений по результатам случайных экспериментов. По своей важности и своеобразию задач и методов выделилась в отдельную науку.

У

УМНОЖЕНИЯ ВЕРОЯТНОСТЕЙ ТЕОРЕМЫ

см. в ст. Теоремы умножения вероятностей

УСЛОВНАЯ ВЕРОЯТНОСТЬ

см. в ст. Вероятность условная

УСЛОВНОЕ РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ

распределение подмножества $k_1 < k$ случайных величин из распределения k случайных величин, когда остальные $(k - k_1)$ случайные величины принимают постоянные значения. Условное распределение вероятностей также называется условным законом распределения, который можно задать как функцией распределения, так и плотностью распределения.

Условная плотность распределения определяется по формуле

$$f(x_1, x_2, ..., x_{k_1} / x_{k_1+1}, ..., x_k) = \frac{f(x_1, x_2, ..., x_k)}{f_1(x_{k_1+1}, ..., x_k)}.$$

Для системы $(X_1, X_2, ..., X_k)$ независимых случайных величин плотность распределения равна произведению плотностей распределения отдельных величин, входящих в систему:

$$f(x_1, x_2, ..., x_k) = f_1(x_1) f_2(x_2) ... f_k(x_k).$$

Законы распределения системы n зависимых случайных величин, являющиеся функциями многих аргументов, чаще всего неудобны в практическом применении и к тому же для своего определения (хотя бы приближенного) требуют очень большого объема экспериментальных данных.

Для распределения вероятностей двух случайных величин X, Y существуют:

– условные распределения X: некоторое конкретное распределение представляют как «распределение X при Y=y»;

– условные распределения Y: некоторое конкретное распределение представляют как «распределение Y при X=x».

Плотность распределения для случайной величины X при условии, что случайная величина Y приняла определенное значение, имеет вид:

$$f_1(x/y)=\frac{f(x,y)}{f_2(y)}.$$

Аналогично для случайной величины Y:

$$f_2(y/x)=\frac{f(x,y)}{f_1(x)}.$$

Плотность распределения системы можно определить по зависимости:

$$f(x,y)=f_1(x)f_2(y/x)=f_2(y)f_1(x/y).$$

Данное равенство часто называют теоремой умножения законов распределения. Условная плотность распределения обладает всеми свойствами безусловной плотности распределения.

Ф

ФУНКЦИЯ ПЛОТНОСТИ ВЕРОЯТНОСТИ

(плотность распределения, плотность вероятностей) – производная от функции распределения

$$f(x)=\frac{dF(x)}{dx}.$$

Функцию $f(x)$ называют также дифференциальной функцией распределения, она является одной из форм закона распределения *случайных величин* и существует только для непрерывных случайных величин. Свойства Ф.п.в.:

1) $f(x)\ge 0$;

2) $p(x_1\le X\le x_2)=\int_{x_1}^{x_2} f(x)dx$;

3) $F(x)=\int_{-\infty}^{x} f(x)dx$; 4) $\int_{-\infty}^{\infty} f(x)dx=1$.

ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

функция $F(x)$ *случайной величины X*, значение которой в точке *x* определяет вероятность события

$(X<x)$ $F(x)=P(X<x).$;

также называют интегральной функцией распределения. Ф.р.в. существует для *случайных величин дискретных* и *непрерывных*, она полностью характеризует случайную величину с вероятностной точки зрения и является одной из форм закона распределения вероятностей.

Ф.р.в. дискретной случайной величины разрывна, имеет вид ступенчатой функции и возрастает скачками при переходе через точки возможных её значений x_i, причём величина скачка равна вероятности соответствующего значения и её можно определить:

$$F(x)=\sum_{x<x_i} p(x=x_i).$$

Свойства Ф.р.в.: $F(x)$ неотрицательная функция, $0\le F(x)\le 1$; $F(x)$ неубывающая функция своего аргумента

$F(x_2)\ge F(x_1)$,

если $x_2>x_1$; $F(-\infty)=0$, $F(+\infty)=1$;

Вероятность появления с.в. X в интервале

$[x_1,x_2]$ $P(x_1\le X\le x_2)=F(x_2)-F(x_1)$.

С помощью Ф.р.в. можно вычислить вероятность противоположного события

$P(X\ge x)=1-F(x)$.

ФУНКЦИЯ ХАРАКТЕРИСТИЧЕСКАЯ

один из способов задания закона распределения *случайной величины*. Ф.х. могут быть удобнее в тех случаях, когда, напр., *плотность* или функция распределения имеют очень сложный вид. Также Ф.х. – удобный инструмент для изучения вопросов сходимости по распределению.

Ф.х. $\varphi(t)$ случайной величины X (или характеристической функцией соответствующего распределения) называется *математическое ожидание* случайной величины

e^{itx}: $\varphi_X(t)=M\left[e^{itx}\right]$,

где t – вещественная переменная; $i=\sqrt{-1}$ – мнимая единица. Из данного определения следует, что если

$Y=aX+b$,

то $\varphi_Y(t)=e^{itb}\varphi_X(at)$.

Кроме этого, если $X_1,\ldots,X_n$ – независимые случайные величины и

$S_n=X_1+\ldots+X_n$,

то $\varphi_{S_n}(t)=\prod_{j=1}^{n}\varphi_{X_j}(t)$.

Если случайная величина X имеет функцию плотности $f(x)$, то Ф.х. будет иметь вид:

$$\varphi(t)=\int_{-\infty}^{+\infty}e^{itx}f(x)dx.$$

Из формулы видно, что характеристическая функция является преобразованием Фурье функции плотности $f(x)$. На основании обратного преобразования Фурье получаем выражение для функции плотности:

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{itx}\varphi(t)dt.$$

Формула показывает, что функция плотности случайной величины однозначно определяется её Ф.х. Следовательно, Ф.х. также является исчерпывающей характеристикой случайной величины и при решении различных задач вместо функции плотности или функции распределения может быть использована соответствующим образом характеристическая функция, если при этом упрощаются математические выкладки.

Для случайной величины X дискретного типа Ф.х. $\varphi_X(t)$ имеет вид:

$$\varphi_X(t)=M\left[e^{itx}\right]=\sum_j x_j p_j,$$

где $p_j=P(X=x_j)$.

Пусть X – случайная величина с функцией распределения

$F=F(x)$ и $\varphi(t)=M\left[e^{itx}\right]$ – её Ф.х. Свойства Ф.х.:

$|\varphi(t)|\le\varphi(0)=1;\ \varphi(t)$

равномерно непрерывна по $t\in\mathbf{R}$;

$\varphi(t)=\overline{\varphi(-t)}$; $\varphi(t)$

является действительно значимой функцией тогда и только тогда, когда распределение F симметрично; если для некоторого $n\ge 1$

$M\left[|X|^n\right]<\infty$,

то при всех $r\le n$ существуют производные

$\varphi^{(r)}(t)$

и $\varphi^{(r)}(t)=\int_{-\infty}^{+\infty}(ix)^r e^{itx}dF(x)$,

$$\alpha_n=M\left[X^n\right]=\frac{\varphi^{(r)}(0)}{i^r},$$

$$\varphi(t)=\sum_{r=0}^{n}\frac{(it)^r}{r!}M\left[X^r\right]+\frac{(it)^n}{n!}\varepsilon_n(t),$$

где $|\varepsilon_n(t)|\le 3M\left[|X|^n\right]$ и $\varepsilon_n(t)\to 0$, $t\to 0$;

если существует и является конечной

$\varphi^{(2n)}(0)$, то $M\left[X^{2n}\right]<\infty$.

Ряд свойств Ф.х. выражают факт возможности некоторой функции $\varphi(t)$ быть характеристической. Теорема Бохнера-Хинчина: пусть $\varphi(t)$ – непрерывная функция, $t\in\mathbf{R}$, и $\varphi(0)=1$. Для того чтобы $\varphi(t)$ была характеристической, необходимо и достаточно, чтобы она была неотрицательно определённой, т.е. для любых действительных $t_1,\ldots,t_n$ и любых комплексных чисел

$\lambda_1,\ldots,\lambda_n$, $n=1,2,\ldots$,

$$\sum_{i,j=1}^{n}\varphi(t_i-t_j)\lambda_i\bar{\lambda}_j\ge 0.$$

Теорема Марцинкевича: если Ф.х. $\varphi(t)$ имеет вид $e^{Q(t)}$, где $Q(t)$ – полином, то степень этого полинома не может быть больше двух.

ФУНКЦИЯ СТАНДАРТНОГО НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

см. в ст. Распределение нормальное (Гауссово)

X

ХОТЕЛЛИНГА РАСПРЕДЕЛЕНИЕ

см. в ст. Распределение Хотеллинга

Ц

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

– группа теорем, предметом изучения которых являются предельные законы распределения, иногда эти теоремы называют "количественной формой *закона больших чисел*".

Все формы Ц.п.т. посвящены установлению условий, при которых возникает нормальный закон распределения. Т.к. эти условия на практике весьма часто выполняются, закон *распределения нормального* – самый распространённый из законов распределения, наиболее часто встречающийся в случайных явлениях природы. Он возникает во всех случаях, когда исследуемая *случайная величина* может быть представлена в виде суммы достаточно большого числа независимых (или слабо зависимых) элементарных слагаемых, каждое из которых в отдельности сравнительно мало влияет на сумму.

Различные формы Ц.п.т. отличаются между собой условиями, накладываемыми на распределения образующих сумму случайных слагаемых. Сформулируем сначала одну из самых простых форм Ц.п.т., относящуюся к случаю одинаково распределенных слагаемых.

Если $X_1, X_2, \ldots, X_n, \ldots$

– независимые случайные величины, имеющие один и тот же закон распределения с математическим ожиданием m и дисперсией D, то при неограниченном увеличении n закон распределения суммы

$$Y_n = \sum_{i=1}^{n} X_i$$

неограниченно приближается к нормальному закону.

Отметим, что Ц.п.т. справедлива и для неодинаково распределённых слагаемых. Напр., А.М. Ляпунов доказал Ц.п.т.: если для последовательности независимых случайных величин $X_1, X_2, \ldots, X_n, \ldots$ выполняется условие:

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} b_k}{\left(\sum_{k=1}^{n} D_k\right)^{\frac{3}{2}}} = 0,$$

где $b_k = M\left[\left|X_k - M[X_k]\right|^3\right]$ $(k = 1,2,\cdots,n)$, D_k – дисперсия величины X_k, то при неограниченном увеличении n закон распределения суммы

$$Y_n = \sum_{i=1}^{n} X_i$$

неограниченно приближается к нормальному закону.

Наиболее общее (необходимое и достаточное) условие справедливости Ц.п.т. – условие Линдеберга: при любом $\tau > 0$

$$\lim_{n\to\infty} \frac{1}{B_n^2} \sum_{k=1}^{n} \int_{|x-m_k|>\tau B_n} (x-m_k)^2 \, p_k(x)dx = 0,$$

где m_k – математическое ожидание, $p_k(x)$ – плотность распределения случайной величины X_k, $B_n = \sqrt{\sum_{k=1}^{n} D_k}$.

Частные случаи выражения Ц.п.т. – *теоремы Муавра–Лапласа*.

ЦЕНТРАЛЬНЫЙ МОМЕНТ ПОРЯДКА Q

математическое ожидание 2-й степени центрированной *случайной величины* для одномерного распределения

$$M\left[(X-\mu_x)^q\right].$$

Для случайной величины *дискретной* центральный момент выражается суммой

$$m_q^{(0)} = \sum_{i=1}^{n} \left(x_i^0 - \mu_x\right)^q p_i,$$

а для *непрерывной* – интегралом

$$m_q^{(0)} = \int_{-\infty}^{\infty} (x-\mu_x)^q \, f(x)dx.$$

Особое место занимает центральный момент второго порядка, который представляет собой дисперсию случайной величины.

Кроме центрального момента второго порядка в теории вероятностей для описания случайных величин применяются центральные моменты третьего и четвёртого порядков.

Третий центральный момент $m_3^{(0)}$ служит для характеристики асимметрии («скошенности») распределения и присутствует в формуле для расчёта *коэффициента асимметрии*. Четвертый центральный момент $m_4^{(0)}$ служит для характеристики островершинности («крутости») распределения и присутствует в формуле для расчёта *коэффициента эксцесса*.

Раскрывая $(x-\mu_x)^q$ под знаком интеграла (или суммы), легко установить связи, существующие между центральными и начальными моментами:

$m_1^{(0)} = 0;\ m_2^{(0)} = m_2 - m_1^2;$

$m_3^{(0)} = m_3 - 3m_1m_2 + 2m_1^3;$

$m_4^{(0)} = m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4$ и т.д.

ЦЕПИ МАРКОВА

см. в ст. Марковская цепь

ЦЕПИ МАРКОВА НЕПРИВОДИМЫЕ

см. в ст. Марковская цепь

ЦЕПИ МАРКОВА ПЕРИОДИЧЕСКИЕ

см. в ст. Марковская цепь

ЦЕПИ МАРКОВА ЭРГОДИЧЕСКИЕ

см. в ст. Марковская цепь

Ч

ЧАСТНАЯ (МАРГИНАЛЬНАЯ) ПЛОТНОСТЬ ВЕРОЯТНОСТИ

см. в ст. Распределение маргинальное (частное)

ЧАСТНАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

см. в ст. Распределение маргинальное (частное)

Э

ЭКСПОНЕНЦИАЛЬНОЕ (ПОКАЗАТЕЛЬНОЕ) РАСПРЕДЕЛЕНИЕ

см. в ст. Распределение экспоненциальное

ЭЛЕМЕНТАРНОЕ СОБЫТИЕ

непосредственный исход случайного эксперимента; также называют «точками», «элементами», «случаями». Э.с. рассматриваются как неразложимые и взаимоисключающие исходы случайного эксперимента.

Множество всех Э.с. образует *пространство элементарных событий.* Входящие в него Э.с. обозначаются строчными буквами греческого алфавита, при необходимости с индексами: ω, ω_1, ω_2, ω_3,...

Выделение Э.с. – первый шаг при построении *вероятностной модели* реального явления или процесса. Напр., при бросании игральной кости (проведении случайного эксперимента) получаем 6 непосредственных исходов (Э.с.) ω_1, ω_2, ω_3, ω_4, ω_5, ω_6. Событие ω_i означает, что в результате бросания кости выпало i очков, i=1, 2, 3, 4, 5, 6.

Рубрика 2.1.2. Математико-статистические методы

А

АЛЬТЕРНАТИВНАЯ ГИПОТЕЗА

гипотеза H_1, являющаяся логическим отрицанием *нулевой* (или осн.) *гипотезы* H_0. А.г. может быть простой и сложной. Напр., для нулевой гипотезы «математическое ожидание а ген. совокупности равно некоторому конкретному значению a_0» $(H_0 : a = a_0)$, альтернативными могут быть гипотезы:

$$H_1: a = a_1, \quad H_1: a \neq a_0, \quad H_1: a > a_0, \quad H_1: a < a_0.$$

При этом только первая из приведённых А.г. является простой, т.к. ей соответствует одна точка пространства параметров – величина a_1; остальные А.г. будут сложными, т.к. предполагают выбор к.-л. одного из целого множества значений. Напр., сложная гипотеза $H_1: a \neq a_0$ может быть представлена в виде множества простых гипотез

$$H_1: \{a = a^*, \quad a^* \in (-\infty, a_0] \cup [a_0, \infty)\},$$

ей соответствует какое-либо значение параметра а из множества – объединения полуинтервалов $(-\infty, a_0]$ и $[a_0, \infty)$.

Также, как и нулевая гипотеза, А.г. может быть параметрической, если в ней речь идёт о числовых значениях некоторых параметров, или непараметрической, если в ней высказывается предположение обо всём распределении: о его виде или его общих свойствах.

См. также *Гипотеза статистическая*.

АПОСТЕРИОРНЫЙ БАЙЕСОВСКИЙ РИСК

характеристика качества принимаемого решения при *байесовском подходе к оцениванию* в теории статистических решений. А.б.р. задается выражением:

$$R_{ps}(\mathbf{x}) = \int_{\Theta} C(\hat{\theta}, \theta) f_{ps}(\theta \mid \mathbf{x}) d\theta,$$

где $C(\hat{\theta}, \theta)$ – функция потерь, $f_{ps}(\theta \mid \mathbf{x})$ – апостериорная плотность распределения параметра θ при заданном наборе наблюдений $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Апостериорная вероятность $f_{ps}(\theta \mid \mathbf{x})$ может быть определена на основании известного априори распределения $p(\theta)$ параметра θ и условной плотности вероятности распределения $L(\mathbf{x} \mid \theta)$ выборочных данных при фиксированном значении неизвестного параметра θ (её называют функцией правдоподобия):

$$f_{ps}(\theta \mid \mathbf{x}) = \frac{p(\theta) L(\mathbf{x} \mid \theta)}{\int_{\Theta} p(\theta) L(\mathbf{x} \mid \theta) d\theta}.$$

В качестве функции потерь могут быть использованы: простая –

$$C(\hat{\theta}, \theta) = c_0 - \delta(\hat{\theta} - \theta),$$

где $\delta(z)$ – дельта-функция; линейная по модулю

$$C(\hat{\theta}, \theta) = |\hat{\theta} - \theta|;$$

квадратичная $C(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$,

экспоненциальная

$$C(\hat{\theta}, \theta) = 1 - \exp[-(\hat{\theta} - \theta)^2 / 2\eta^2]$$

и другие. Т.о., для заданного априорного распределения параметра $p(\theta)$ для каждого возможного решения (напр., о значении параметра θ) вычисляется средняя потеря (апостериорный байесовский риск) относительно этого апостериорного распределения

$f_{ps}(\theta \mid \mathbf{x})$.

В качестве искомого оптимального решения выбирается то, для которого А.б.р. минимален. Если случайная величина X дискретна, то вместо интегралов пишут соответствующие суммы.

АПРИОРНАЯ СТАТИСТИЧЕСКАЯ (ВЫБОРОЧНАЯ) ИНФОРМАЦИЯ

информация, основанная на выборочных данных, имеющихся в распоряжении исследователя до проведения предполагаемого статистического анализа. Эта информация, характеризующая некоторый процесс или результаты функционирования системы, может быть представлена в различных видах. В математической постановке задачи на «входе» исследователь имеет либо матрицы типа «объект-свойство», либо матрицы парных сравнений объектов. В зависимости от целей проводимого анализа исследователю могут понадобиться данные несколько иного типа. Так, если предполагается решать задачу классификации, то может понадобиться информация о структуре и свойствах классов – групп объектов. Если подобная информация имеется в наличии, т.е. определены некоторые группы объектов, принадлежность

которых к определенному классу заранее известна, то в этом случае дополнительная А.с.и. представлена т.н. обучающими выборками.

См. также *Исходные статистические данные.*

АСИМПТОТИЧЕСКАЯ ОТНОСИТЕЛЬНАЯ ЭФФЕКТИВНОСТЬ

позволяет оценить эффективность полученной оценки некоторого параметра при неограниченном увеличении числа наблюдений, а также сравнить статистические оценки по их эффективности при $n \to \infty$, и определяется пределом отношения асимптотических дисперсий оценок:

$$K = \lim_{n\to\infty} \frac{D(\theta_1)}{D(\theta_2)}.$$

Если $K < 1$, то оценка θ_1 обладает меньшей дисперсий по сравнению с оценкой θ_2 в пределе при $n \to \infty$. Если в качестве θ_2 для сравнения взята эффективная оценка, то по величине K судят об эффективности оценки θ_1. Так, напр., доказано, что среднее арифметическое – наилучшая (состоятельная, несмещённая, эффективная) оценка математического ожидания a, тогда как выборочная медиана Me, также являющаяся несмещенной оценкой a, не является асимптотически наилучшей, т.к.

$$K = \lim_{n\to\infty} \frac{D(\bar{x})}{D(Me)} = \frac{2}{\pi} < 1.$$

Тем не менее использование Me имеет свои положительные стороны. Так, если истинное распределение не является в точности нормальным, а несколько отличается от него, то дисперсия $\bar{x}$ может резко возрасти, а дисперсия Me остается почти той же, т. е. Me будет более устойчивой характеристикой в данной ситуации.

АСИМПТОТИЧЕСКИЕ СВОЙСТВА ОЦЕНОК

неизвестных параметров распределения, определённых по выборочным данным – свойства, которыми они обладают при неограниченном увеличении объёма выборки (при $n \to \infty$). Одно из осн. свойств статистических оценок – свойство состоятельности: по мере роста числа наблюдений (при $n \to \infty$) оценка

$$\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$$

сходится по вероятности к оцениваемому значению

$$\theta: \lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1.$$

Доказано, что выборочные моменты (начальные и центральные) $\hat{\nu}_k$ и $\hat{\mu}_k$, эмпирические функции распределения $\hat{F}(x)$, функция плотности $\hat{f}(x)$ и относительные частоты $\hat{p}$ при неограниченном увеличении объёма выборки стремятся по вероятности к соответствующим теоретическим характеристикам, т.е. обладают статистической устойчивостью. Это даёт основание использовать выборочные характеристики для приближенного описания свойств всей ген. совокупности. Несмотря на то, что осн. свойства выборки объёма n зависят от распределения ген. совокупности, по мере увеличения объёма выборки выборочные характеристики начинают вести себя одинаковым образом независимо от специфики ген. совокупностей, по выборкам из которых они были вычислены. На основании *центральной предельной теоремы* доказано, что при больших объёмах выборок все осн. выборочные характеристики ведут себя как нормально распределённые величины., т.е. являются асимптотически нормальными. Конкретизация закона распределения для каждой конкретной выборочной характеристики может быть произведена путём вычисления её среднего значения и дисперсии. Так для осн. выборочных характеристик – среднего $\bar{x}$, дисперсии $\hat{\sigma}^2$ и функции распределения $\hat{F}(x)$ доказано:

$$\bar{x} \to N\left(a, \frac{\sigma^2}{n}\right);$$

$$\hat{\sigma}^2 \to N\left(\frac{n-1}{n}\sigma^2, \frac{2(n-1)\sigma^4}{n^2}\right);$$

$$\hat{F}(x) \to N\left(F(x), \frac{F(x)(1-F(x))}{n}\right).$$

Желательным свойством «хорошей» оценки является её эффективность. Однако не все выборочные характеристики – *оценки эффективные*. Так, *выборочная дисперсия* и «исправлен-

ная» выборочная дисперсия являются только асимптотически эффективными, т.е. эффективными при $n \to \infty$.

Б

БАЙЕСОВСКАЯ КЛАССИФИКАЦИЯ

отнесение объектов к определённым классам, минимизирующее *байесовский риск*

$$\bar{r} = \sum_i \sum_j r(s_i, \hat{s}_j) P(s_i, \hat{s}_j),$$

компонентами которого являются стоимость (риск) $r(s_i, \hat{s}_j)$ отнесения j-го объекта к i-му классу и вероятность такой ситуации. При простой функции стоимости

$$r(s_i, \hat{s}_j) = \begin{cases} -1 \ при \ \hat{s}_j = s_i \\ 0 \ при \ \hat{s}_j \neq s_i \end{cases}$$

минимизация байесовского риска эквивалента максимизации вероятности правильной классификации объектов, равной сумме

$$\sum_{i=1}^{k} \pi_i P(\hat{s}_j \mid s_i)$$

произведений вероятности появления объекта i-го класса (равной доле этих объектов в общей их совокупности) π_i и условной вероятности правильного его отнесения к i-му классу $P(\hat{s}_j \mid s_i)$.

Правила Б.к. с определёнными ограничениями реализуются в виде параметрических процедур *дискриминантного анализа* и классификации на основе расщепления смесей унимодальных распределений. Они находят широкое применение и в непараметрических процедурах классификации, таких как метод парзеновского окна, метод ближайших соседей и других.

БАЙЕСОВСКАЯ ОЦЕНКА

оценка $\hat{\theta}$ неизвестного параметра θ, который при *байесовском подходе к оцениванию* считают случайным, вычисляемая по выборке

$X = (x_1, x_2, \ldots, x_n)$

из ген. совокупности в соответствие с критерием минимума среднего риска ошибок оценивания, или *байесовского риска*

$$\bar{r} = \int_{\theta} \int_{\hat{\theta}} r(\theta, \hat{\theta}) f(\theta, \hat{\theta}) d\theta d\hat{\theta}$$

для выбранной функции стоимости ошибок $r(\theta, \hat{\theta})$.

Пусть θ и X – непрерывные величины, $f_0(\theta)$ – априорная функция плотности вероятности параметра θ, а $f(X \mid \theta)$ – плотность условного распределения выборки X при данном значении параметра θ. Апостериорная плотность распределения $f_1(\theta \mid X)$ параметра θ после наблюдения выборки X определяется формулой Байеса:

$$f_1(\theta \mid X) = \frac{f(X \mid \theta) f_0(\theta)}{\int_{\theta} f(X \mid \theta) f_0(\theta) d\theta}.$$

Б.о. при квадратичной функции стоимости ошибок оценивания

$$r(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

равна математическому ожиданию величины θ, вычисленному по апостериорному распределению

$$\hat{\theta}_Б(X) = \int_{\theta} f_1(X \mid \theta) f_0(\theta) d\theta.$$

Оценка $\hat{\theta}_Б(X)$,

как и оценки, получаемые другими методами, является функцией выборки. В случае симметричной унимодальной функции плотности вероятности $f(\theta, \hat{\theta})$ и равномерной априорной плотности вероятности Б.о. оценка совпадает с небайесовской оценкой макс. правдоподобия. Однако наличие дополнительной априорной информации об оцениваемом параметре позволяет повысить точность оценки.

БАЙЕСОВСКИЙ ПОДХОД К ОЦЕНИВАНИЮ

один из способов операционализации априорной информации об изучаемом процессе (объекте, системе) при принятии статистических решений, которые вырабатываются на основании информации двух типов: априорной и содержащейся в исходных статистических данных (наблюдениях). Идеология именно байе-

совского способа операционализации априорной информации об изучаемом процессе основана на двух принципах: а) степень «разумной уверенности» исследователя в справедливости некоторого утверждения (напр., утверждения относительно возможного значения оцениваемого параметра модели) численно выражается в виде вероятности; это означает, что вероятность в байесовском подходе интерпретируется в рамках субъективной школы теории вероятностей; б) априорная информация об оцениваемом параметре модели предоставлена исследователю в виде некоторого априорного распределения вероятностей этого параметра, которое отражает степень уверенности исследователя в том, что оцениваемый параметр примет то или иное значение ещё до использования исходных статистических данных. По мере поступления этих данных исследователь уточняет (пересчитывает) с помощью формулы Байеса это распределение, переходя от априорного распределения к апостериорному.

Рассмотрим подробнее упомянутые выше понятия и общую логическую схему реализации байесовского подхода применительно к задаче статистического оценивания неизвестного параметра модели.

Априорные сведения о параметре Θ основаны на предыстории функционирования анализируемого процесса (если таковая имеется) и на профессиональных теоретических соображениях о его сущности, специфике, особенностях и т. п. В конечном итоге эти априорные сведения должны быть представлены в виде функции $p(\Theta)$, задающей априорное распределение параметра и интерпретируемой как вероятность того, что параметр примет значение, равное Θ, если параметр дискретен, и как функция плотности распределения в точке Θ, если параметр непрерывен по своей природе. В ситуациях, когда априорные сведения об анализируемом параметре слишком скудны, в качестве априорного распределения $p(\Theta)$ используют, напр., равномерное на отрезке

$[\Theta_{\min}, \Theta_{\max}]$ распределение,

где $[\Theta_{\min}, \Theta_{\max}]$ – априорный диапазон варьирования возможных значений оцениваемого параметра,

т.е.:

$$p(\Theta) = \begin{cases} \dfrac{1}{\Theta_{\max} - \Theta_{\min}} & \text{при } \Theta_{\min} \le \Theta \le \Theta_{\max}; \\ 0 & \text{при } \Theta \notin [\Theta_{\min}, \Theta_{\max}]. \end{cases}$$

Исходные статистические данные $x_1, x_2, \ldots, x_n$ – выборка объёма п из анализируемой ген. совокупности. Получая такие данные, к имевшейся ранее априорной информации о параметре присоединяем выборочную (эмпирическую) информацию.

Вычисление функции правдоподобия

$L(x_1, x_2, \ldots, x_n \mid \Theta)$

производится для независимых наблюдений $x_1, \ldots, x_n$ по формуле:

где $f(x \mid \Theta)$ – функция плотности (или вероятность $P\{\xi = x \mid \Theta\}$), описывающая закон распределения вероятностей анализируемой ген. совокупности в предположении (или при условии), что значение оцениваемого параметра равно Θ.

Вычисление апостериорного распределения

$\varphi(\Theta \mid x_1, \ldots, x_n)$

осуществляется с помощью формулы Байеса:

$$P\{A_i \mid B\} = \frac{P\{A_i\}P\{B \mid A_i\}}{\sum_{j=1}^{k} P\{B \mid A_j\} \cdot P\{A_j\}},$$

в которой роль события A_i ; играет событие, заключающееся в том, что значение оцениваемого параметра равно Θ, а роль условия В – событие, заключающееся в том, что значения *n* наблюдений, произведённых в анализируемой ген. совокупности, зафиксированы на уровнях $x_1, x_2, \ldots, x_n$.

Соответственно имеем:

$$\varphi(\Theta \mid x_1, \ldots, x_n) = \frac{p(\Theta)L(x_1, \ldots, x_n \mid \Theta)}{\int L(x_1, \ldots, x_n \mid \Theta) \cdot p(\Theta) d\Theta}. \qquad (1)$$

Построение байесовских точечных и интервальных оценок основано на использовании знания апостериорного распределения

$$\varphi(\Theta \mid x_1,\ldots,x_n),$$

задаваемого соотношением (1). В частности, в качестве байесовских точечных оценок $\hat{\Theta}^{(6)}$ используют среднее или модальное значение этого распределения, т.е.:

$$\hat{\Theta}_{\text{ср}}^{(\text{Б})} = E(\Theta \mid x_1,\ldots,x_n) = \int \Theta \varphi(\Theta \mid x_1,\ldots,x_n)\, d\Theta,$$

$$\hat{\Theta}_{\text{ср}}^{(\text{Б})} = \arg\max_{\Theta} \varphi(\Theta \mid x_1,\ldots,x_n).$$

Для вычисления этих оценок нам достаточно знать только числитель правой части (1), так как знаменатель этого выражения играет роль нормирующего множителя и от Θ не зависит, это существенно упрощает процесс практического построения оценок $\hat{\Theta}_{\text{ср}}^{(6)}$ *и* $\hat{\Theta}_{\text{мод}}^{(6)}$.

Оценка $\hat{\Theta}_{\text{ср}}^{(\text{Б})}$ обладает одним важным оптимальным свойством. Пусть $\hat{\Theta}(x_1,\ldots,x_n)$ – любая оценка параметра Θ. Оказывается, если качество любой оценки $\hat{\Theta}(x_1,\ldots,x_n)$ измерять т.н. апостериорным байесовским риском:

$$R^{(\text{Б})}(x_1,\ldots,x_n) = E\left\{(\hat{\Theta}(x_1,\ldots,x_n)-\Theta)^2 \mid x_1,\ldots,x_n\right\} = \int\left(\hat{\Theta}(x_1,\ldots,x_n)-\Theta\right)^2 \varphi(\Theta \mid x_1,\ldots,x_n)d\Theta$$

или его средним (усреднение – по всем возможным выборкам $x_1,\ldots,x_n$) значением

$$R_{\text{ср}}^{(\text{Б})},$$

то байесовская оценка (2) является наилучшей и в том и в другом смысле.

Для построения байесовского доверительного интервала для параметра Θ необходимо вычислить по формуле (1) функцию

$$\varphi(\Theta \mid x_1,\ldots,x_n),$$

характеризующую апостериорный закон распределения параметра Θ, а затем по заданной доверительной вероятности *P* определить

$$100\frac{1+P}{2}$$ – и $$100\frac{1-P}{2}\%$$-ные

точки этого закона, которые и дают соответственно левый и правый концы искомой интервальной оценки.

Существенную роль в реализации Б.п. к о. играют т.н. «сопряжённые» (или «естественно сопряжённые») семейства априорных распределений. Семейство зависящих от параметров Λ априорных распределений

$$G = \{P(\Theta; \Lambda)\}$$

называется сопряжённым по отношению к функции правдоподобия

$$L(x_1, x_2,\ldots, x_n \mid \Theta),$$

если и апостериорное распределение

$$\varphi(\Theta \mid x_1,\ldots,x_n),$$

вычисленное по формуле (1), снова принадлежит этому же семейству *G*. Т.о., использование в качестве априорных законов распределения $P(\Theta)$ сопряжённых по отношению к L законов облегчает статистику задачу вычислительной реализации формулы (1): ему остается лишь уметь пересчитывать значения параметров Λ. Естественность и востребованность сопряжённых семейств априорных распределений можно объяснить, в частности, следующим обстоятельством: оказывается, если сопряжённое априорное распределение существует и если в Б.п. к о. стартовать с априорного распределения, не несущего никакой дополнительной, по отношению к имеющимся исходным статистическим данным, полезной информации об оцениваемом параметре Θ (такие распределения являются некоторым обобщением равномерного распределения на любой, в т.ч., и бесконечной области возможных значений Θ), то первый же переход от него по формуле (1) к апо-

стериорному распределению приведет нас к семейству распределений, сопряжённому с наблюдаемой ген. совокупностью.

Байесовский способ оценивания может давать весьма ощутимый выигрыш в точности при ограниченных объёмах выборок. В процессе же неограниченного роста объёма выборки n оба подхода (байесовский и классический, основанный на методе макс. правдоподобия) будут давать, в силу их состоятельности, всё более похожие результаты.

БАЙЕСОВСКИЙ РИСК

средний ожидаемый риск ошибки при реализации статистических процедур оценивания и классификации. Определение Б.р. при статистическом оценивании случайного параметра θ требует знания вида совместного закона распределения $f(\theta,\hat{\theta})$ этого параметра θ и его оценки $\hat{\theta}$, а также задания функции стоимости ошибки $r(\theta,\hat{\theta})$.

Наиболее распространённые виды функции стоимости – линейная $r(\theta,\hat{\theta})=\left|\theta-\hat{\theta}\right|$, квадратичная

$r(\theta,\hat{\theta})=\left(\theta-\hat{\theta}\right)^2$ и простая.

Простая функция стоимости различна для случаев оценивания непрерывного и дискретного параметров. В непрерывном случае она определяется с помощью δ-функции Дирака:

$$r(\theta,\hat{\theta})=c-\delta(\theta-\hat{\theta}),$$

в которой константа c представляет собой одинаковый риск для любой ошибки оценивания и бесконечно большую величину отрицательного риска («премии») за абсолютно точную оценку $\hat{\theta}=\theta$. В дискретном случае простая функция стоимости может быть записана в том же виде, но $\delta(\theta-\hat{\theta})$ в ней является символом Кронекера

$$\delta(\theta-\hat{\theta})=\begin{cases}1 \text{ при } \hat{\theta}=\theta\\ 0 \text{ при } \hat{\theta}\neq\theta\end{cases}.$$

При оценивании непрерывного параметра Б.р. определяется путём интегрирования всех возможных значений параметра и его оценки с функцией стоимости в качестве весовой функции:

$$\bar{r}=\int_{\theta}\int_{\hat{\theta}} r(\theta,\hat{\theta})f(\theta,\hat{\theta})d\theta d\hat{\theta}.$$

В случае оценивания дискретного параметра Б.р. определяется весовой суммой

$$\bar{r}=\sum_{\theta}\sum_{\hat{\theta}} r(\theta,\hat{\theta})P(\theta,\hat{\theta}),$$

где $P(\theta,\hat{\theta})$ – вероятность того, что одновременно истинный параметр и его оценка примут соответственно значения θ и $\hat{\theta}$.

В задачах классификации Б.р. – средний риск ошибочного принятия решения $\hat{s}_j$ отнесения к классу s_j объекта, относящегося к классу s_i:

$$\bar{r}=\sum_{i}\sum_{j} r(s_i,\hat{s}_j)P(s_i,\hat{s}_j).$$

Б.р. – универсальный байесовский критерий. Построение оптимальных процедур измерения и классификации при использовании *байесовского подхода* предполагает минимизацию Б.р.

БАЙЕСОВСКИЙ РИСК АПОСТЕРИОРНЫЙ

см. в ст. Апостериорный байесовский риск

БЕРЕНСА-ФИШЕРА ПРОБЛЕМА

задача сравнения средних двух нормально распределённых выборок при неизвестных и неравных дисперсиях, т.е. задача проверки определённой *статистической гипотезы* $H_0:\mu_1=\mu_2$. Отметим, что успешно решены задачи проверки гипотез о равенстве ген. средних двух совокупностей в случае известных, или неизвестных, но равных дисперсий этих совокупностей. Точного же решения Б.-Ф.п. до настоящего времени нет. На практике используются различные приближения. Пусть имеются две выборки

$x=(x_1,x_2,\ldots,x_n)$

и $y=(y_1,y_2,\ldots,y_m)$.

В качестве статистики критерия рассматривается величина

$$t=\frac{\bar{x}-\bar{y}}{s},$$

где $s^2 = \frac{1}{n}s_x + \frac{1}{m}s_y$,

$$s_x^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$s_y^{\,2} = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$$

– выборочные дисперсии;

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \quad \bar{y} = \frac{1}{m}\sum_{i=1}^{m}y_i$$

– выборочные средние. В зависимости от вида *альтернативной гипотезы* H_1, критическая область, при попадании в которую статистики критерия, нулевая гипотеза отвергается, можно выделить три осн. случая:

$t > t_{кр} = t^*_{\alpha/2}$ при $H_1 : \mu_x \neq \mu_y$,

$t < t_{кр} = t^*_{\alpha}$ при $H_1 : \mu_x < \mu_y$,

$t > t_{кр} = t^*_{1-\alpha}$ при $H_1 : \mu_x > \mu_y$,

где квантили t^*_α определяются по-разному для различных приближений, α - заданный уровень значимости. В критерии Кокрана-Кокса

$$t^*_\alpha = \frac{\nu_x t_{\alpha,n-1} + \nu_y t_{\alpha,m-1}}{\nu_x + \nu_y}, \quad \nu_x = \frac{s_x^{\,2}}{n}, \quad \nu_y = \frac{s_y^{\,2}}{m},$$

где $t_{\alpha,k}$ – α-квантиль распределения Стьюдента с k степенями свободы. Для критерия Сатервайта t^*_α – α-квантиль распределения Стьюдента с числом степеней свободы

$$f = s^4\left[\frac{1}{n-1}\left(\frac{s_x^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_y^2}{m}\right)^2\right]^{-1}.$$

В критерии Крамера-Уэлча число степеней свободы отличается от предыдущего критерия на двойку:

$l = f - 2$.

БУТСТРЕП МЕТОД

см. в ст. Бутстреп - моделирование регрессионных моделях.

БУТСТРЕП-МОДЕЛИРОВАНИЕ

создание моделей с помощью тиражирования исходных данных при недостаточности их исходного объёма. Методы Б.-м. получили развитие с появлением возможности быстро производить большие объёмы вычислений. Из одного множества данных генерируются «новые» множества путём тиражирования выборки чисел. Эти методы используются во многих ситуациях. Б.-м. актуально при изучении процессов, для которых характерны экспериментальные данные ограниченной выборки с большой неоднородностью. Помимо нахождения оценок параметров, важной задачей, решаемой при помощи тиражирования выборки, является получение хороших оценок стандартных ошибок для распределений, генерируемых в ходе тиражирования. Это особенно ценно в ситуациях, когда выражения для стандартных ошибок нельзя непосредственно вывести теоретически. Для регрессионных моделей можно использовать две осн. процедуры тиражирования выборки: Б.-м. с использованием остатков и Б.-м. с использованием пар. В первом методе подбирается линейная модель и получаются n остатков. Из них составляется выборка объёмом n по следующему принципу: производится выбор с возвращением, причём вероятность выбора для каждого остатка составляет 1/n. Отобранные n значений прибавляются к предсказанным значениям Y_i, в результате получается новое множество переменных Y. Т.о., если модель имеет вид

$Y = X\beta + e$, и $\hat{Y} = Xb$,

то новыми Y-значениями являются

$Y^* = Xb + e^*$,

где e* – множество, полученное с помощью тиражирования выборки из вектора

$e = Y - \hat{Y}$.

Далее *методом наименьших квадратов* подбирается регрессионное уравнение

$Y^* = X\beta + e$

и получается оценка. Можно выполнить любое желаемое количество итераций и обычным способом найти выборочное среднее и стандартное отклонение для каждого элемента этих векторных оценок. В процедуре Б.-м. с использованием пар может также производиться тиражирование выборки парами (Y_i; x_i. где Yi – i-е наблюдение, а x_i – i-я строка матрицы X, т.е. значений регрессоров для i-го наблюдения. По-

вторная выборка заключается в выборе множества, состоящего из n пар (Y_i; x_i), каждая из которых выбирается с вероятностью 1/n в выборе с возвращениями, чтобы получить новый набор значений зависимой переменной и Y' и регрессоров X'. Затем методом наименьших квадратов подбирается модель вида

$$Y' = X'\beta + e .$$

После выполнения любого желаемого числа итераций можно исследовать свойства соответствующих значений. Независимо от того, какая из процедур выполняется, для оценки коэффициентов модели могут использоваться не только метод наименьших квадратов, но и робастные методы.

В

ВАРИАЦИОННЫЙ РЯД

значения n независимых наблюдений x_1, x_2,…, x_n случайной величины X, расположенные в порядке возрастания значений признака:

$x_{(1)} \le x_{(2)} \le \ldots \le x_{(k)} \le x_{(k+1)} \le \ldots \le x_{(n)}$.

Элемент $x_{(i)}$ называется *i*-й порядковой статистикой, гл. порядковые статистики –

$x_{(1)}=min\{x_{(i)}\}$ – наименьшее значение и

$x_{(n)}=\max\{x_{(i)}\}$ – наибольшее значение,

разность которых называется размахом вариационного ряда: $R=x_{(n)}-x_{(1)}$. Размах служит самостоятельной характеристикой разброса значений изучаемого признака и используется в математической статистике для оценки неизвестного среднего квадратического отклонения σ.

В.р. необходим для определения выборочной медианы – равной k-й порядковой статистике

$$\hat{Me} = x_{(k)}$$

при нечётном количестве членов вариационного ряда (n=2k+1) и среднему арифметическому k-й и (k+1)-й порядковых статистик

$$\hat{Me} = \frac{x_{(k)} + x_{(k+1)}}{2}$$

при четном числе наблюдений (n=2k). В.р. нужен для построения важнейшей характеристики неизвестного распределения случайной величины X – эмпирической функции распределения:

$$F_э(x_{(i)}) = P(X < x_{(i)}) = \frac{количество\ \ (x_{(j)} < x_{(i)})}{n} .$$

В.р. используется в математической статистике для получения выборочной оценки плотности вероятностей (для непрерывной случайной величины) или функции вероятностей (для дискретной случайной величины). В первом случае для этого строят интервальный В.р., а во втором – дискретный.

С целью улучшения представления эмпирических данных при большом числе наблюдений их группируют, получая сгруппированный В.р.

Для группировки непрерывных случайных величин весь вариационный размах

$$R=x_{(n)}-x_{(1)}$$

разбивают на некоторое количество интервалов l (напр., с помощью формулы Стерджеса ширину интервала определяют как

$$h = \frac{R}{1+\log_2 n} \approx \frac{R}{1+3{,}322 \cdot \lg n} .$$

Нижнюю границу первого интервала обычно принимают равной

$$x_{(1)} - \frac{h}{2})$$

и подсчитывают частоты попадания m_i случайной величины в каждый из построенных интервалов. В этом случае за значения x_i принимаются середины интервалов.

Для группировки дискретных В.р. подсчитывают частоту встречаемости m_i каждого признака x_i. При достаточно большом числе значений сгруппированный вариационный ряд может быть подвергнут дальнейшей группировке и преобразован в интервальный.

Сгруппированный В.р. – значения признака (для дискретных случайных величин) или середины интервалов (для непрерывных), указанные вместе с соответствующими частотами m_i или частостями $w_i=m_i/n$ (см. табл. 1). Эти частоты называют эмпирическими.

В.р.

Значение признака x_i	$x_{(1)}$	$x_{(2)}$	…	$x_{(i)}$	…	$x_{(l)}$

Частота встречаемости m_i	m_1	m_2	…	m_i	…	m_l

Сгруппированный В.р. графически представляют в виде *гистограммы* или *полигона*, отражающих распределение относительных частот

$$\frac{m_i}{n \cdot h},$$

отстоящих друг от друга на интервалы шириной *h*, и являющихся выборочной оценкой плотности вероятностей (для непрерывной случайной величины) или функцией вероятностей (для дискретной случайной величины). Сгруппированный В.р. также служит для построения эмпирической функции распределения, которая равна 0 для $x \le x_{(1)}$ и 1 для $x > x_{(l)}$, а для остальных *x* находится как отношение накопленной частоты до текущего значения признака (интервала) к объёму выборки n:

$$F_э(x_{(i)}) = \begin{cases} 0, \ при \ x \le x_{(1)} \\ \dfrac{m_{нi}}{n} = \dfrac{\sum_{k=1}^{i-1} m_k}{\sum_{k=1}^{l} m_k}, \quad при \ x_{(i-1)} < x \le x_{(i)} \qquad i = 2,...,l \\ 1, \ при \ x > x_{(l)} \end{cases}$$

С помощью эмпирической функции распределения можно проверить исследуемую совокупность данных на соответствие любому теоретическому закону распределения с использованием соответствующих статистических критериев (*критериев согласия*).

ВЕРОЯТНОСТНАЯ МОДЕЛЬ

математическая модель экономического процесса, в которой параметры, условия функционирования и характеристики состояния моделируемого объекта представлены *случайными величинами* и связаны стохастическими (т.е. случайными, нерегулярными) зависимостями, либо исходная информация также представлена случайными величинами. Следовательно, характеристики состояния в модели определяются не однозначно, а через законы распределения их вероятностей. При построении В.м. применяются методы корреляционного и регрессионного анализов, другие статистические методы.

Т.о., при задании на входе модели некоторой совокупности значений, на её выходе могут получаться различающиеся между собой результаты в зависимости от действия случайного фактора. Другие названия В.м. – недетерминированная, стохастическая модель.

ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКАЯ МОДЕЛЬ

вероятностная модель, значения отдельных характеристик которой оцениваются по исходным статистическим данным, характеризующим функционирование моделируемого конкретного явления. В.-с.м., описывающая механизм функционирования экономической или социально-экономической системы, называется эконометрической. Если математическая модель описывает механизм функционирования некой гипотетической экономической или социально-экономической системы, то модель называется экономико-математической или просто экономической. Для построения и экспериментальной проверки В.-с.м. использует одновременно информацию двух типов: априорную информацию о природе и содержательной сущности анализируемого явления, представленную в виде тех или иных теоретических закономерностей, ограничений, гипотез; исходные статистические данные, характеризующие процесс и результаты функционирования конкретного анализируемого явления или кон-

кретной системы. Осн. этапы вероятностно-статистического моделирования: на первом этапе (постановочном) происходит определение конечных целей моделирования, набора факторов и показателей, их взаимосвязи и роли (объясняющие и объясняемые); второй этап (априорный, предмодельный) состоит в анализе содержательной сущности моделируемого явления или системы, формировании и формализации имеющейся априорной информации о них в виде ряда гипотез и исходных допущений; третий этап (информационно-статистический) посвящён сбору необходимой статистической информации, т.е. регистрации значений факторов и показателей на различных временных и/или пространственных тактах функционирования моделируемого явления или системы; четвёртый этап (спецификация модели) включает в себя непосредственный вывод общего вида модельных соотношений, связывающих входные и выходные переменные. На этом этапе определяется лишь структура модели, ее символическая аналитическая запись; пятый этап (идентифицируемость и параметризация модели) предназначен для проведения статистического анализа модели с целью определения возможности однозначного восстановления неизвестных параметров модели по исходным статистическим данным (идентифицируемость) и, при положительном решении данного вопроса, предложения и реализации математически корректных процедур оценивания вышеуказанных параметров (параметризация). Если проблема идентифицируемости решается отрицательно, то возвращаются к четвёртому этапу и вносят необходимые коррективы в решение задачи спецификации; шестой этап (верификация модели) заключается в использовании различных процедур сопоставления модельных заключений, оценок, следствий и выводов с реально наблюдаемой действительностью. Этот этап называют также этапом статистической точности и адекватности модели. При пессимистическом характере результатов этого этапа необходимо возвратиться к пятому этапу (использовать другой способ оценки параметров), к четвёртому (изменение спецификации) или даже к первому (пересмотр набора факторов и показателей). В результате процесс построения В.-с.м. носит последовательный итеративный характер. Построение и анализ модели основаны только на априорной информации и не предусматривают третьего и пятого этапов. В этом случае модель не является В.-с.м.

ВИЗУАЛИЗАЦИЯ ДАННЫХ

использование геометрических образов (точек, линий, плоских фигур и т.п.) для изображения числовых величин и их соотношений. Графические изображения служат одним из важнейших технических и познавательных средств статистики. Использование графиков для изложения статистических показателей позволяет придать последним наглядность и выразительность, облегчает их восприятие, а во многих случаях помогает уяснить сущность изучаемого явления, его закономерности и особенности, увидеть тенденции его развития, взаимосвязь характеризующих его показателей. Статистические графики можно классифицировать по разным признакам: назначению (содержанию), способу построения, форме графического образа и решаемым задачам. По способу построения и задачам изображения статистические графики разделяются на диаграммы и статистические карты. Диаграммы – наиболее распространенный способ графических изображений. Они применяются для наглядного сопоставления в различных аспектах (пространственном, временном и т.п.) независимых друг от друга совокупностей. В зависимости от целей представляемых данных подразделяются на диаграммы сравнения, динамики, структурные. По характеру геометрического образа различают графики точечные, линейные, плоскостные и пространственные (объёмные). При построении точечных диаграмм в качестве графических образов применяются совокупности точек; при построении линейных – линии. Осн. принцип построения всех плоскостных диаграмм сводится к тому, что статистические величины изображаются в виде геометрических фигур. Плоскостные диаграммы подразделяются на столбиковые, полосные, круговые, квадратные и фигурные. Столбиковая диаграмма чаще всего

используется для сравнения одноименных показателей, а также для изображения структуры явлений. Значения сравниваемых показателей изображаются при этом в виде прямоугольных столбиков, имеющих одинаковую ширину и расположенных на общей линии. Высота каждого столбика в определённом масштабе соответствует величине изображаемого показателя (см. рис.1).

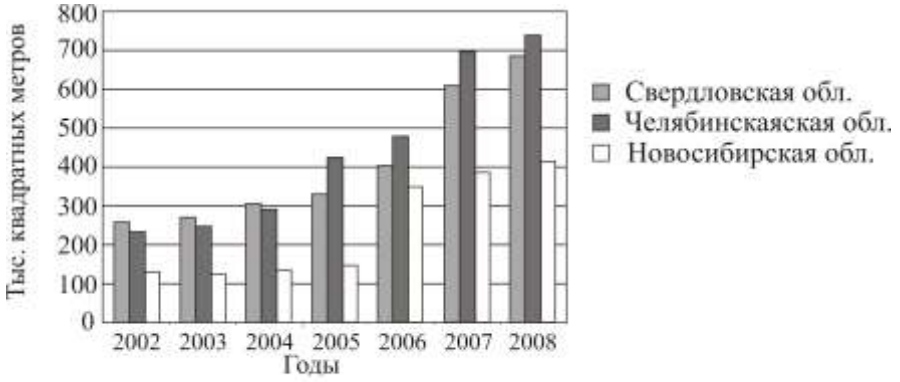


Рис.1. Индивидуальные жилые дома, построенные населением за свой счет и с помощью кредитов, 2002 - 2008 гг

Разновидностью столбиковых диаграмм являются ленточные или полосовые диаграммы. У них масштабная шкала расположена по горизонтали сверху или снизу и определяет величину явления по длине полосы. При характеристике структуры совокупности (в относительных величинах) все столбики (полосы) в диаграмме имеют одинаковую высоту и соответствуют 100%. Каждый столбик разбивается на части пропорционально удельному весу отдельных частей во всей совокупности. 2) Для простого сравнения независимых друг от друга показателей могут также использоваться диаграммы, принцип построения которых состоит в том, что сравниваемые величины изображаются в виде правильных геометрических фигур (круги, квадраты) или фигур-знаков (воспроизводят в некоторой степени внешний образ данных), которые выражают величину изображаемого явления размером своей площади. Фигурные диаграммы можно строить различной численностью фигур одинакового размера или фигурами различных размеров. Следует различать два вида использования круга при построении диаграмм. В одном случае круг используется для сравнения пл. нескольких кругов друг с другом, такого рода диаграммы называются круговыми (см. рис. 2). В другом случае круг используется для сравнения пл. отдельных секторов друг с другом, такая диаграмма называется секторной и применяется для наглядной иллюстрации структуры к.-л. явления, для характеристики удельных весов отдельных частей целого (см. рис.3), для выявления структурных сдвигов.

Рис.2. Численность населения на одного врача, 2007 г.

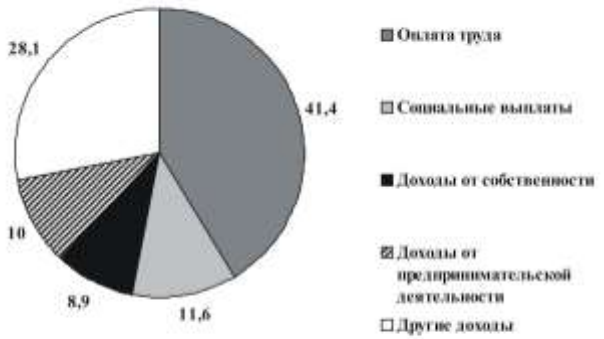


Рис.3. Структура денежных доходов населения,%,Россия,2007 г.

Прямоугольные диаграммы (но не квадратные!) находят применение при графическом изображении двухмасштабных сравнений: один масштаб для основания, другой – для высоты. Эти диаграммы носят название знаки Варзара. Обычно такого рода диаграммы применяются в тех случаях, когда показатель является произведением двух других. Его можно графически изобразить так, чтобы были видны его сомножители. Для этого поступают следующим образом: один множитель принимают за основание, другой – за высоту, устанавливают масштабы и, располагая значением статистического показателя основания и высоты, строят прямоугольники. Широкое распространение в экономике имеют диаграммы, которые можно назвать координатными, так как они основаны на системе прямоугольных координат. В отличие от столбиковых и полосовых диаграмм они требуют не одного, а двух масштабов: одного по оси абсцисс, другого по оси ординат. Для характеристики изменений явлений во времени чаще всего используются линейные диаграммы. Они незаменимы в тех случаях, когда на одном графике нужно показать динамику нескольких явлений. В статистической практике чаще всего применяются графические изображения динамики с равномерными шкалами. По оси абсцисс они пропорциональны числу периодов времени, а по оси ординат пропорциональны самим уровням.

Статистические карты – графики количественного распределения по поверхности. По своей осн. цели они близко примыкают к диаграммам и специфичны лишь в том отношении, что представляют собой условные изображения статистических данных на контурной географической карте, т.е. показывают пространственное размещение или пространственную распространённость статистических данных. Они делятся по географическому образу на картограммы и картодиаграммы. Картограмма – схематическая географическая карта, на которой штриховкой различной густоты, точками или окраской определённой степени насыщенности показывается сравнительная интенсивность к.-л. показателя в пределах каждой единицы нанесённого на карту терр. деления. Картограммы делятся на фоновые и точечные. Картограмма фоновая – вид картограммы, на которой штриховкой различной густоты или окраской определенной степени насыщенности показывают интенсивность к.-л. показателя в пределах терр. единицы (см. рис. 4). Картограмма точечная – вид картограммы, где уровень к.-л. явления изображается с помощью точек. Точка изображает одну единицу совокупности или некоторое их количество. Путём нанесения точек на географической карте показывается плотность или частота появления определённого явления. Фоновые картограммы, как правило, используются для изображения средних или относительных показателей, а точечные – для объёмных (количественных) показателей (численность нас., поголовье скота и т.д.). Вторую большую группу статистических карт составляют картодиаграммы – сочетание диаграмм с географической картой. В качестве изобразительных знаков в картодиаграммах используются диаграммные фигуры (столбики, квадраты, круги, фигуры, полосы), которые размещаются на контуре

географической карты, картодиаграммы дают возможность географически отразить более сложные статистико-географические построения, чем картограммы. Среди картодиаграмм следует выделить картодиаграммы простого сравнения, пространственных перемещений и изолиний. На картодиаграмме простого сравнения диаграммные фигуры, изображающие величины исследуемого показателя, разносятся по всей карте в соответствии с тем районом, областью или страной, которые они представляют. Элементы простейшей картодиаграммы можно обнаружить на политической карте, где города отличаются различными геометрическими фигурами в зависимости от числа жителей. Изолинии – линии равного значения к.-л. величины в её распространении на поверхности, в частности на географической карте или графике. Изолинии отражают непрерывное изменение исследуемой величины в зависимости от двух других переменных. Они применяются при картографировании природных и социально-экономических явлений; могут быть использованы для получения их количественной характеристики и для анализа корреляционных связей между ними. Перечисленные виды графиков не являются исчерпывающими, но являются наиболее часто употребляемыми.

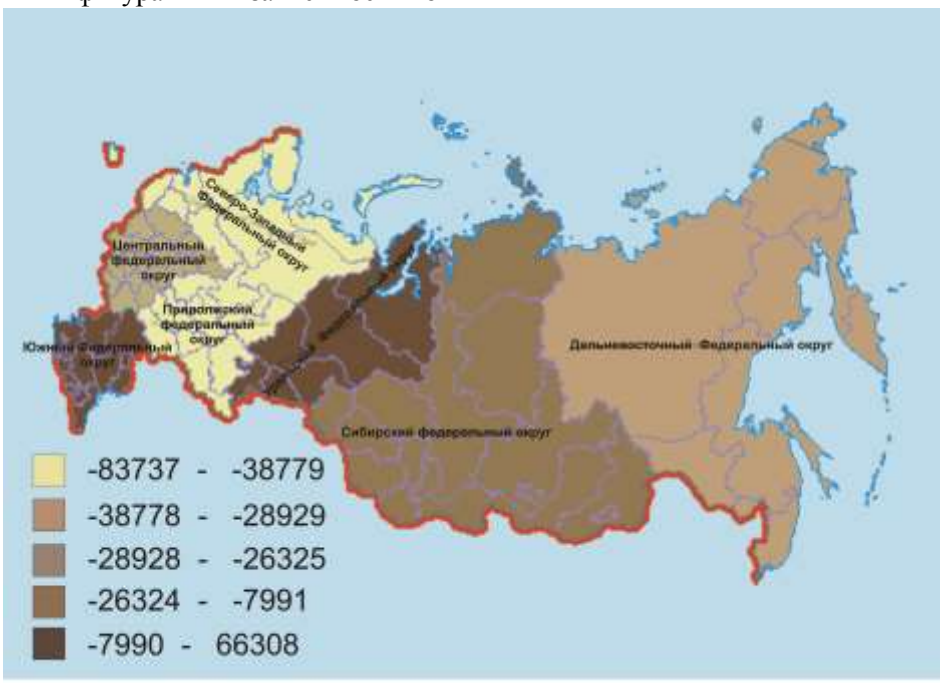


Рис.4. Общий прирост постоянного населения, человек, 2007 г.

ВЫБОРКА

любое подмножество элементов изучаемой *ген. совокупности*, отобранных для наблюдения. В контексте проведения опросов и обследований ген. совокупность – обычно многочисленное, но конечное множество реально существующих элементов, обладающих рядом представляющих интерес характеристик, которое полностью охватывает изучаемое социально-экономическое явление. Элементами ген. совокупности наиболее часто являются индивиды, домохоз-ва, пр-тия, а также могут быть терр. единицы и др., что может быть строго определено. В. непосредственно отбирается из основы В., т.е. из составленного организатором обследования списка относящихся к ген. совокупности элементов с базовой информацией. Под базовой информацией понимается набор характеристик, известных до проведения обследования для каждого элемента основы выборки. Такими характеристиками могут быть, напр., наименование орг-ции (юридического лица), адрес места нахождения, контактный телефон, вид осуществляемой экономической деятельности, численность персонала и пр. В определении В., которое было дано выше, не использовалось понятие вероятности. Причина этого состоит в том, что известных и используемых на практике способов формирования В. достаточно мно-

го. Причём только в некоторых из них присутствует элемент случайного отбора. Вероятностная или случайная В. предполагает такую процедуру отбора, при которой каждый элемент ген. совокупности имеет известный неравный нулю шанс оказаться включённым в выборку. Детально разработано и обычно применяется для проведения многоцелевых обследований ограниченное число вариантов вероятностной В., что не ограничивает широты и надёжности применения выборочного метода. Базовый вариант – *В. простая случайная*, которая также часто применяется для непосредственного отбора элементов на конечной стадии формирования более сложной В. Важность вероятностной В. состоит в том, что она позволяет научно обоснованно рассчитать ошибки В. И доверительные интервалы для статистик, вычисленных по данным самой В. Также она позволяет проверять критерии и делать статистически значимые выводы о ген. совокупности на основе выборочных результатов. Для неслучайных В. выборочные результаты (какими бы аккуратными они не представлялись) научно обоснованно можно применять только к совокупности элементов самой В., но ни к какой-либо большей группе объектов. В теории В. рассматриваются вероятностные В., отбираемые из конечной ген. совокупности, включающей некоторое число N различимых и опознаваемых между собой элементов или единиц. Общее число выборок объема n, которые могут быть извлечены из ген. совокупности объёма N, равно числу различных сочетаний элементов совокупности по n единиц (C_N^n).

Вероятностная схема формирования списка выборочной совокупности – план (дизайн) В. Формально план В. $\mathrm{p(s)}$ можно определить как *закон распределения вероятностей* отбора всех непустых подмножеств элементов ген. совокупности $\{U\}$, такой что

$$\forall s \in \{s\} \ \mathrm{p(s)} \geq 0 \text{ и } \sum_{s \subset \{U\}} p(s) = 1,$$

где s – В.

Для формирования списка элементов В. определяются приемлемый объём В., зависящий от имеющихся ресурсов, и схема случайного отбора, которая приведет к формированию В. с наилучшими свойствами в смысле обеспечения наименьшего уровня ошибок оценок, связанных с выборкой. Этот процесс называется планированием В. Хотя В. используется для многих целей, чаще всего интерес представляют такие показатели, как среднее или суммарное значение наблюдаемого признака, отношение суммарных или средних значений, а также доля единиц в совокупности, отвечающих некоторому критерию. Оценивание параметров ген. совокупности по данным В. основывается на следующей базовой формуле оценки суммарного показателя, называемой π-оценкой:

$$\hat{Y}_\pi = \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \sum_{k=1}^{n} w_k y_k ;$$

здесь y_k – значение признака элемента *k*; π_k – вероятность включения элемента *k* в выборку;

$w_k = 1/\pi_k$ – выборочный вес элемента *k*;

$$Y = \sum_{k=1}^{N} y_k$$

– суммарный показатель признака *y* по ген.совокупности объёмом N; $\hat{Y}_\pi$ – оценка по данным В. объёмом n значения суммарного показателя признака y.

Соответственно оценка среднего значения признака y равна отношению оценки суммарного показателя *($\hat{Y}$)* и объёма ген. совокупности *(N)*, если он точно известен. В противном случае объём ген. совокупности можно оценить по базовой формуле π-оценки как суммарный показатель признака тождественно равного «1» для всех элементов ген. совокупности.

Дисперсию базовой π-оценки суммарного показателя оценивают по данным В. с помощью общей формулы:

$$Var\left(\hat{Y}_\pi\right) = \sum_{k=1}^{n} \frac{y_k^2}{\pi_k^2}(1-\pi_k) + \sum_{k=1}^{n} \sum_{\substack{l=1 \\ l \neq k}}^{n} \frac{y_k y_l}{\pi_k \pi_l} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}},$$

где π_{kl} – совместная вероятность включения пары элементов *k* в *l* в выборку.

Важно подчеркнуть, что приведённые выражения оценки суммарного показателя и оценки дисперсии этой оценки – универсальны, т.е. применимы к данным любой вероятностной В.

Для их использования помимо данных В. необходимо знать вероятности включения элементов в выборку (π_k), а также парные вероятности включения (π_{kl}), что не очень удобно. Поэтому для каждого конкретного плана В., применяемого на практике, в теории получены более простые и удобные индивидуальные выражения для вычисления дисперсии оценок.

Достоверность результатов, получаемых по В., основывается на известном распределении значений оценки изучаемого параметра по множеству всевозможных выборок. В случае конечной ген. совокупности применима Центральная придельная теорема. А именно, при достаточно большом объёме В. оценки параметров, рассчитанные по выборке, близки к истинным значениям, причем ошибка В., т.е. отклонение оценки от истинного значения, распределена приблизительно по нормальному закону распределения. Поэтому по данным В. может быть рассчитан доверительный интервал, в пределах которого с заданной вероятностью (обычно 0,95 или 0,9) находится истинное для ген. совокупности значение оцениваемого параметра. Половина длины доверительного интервала называется предельной ошибкой В. Это величина, которую с заданной доверительной вероятностью не превышает отклонение рассчитанной по В. оценки параметра от его истинного значения. Кроме этого, имеется возможность определения необходимого объёма В. для обеспечения требующейся точности результатов, т.е. чтобы значение предельной ошибки В. не превосходило заданной величины.

Напр., в случае простой случайной В. достоверность результатов (L – предельная ошибка В.) оценки доли (P) признака в совокупности и необходимый для этого объём В. (n) связаны приближенным соотношением:

$$n \cong \frac{4p(1-p)}{L^2}.$$

Квадратичная функция $y = p(1-p)$ достигает своего абсолютного максимума при значении доли $p = 0{,}5$. С учётом этого на основе приведенного соотношения можно вычислить, что при фиксированной величине предельной ошибки В.: L = 0,01 и любой истинной доле, в т.ч. при P = 0,5, нужный объём В. составляет не более чем n = 10000 элементов для наблюдения. Также простая случайная В. объёмом n =2500 единиц при истинном значении оцениваемой доли $p = 0{,}5$ может обеспечить только 2%-ую точность оценки по данным В. (L = 0,02).

См. также *Выборка* в разделе 1. Рубрика 1.2.2. Теория статистического наблюдения. Этапы статистического исследования

ВЫБОРКА КЛАСТЕРНАЯ

простой случайный отбор групп элементов ген. совокупности. Такие единицы принято называть *кластерами*. Все элементы отобранных кластеров подлежат наблюдению. Предпосылкой для применения этого метода выборочного обследования может служить отсутствие актуальной списочной основы выборки (эффективного списка элементов ген. совокупности), причем её создание оказывается весьма трудоёмкой и дорогостоящей операцией. Если кластеры образованы исходя из терр. близости элементов, то в этом случае достигается снижение транспортных расходов при интервьюировании респондентов, а сама выборка называется терр. Кроме простого случайного метода непосредственный отбор кластеров может осуществляться любым другим методов случайной выборки – расслоенным, с вероятностью пропорциональной размеру в данном случае кластеров и т.д. Распространение данных выборки на ген. совокупность осуществляется соответственно использованному методу отбора. Причём эффективность в смысле точности результатов кластерной выборки обычно ниже, чем при простой случайной. Это обычно связно с различием кластеров по размеру и по величине значений показателей обследования. Однако указанные выше преимущества, в т.ч. сокращение затрат на создание основы выборки и транспортных расходов, обычно оказываются существенно более весомыми. В.к. широко применяется при выборочных обследованиях, проводимых на транспорте (кластеры – транспортные единицы), в сел. хоз-ве (кластеры – терр. единицы с.-х. угодий), в обследованиях

нас. (кластеры – населённые пункты), а также для проведения глубоких аналитических исследований. Последнее связано с преимуществами сплошного охвата элементов в каждом отобранном кластере.

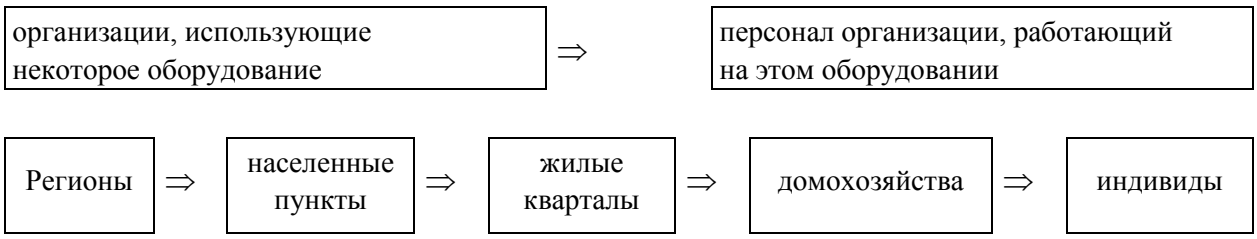
ВЫБОРКА КЛАСТЕРНАЯ СЛУЧАЙНАЯ

см. в ст. Выборка кластерная.

ВЫБОРКА МНОГОЭТАПНАЯ СЛУЧАЙНАЯ

метод формирования выборки, который является обобщением метода *выборки кластерной*. Если объективные причины обуславливают проведение обследования с помощью кластерной выборки, то может оказаться удобным в целях экономии ресурсов не проводить наблюдение всех элементов отобранных в выборку кластеров на первом этапе, а в каждом из них независимо отобрать случайную выборку из списка элементов данного кластера. И, тем самым, осуществить двухэтапный отбор. Этот процесс случайного двухэтапного отбора может быть продолжен, приводя к трехэтапному отбору и т.д. Число этапов отбора, предусмотренных планом выборки, обычно зависит от того, когда становится возможным с учётом ресурсных ограничений создание списков элементов ген. совокупности в отобранных на очередном этапе кластерах. Т.о., на всех этапах отбора, кроме последнего, выполняется отбор кластеров в выборку соответствующего этапа. На последнем этапе может быть отобрана выборка элементов ген. совокупности в каждом кластере предпоследнего этапа отбора, а может быть осуществлен сплошной охват элементов в этих кластерах. В последнем случае обычно план выборки называется многоэтапной кластерной выборкой.

Примерами кластеров на этапах отбора и конечными элементами генеральной совокупности могут служить следующие единицы многоэтапной выборки:



Формально к методу многоэтапной выборки предъявляется требование инвариантности. Во-первых, на текущем этапе план выборки не зависит от предыдущих этапов отбора. И, во-вторых, отбор в каждом кластере производится независимо от состава выборки кластеров на данном этапе и от отбора в других кластерах. Если условие инвариантности не выполняется, то такой план выборки называется многофазным.

Следствием требования инвариантности плана В.м.с. является то, что вероятность включения конечного элемента ген. совокупности в многоэтапную выборку равна произведению вероятностей включения единиц на этапах отбора, содержащих данный элемент. Соответственно, выборочный вес конечного элемента ген. совокупности равен произведению выборочных весов единиц на этапах отбора. Кроме того, имеет место правило разложения дисперсии оценки параметра ген. совокупности в сумму компонентов дисперсии, связанных с методом отбора на каждом этапе формирования многоэтапной выборки. Первый компонент дисперсии, связанный с планов выборки на первом этапе отбора, обычно существенно больше второго, который в свою очередь превосходит третий и т.д. Причём с увеличением объёма выборки на первом этапе уменьшаются также и второй компонент дисперсии и последующие. Увеличение объёма выборки на втором (и любом последующем) этапе отбора уменьшает только второй (или соответствующий последующий) компонент дисперсии оценки. Поэтому на

практике стремятся к тому, чтобы на первом этапе в основе выборки было как можно больше небольших кластеров, и объем выборки на первом этапе был бы по возможности большим. А на последующих этапах – ограниченным. Если кластеры приблизительно одинаковые по размеру, то можно использовать простую или систематическую выборку кластеров. Если же кластеры существенно варьируют по размеру, то на первом этапе применяется *выборка расслоенная* или выборка с вероятностями пропорциональными размеру кластеров.

ВЫБОРКА НЕСЛУЧАЙНАЯ

любое подмножество элементов изучаемой ген. совокупности, предполагающая, что для элементов ген. совокупности невозможно рассчитать вероятности отбора. Поэтому невозможно обоснованно применить статистическую теорию для анализа данных такой выборки. Напр., не имеет содержательного смысла вычислять статистические критерии для проверки гипотез. В.н. даёт информацию только об элементах, самой выборки, но ни о какой большей (ген.) совокупности, из которой выборка была отобрана.

Существуют ситуации, когда использование неслучайной выборки оправдано. Напр., при обследовании ограниченной и хорошо известной исследователю ген. совокупности или при наблюдении так называемых фокус групп в маркетинговых исследованиях. Формирование фокус групп ориентировано на то, чтобы детально изучить мнения или характеристики различных представителей ген. совокупности. Но поскольку размер фокус групп обычно ограничен, использование вероятностной процедуры для отбора участников может не обеспечить представленность всего разнообразия мнений. Тем не менее, использование В.н. ограничивается специальными ситуациями. Обычно их недостатки перевешивают преимущества, такие как удобство отбора элементов и экономия ресурсного обеспечения проведения обследования.

В имеющемся большом числе разновидностей В.н. можно выделить наиболее распространенные типы выборки: удобная, целевая и квотная, методом «снежного кома».

Удобная выборка формируется исходя из удобства исследователя. В частности интервью, взятые репортером у прохожих на улице, обычно основываются на удобной выборке. Сюда же следует отнести и опрос, напр., первых 20 вышедших человек, посетивших концерт. В этом случае нет никакой гарантии, что удобная выборка представительна для изучаемой ген. совокупности. Однако удобные выборки могут использоваться для проведения пилотных обследований, напр., для предварительного тестирования анкет и вопросников, а также для наблюдения фокус групп.

Для формирования квотной выборки ген. совокупность разбивается на важные для целей исследования подгруппы, а затем для каждой подгруппы устанавливается квота на количество представителей в выборке. Квоты обычно устанавливаются в соответствии со структурой ген. совокупности или в соответствии с нужными характеристиками. Напр., количество мужчин и женщин в выборке пропорционально количеству мужчин и женщин в ген. совокупности. Квотная выборка была введена в практику для преодоления общей проблемы, состоящей в том, что случайная выборка может не соответствовать распределению ген. совокупности по ключевым характеристикам. Для обеспечения установленной квоты на этапе сбора сведений интервьюерам предоставляется свобода опрашивать респондентов по собственному усмотрению. Напр., если интервьюеры уже опросили нужное число мужчин, чтобы обеспечить установленную квоту, то опрос мужчин прекращается, а женщин продолжается до момента обеспечения квоты. Существует мнение, что применение квотных выборок оправдано, так как они гарантируют, что труднодостижимые группы респондентов (напр., состоятельные домохоз-ва) будут включены в выборку в пропорции, равной их пропорции в ген. совокупности. Однако при квотной выборке возникает ряд проблем. Поскольку интервьюерам предоставлена возможность самим выбирать респондентов, возможно возникновение смещения в результатах из-за того,

что характеристики респондентов могут влиять на их шансы быть отобранными. В отсутствии строгих инструкций (применение которых, в любом случае, было бы весьма проблематичным) интервьюеры могут выбирать более дружелюбных людей или тех, кто находится дома в определённое время. Кроме этого, при использовании квотной выборки оказывается замаскированной проблема недоступности или отказа респондентов предоставлять сведения. Таких респондентов фактически заменяют другими, которые могут от них отличатся по существенным для обследования характеристикам. И, главное, поскольку на последнем этапе формирования выборки не используется вероятностный отбор, то расчёт стандартных ошибок статистик оказывается практически невозможным.

При формировании списка элементов целевой (по суждению) выборки исследователь отбирает каждый элемент по своему усмотрению, основываясь на критериях важных в этом исследовании. Напр., если изучаются орг-ции (юридические лица), осуществляющие деятельность в некотором городе, то может понадобиться включить в выборку наиболее значимые для данного населённого пункта, т.е. градообразующие орг-ции. Чаще всего целевые выборки используются для предварительных тестовых опросов, в которых важно протестировать вопросник на различных респондентах, представляющих различные мнения и жизненный опыт.

Выборка по принципу «снежного кома» (по рекомендации). Для труднодостижимых элементов некоторых ген. совокупностей может использоваться такая техника отбора, при которой сначала целенаправленно ведется поиск ограниченного числа представителей ген. совокупности, а затем у найденных представителей выясняется информация о других представителях этой ген. совокупности. У них, в свою очередь, также запрашивают сведения о других представителях ген. совокупности и так далее. Ясно, что такая выборка не является вероятностной, но если требуется исследовать такие социальные группы, как представители криминального мира, наркоманы, алкоголики, а также ген. совокупность домохоз-в с высоким уровнем доходов, списки которой сложно или невозможно создать, такой подход может оказаться единственным. Представители такой ген. совокупности должны быть связаны друг с другом (должны знать друг друга). Выборка по принципу «снежного кома» наиболее эффективно применяется в пилотных исследованиях и при изучении небольших ген. совокупностей.

В современной практике проведения крупномасштабных многомерных и многоцелевых выборочных обследований безусловное предпочтение отдается вероятностной выборке. При таком подходе исключается влияние субъективного фактора на результаты выборочного обследования. В отличие от неслучайной вероятностная выборка позволяет научно обоснованно оценить параметры ген. совокупности и рассчитать ошибку выборки для оцениваемых статистик.

ВЫБОРКА ПРОСТАЯ СЛУЧАЙНАЯ

выборка, план которой предусматривает, что только каждая выборка фиксированного объёма (n) может быть отобрана с равной вероятностью из конечной *ген. совокупности* объёма (N). В этом случае вероятности включения в выборку (π_k) для всех единиц совокупности объёма (N) равны между собой:

$$\pi_k = \frac{n}{N}, \text{где } k = 1, \dots, N.$$

На практике В.п.с. получают, отбирая последовательно единицу за единицей. Единицы в совокупности нумеруются числами от 1 до N, после чего случайно выбирается число, заключенное между 1 и N. Единица совокупности, имеющая этот номер, включается в выборку. На каждом последующем шаге уже отобранные номера исключаются из списка, поскольку иначе одна и та же единица могла бы попасть в выборку более одного раза. Поэтому такой отбор называется отбором без возвращения. Отбор с возвращением легко осуществим, но им, за исключением особых случаев, пользуются редко, поскольку уже опрошенная единица не может предоставить дополнительную полезную информацию. Кроме того известно, что выборка без возвращения более эффективна (дисперс-

сия оценок меньше), т.е. её результаты более точные.

При простом случайном отборе для оценки среднего значения признака совокупности используют выборочное среднее ($\bar{y}$), а точность характеризуется оценкой дисперсии оценки среднего признака ($\mathrm{var}(\bar{y})$), рассчитанной по выборке:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

и $\mathrm{var}(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$,

где N – количество элементов ген. совокупности;

n – количество элементов выборочной совокупности;

y_i – значение признака у i-го элемента, i = 1,2… n;

s^2 – дисперсия признака у в выборке.

ВЫБОРКА РАССЛОЕННАЯ

выборка, предусматривающая предварительное расслоение (стратификацию) *ген. совокупности* на подсовокупности, называемые слоями или стратами. После завершения расслоения в каждом слое независимо осуществляется простой случайный отбор элементов. Расслоение осуществляется на основе имеющейся пообъектной базовой информации – категориальные и/или количественные признаки, значения которых известны для каждого элемента основы выборки до проведения обследования.

В случае обследования орг-ций розничной торг. в качестве базовой информации для расслоения могут использоваться такие характеристики, как адреса места расположения торговых дислокаций, размер торговой пл., численность персонала и т.п. В результате расслоения в слои объединяются сходные по базовым характеристикам элементы основы выборки. Однако в отличие от *кластерного анализа* при расслоении слои определяются такими, чтобы минимизировать дисперсию оценки представляющего интерес показателей ген. совокупности. Применение В. Р. случайной может дать существенный (на 1–2 порядка) выигрыш в точности оценивания параметров ген. совокупности по сравнению с простой случайной выборкой. Условием этого является коррелированность переменных расслоения с изучаемыми признаками. Т.о. при В. Р. случайной ген. совокупность, содержащая *n* единиц, сначала расслаивается на h слоев, состоящих соответственно из N_1, N_2, …, N_H единиц. Слои не содержат общих единиц и вместе исчерпывают всю ген. совокупность, так что

$N_1 + N_2 + \ldots + N_H$ = n.

В каждом слое независимо отбирается выборка рассчитанного объёма: n_1, n_2, …, n_H.

Так что $n_1 + n_2 + \ldots + n_H$ = n.

Поэтому в случае расслоенной выборки вероятность включения единицы k в выборку (π_k) зависит от слоя, к которому принадлежит данная единица:

$\pi_k = \frac{n_h}{N_h}$, если единица k включена в слой h.

На практике используется несколько вариантов размещения общего объёма выборки (*n*) по слоям. Пропорциональное, при котором в выборку из каждого слоя отбирается столько элементов

$(n_h, h = 1,\ldots,H)$,

какова доля объёма слоя

$(N_h, h = 1,\ldots,H)$

в объёме ген. совокупности (*N*):

$$\frac{n_h}{n} = \frac{N_h}{N}, h = 1,\ldots,H.$$

Отсюда следует, что все элементы ген. совокупности при В.р. с пропорциональным размещением по слоям имеют одинаковую вероятность быть отобранными. Если элемент *k* принадлежит слою *h*, то вероятность включения этого элемента в выборку:

$$\pi_k = \frac{n_h}{N_h} = \frac{n}{N}.$$

Следовательно, равны выборочные веса всех элементов:

$$w_h = \frac{N}{n}, k = 1,\ldots,N.$$

Такие выборки называются равновзвешенными.

Другой вариант размещения объёма выборки по слоям – оптимальный, при котором достигается мин. дисперсии оценки параметра ген. совокупности, когда фиксированы расслоение и объём выборки (или стоимость обследования). Также имеются еще несколько вариантов оптимального размещения, в т.ч. т.н. степенное, обеспечивающее одинаковую точность результатов выборочного обследования по целевым группам элементов ген. совокупности. В случае В.р. для оценки среднего значения признака совокупности используется формула среднего взвешенного ($\bar{y}_{st}$), а точность оценивания характеризуется соответствующей оценкой дисперсии оценки среднего признака ($\mathrm{var}(\bar{y}_{st})$), рассчитанной по выборке:

$$y_{st} = \frac{1}{N}\sum_{h=1}^{H} N_h y_h$$

и $$\mathrm{var}(y_{st}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h (N_h - n_h)\frac{s_h^2}{n_h},$$

где H – число слоев;

Nh – объём h-го слоя ген. совокупности;

y_h – среднее значение признака y в h-м слое выборки;

N – объём ген. совокупности;

nh – объём h-го слоя выборки;

s_h^2 – $s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - y_h)^2$.

ВЫБОРКА РАССЛОЕННАЯ СЛУЧАЙНАЯ

см. в ст. Выборка расслоенная

ВЫБОРКА СИСТЕМАТИЧЕСКАЯ

отбор из списка основы каждого k-го элемента, начиная с первого, который отбирается случайно. Величина шага отбора k выбирается т.о., чтобы количество отобранных элементов было равно требуемому объёму выборки (*n*). Хотя свойства статистик В.с. несколько сложнее, в большинстве случаев (если элементы в основе выборки располагались в случайной последовательности, напр., индивиды упорядочены по фамилиям в алфавитной последовательности) можно предполагать с достаточной степенью уверенности, что В.с. случайная эквивалентна *выборке простой случайной* того же объёма.

Напр., в случае формирования систематической выборки объёмом 500 элементов из ген. совокупности объёмом 15000 элементов, доля выборки составляет 500/15000 или 1/30. Поэтому шаг отбора k выбирается равным 30. Далее случайным образом выбираем целое число в промежутке от 1 до 30. Пусть этим числом оказалось 14. Тогда элемент с номером 14 в списке включаем в выборку в качестве первого элемента. После этого последовательно прибавляем шаг k (30) к последнему отобранному номеру и получаем следующий номер элемента в списке, который включается в выборку. В результате в выборку будут отобраны

14, 14 + 30 = 44, 44 + 30 = 74

и т.д. Элемент. Эта процедура отбора продолжается пока не достигнут конец списка.

Часто результатом деления объёма ген. совокупности на объём выборки оказывается не целым числом. Для таких случаев разработаны технические приемы (алгоритмы), обобщающие описанный выше алгоритм на указанную ситуацию.

ВЫБОРКА СИСТЕМАТИЧЕСКАЯ СЛУЧАЙНАЯ

см. в ст. Выборка систематическая

ВЫБОРКА ЭЛЕМЕНТОВ

с вероятностями пропорциональными их размеру**.** Если в основе выборки имеется вспомогательный признак, значения которого задают размер элементов ген. совокупности, то отбор в В.э. можно осуществить с неравными вероятностями пропорциональными размеру. Вероятности включения элементов в выборку в случае отбора без возвращения задаются соотношением:

$$\pi_k = n\frac{x_k}{\sum_{i=1}^{N} x_i},$$

здесь π_k – вероятность включения элемента *k* в выборку; x_k – значение вспомогательного при-

знака (x), характеризующего размер элемента k; N – объём ген. совокупности; n – фиксированный объём выборки.

При этом вероятности включения каждого элемента k пропорциональны размеру x_k, так как остальные величины в приведённом соотношения являются константами.

Этот метод отбора позволяет повысить точность оценивания, если используемый для определения вероятностей вспомогательный признак приблизительно пропорционален изучаемым характеристикам. Выборку такого типа выгодно использовать, когда имеются отдельные элементы с большими значениями признака.

Данный метод случайной выборки может использоваться, напр., при проведении обследований пр-тий розничной торг. В этом случае пр-тия, имеющие большую численность работников или торговую пл., обычно также имеют больший розничный товарооборот. Другим примером широкого использования В.э. с вероятностью пропорциональной их служат обследования нас., в которых план выборки на первом этапе обычно предусматривает отбор терр. кластеров. При этом вероятности включения терр. единиц в выборку пропорциональны численности проживающего в них нас.

В случае выборки элементов с вероятностями пропорциональными их размеру для оценивания показателей обследования используются базовые формулы π-оценки. Однако вычисление соответствующих характеристик точности (оценки дисперсии оценки параметра) на практике выполняется на основе аппроксимационных формул. См. также *Выборка кластерная.*

ВЫБОРОЧНАЯ ДИСПЕРСИЯ

характеристика изменчивости (вариации) случайной величины X, определяемая как среднее арифметическое квадратов отклонений наблюдаемых значений $x_1,\dots x_n$ этой величины от их среднего арифметического:

$$\hat{D}(x)=\frac{\sum_{i=1}^{m}(x_i-\overline{X})^2 n_i}{n}=\sum_{i=1}^{m}(x_i-\overline{X})^2\,\hat{p}_i,$$

где $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$.

Если данные наблюдений представлены в виде дискретного ряда, где $x_1,x_2,\dots,x_m$ – наблюдаемые варианты, а $n_1,n_2,\dots,n_m$ – соответствующие им частоты, причем

$$\sum_{i=1}^{m}n_i=n,$$

то В.д. определяется по формуле:

$$\hat{D}(x)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})^2,$$

где $\overline{X}=\frac{\sum_{i=1}^{m}x_i n_i}{n}=\sum_{i=1}^{m}x_i\,\hat{p}_i,$

а $\hat{p}_i=\frac{n_i}{n}$

– выборочная относительная частота.

На практике чаще применяют упрощённую формулу вычисления дисперсии:

$\hat{\ }(\ )=(\overline{\ }^2)\ -(\overline{\ })^2$, где $\overline{(X^2)}=\frac{1}{n}\sum_{i=1}^{n}x_i^{\,2}$.

В.д. является состоятельной, но смещённой оценкой *ген. дисперсии*. На практике чаще используют, т.н., исправленную В.д., которая обладает свойствами состоятельности и несмещённости, но не является эффективной:

$$s^2=\frac{\sum_{i=1}^{m}(x_i-\overline{X})^2}{n-1}=\frac{n}{n-1}\hat{D}(x).$$

ВЫБОРОЧНАЯ ЧАСТОТА СОБЫТИЯ

число n_i, равное количеству повторений значения x_i в выборке. Если всего в выборке k разных значений, то $n_1+n_2+\dots+n_k=n$, где n – объём выборки. Если изучаемая случайная величина является непрерывной, то отдельные значения наблюдаемых данных могут как угодно мало отличаться друг от друга и поэтому в выборке одинаковые значения величины могут встречаться редко, а частоты вариантов мало отличаются друг от друга (практически все равны 1). В подобных случаях строят интервальный *вариационный ряд*: располагают дан-

ные в порядке неубывания, весь интервал варьирования наблюдаемых значений случайной величины разбивают на ряд частичных интервалов и подсчитывают число попаданий значений величины в каждый частичный интервал. Эти числа и будут являться В.ч.с. – n_i, но уже не конкретного значения варианта, а частотами попадания выборочных данных в частичный интервал. В.ч.с. используется для графического изображения вариационных рядов. Наряду с В.ч.с. на практике рассматривают выборочную относительную частоту, или частость, представляющую отношение числа наблюдений n_i в выборке, в точности равных x_i (или попавших в i-й частичный интервал) , к общему объёму выборки *n*:

$$\hat{p}_i = \frac{n_i}{n}.$$

Теорема Бернулли даёт теоретическое обоснование замены неизвестной вероятности события его относительной частотой. Выборочная относительная частота является состоятельной, несмещённой и эффективной оценкой вероятности события.

См. также *Бернулли теорема*.

ВЫБОРОЧНЫЕ ХАРАКТЕРИСТИКИ

характеристики распределения *случайной величины*, вычисляемые по *выборке*, являющиеся статистическими оценками параметров распределения *ген. совокупности*. К осн. В.х., представляющим наибольшую информа-цию о ген. совокупности, из которой получена выборка, относятся: выборочные функции распределения $\hat{F}(x)$ и плотности распределения $\hat{f}(x)$, выборочная относительная частота $\hat{p}_i$ появления *i*-го возможного значения x_i дискретной случайной величины, выборочные начальные и центральные моменты.

Рассмотрим выборочную совокупность значений некоторой случайной величины *X* объема *n,* и каждому варианту из этой совокупности поставим в соответствие его частоту. Если обозначить через *x* некоторое значение случайной величины *X*, а через n_x – число выборочных значений случайной величины *X*, меньших *x*, то число n_x/n – будет относительной частотой появления события $X < x$, его еще называют накопленной частотой. При изменении *x* в общем случае будет изменяться и величина n_x/n. Это означает, что накопленная частота n_x/n является функцией аргумента *x*. Выборочной функцией распределения называется функция $\hat{F}(x)$, задающая для каждого значения *x* относительную частоту события

$X < x$, т.е. $\hat{F}(x) = n_x/n$.

Свойство статистической устойчивости частоты, обоснованное *Бернулли теоремой*, оправдывает целесообразность использования функции $\hat{F}(x)$ при больших значениях *n* в качестве приближённого значения неизвестной функции $F(x)$. Выборочная функция распределения $\hat{F}(x)$ обладает всеми теми же свойствами, что и теоретическая:

$0 \le \hat{F}(x) \le 1$, $\hat{F}(x)$ – неубывающая функция,

$\hat{F}(-\infty) = 0$, $\hat{F}(\infty) = 1$.

Выборочной относительной частотой $\hat{p}_i$ дискретной случайной величины называется отношение числа наблюдений n_i в выборке, в точности равных x_i, к общему объёму выборки *n*: $\hat{p}_i = \frac{n_i}{n}$.

Если данные, наблюдаемые над непрерывной случайной величиной, представлены в виде интервального *вариационного ряда*, то можно определить выборочный аналог дифференциальной функции распределения $f(x)$ - выборочную функцию плотности вероятности:

$$\hat{f}(x) = \frac{\hat{F}(x+\Delta x) - \hat{F}(x)}{\Delta x},$$

где $\hat{F}(x+\Delta x) - \hat{F}(x)$ – относительная частота попадания наблюдаемых значений случайной величины *X* в интервал $[x, x+\Delta x[$.

Для выявления особенности поведения выборочной совокупности значений случайной величины выделяют некоторые постоянные, которые представляют выборку в целом и отражают присущие ей закономерности. Постоянные, вокруг которых концентрируются остальные результаты наблюдений, называются средними величинами. К ним относятся среднее арифметическое,

среднее геометрическое, среднее гармоническое, мода и медиана.

Для характеристики изменчивости наблюдаемых данных, служат показатели вариации: размах варьирования, дисперсия, среднее квадратическое отклонение и т.д.

Среднее арифметическое и *выборочная дисперсия* являются частным случаем более общего понятия – выборочного момента.

Начальным выборочным моментом порядка *k* называется среднее арифметическое *k*-х степеней наблюдаемых значений случайной величины:

$$\nu_k = \frac{\sum_{i=1}^{m} x_i^k n_i}{\sum_{i=1}^{m} n_i}.$$

Из определения следует, что начальный выборочный момент нулевого порядка $\hat{\nu}_0 = 1,$

а начальный выборочный момент первого порядка $\hat{\nu}_1 = \overline{X}$ – средняя арифметическая.

Центральным выборочным моментом порядка *k* называется среднее арифметическое *k*-х степеней отклонений наблюдаемых значений случайной величины от их среднего арифметического:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{m} (x_i - \overline{X})^k n_i}{\sum_{i=1}^{m} n_i}.$$

Из определения следует, что центральный выборочный момент нулевого порядка $\hat{\mu}_0 = 1,$ первого порядка $\hat{\mu}_1 = 0,$ второго порядка

$\hat{\mu}_2 = \hat{D}(X)$.

В качестве характеристик формы распределения (графика плотности вероятности или полигона) используются коэффициенты асимметрии и эксцесса. Выборочным коэффициентом асимметрии называется величина $\hat{A}$, вычисляемая по формуле:

$$\hat{A} = \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}} = \frac{\hat{\mu}_3}{\hat{\sigma}^3}.$$

Для симметричного распределения $\hat{A} \approx 0$. Если распределение асимметрично, то одна из ветвей его графика плотности вероятностей начиная с вершины имеет более пологий «спуск»: при $\hat{A} < 0$ – слева, а при $\hat{A} > 0$ – справа.

Выборочный эксцесс $\hat{E}$ служит для сравнения на «крутость» выборочного распределения с нормальным распределением. Он вычисляется по формуле:

$$\hat{E} = \frac{\hat{\mu}_4}{\hat{\mu}_2^2} - 3 = \frac{\hat{\mu}_4}{\hat{\sigma}^4} - 3.$$

Для более островершинного по сравнению с нормальным распределением $\hat{E} > 0$, а для плосковершинного – $\hat{E} < 0$.

При неограниченном увеличении объёма выборки $(n \to \infty)$ все В.х. стремятся по вероятности к соответствующим теоретическим характеристикам. Это даёт нам основание использовать выборочные характеристики для приблизительного описания свойств всей ген. совокупности. Математическим основанием этого факта служат различные формы *закона больших чисел*, который позволяет теоретически обосновать устойчивость основных выборочных характеристик распределения.

Г

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ

в математической статистике понятие Г.с. трактуется как совокупность всех мыслимых наблюдений, которые могли бы быть произведены при данном реальном комплексе условий. Так, обследовав даже все пр-тия некоторой отрасли по определённым показателям, мы можем рассматривать обследованную совокупность лишь как представителя гипотетически возможной более широкой совокупности пртий, которые могли бы функционировать в рамках того же реального комплекса условий.

Г.с – множество каких-либо однородных элементов (характеристик), из которого по определённому правилу выделяется некоторое подмножество, называемое *выборкой*.

Г.с. называется конечной или бесконечной в зависимости от того, конечна или бесконечна совокупность составляющих её элементов. Если число элементов исследуемого множества конечно, то Г.с. конечна. В противном случае – бесконечна. Напр., по статистическим данным

оценивается доля мальчиков среди детей, родившихся за год. Все родившиеся за год дети – конечная Г.с. Бесконечное непрерывное воспроиз-во нас. – бесконечная Г.с.

ГИПОТЕЗА СТАТИСТИЧЕСКАЯ

всякое высказывание, проверяемое по выборке (результатам наблюдений), о вероятностных закономерностях, которым подчиняется изучаемое явление. Примеры Г.с. – следующие высказывания: *ген. совокупность*, о которой мы располагаем лишь выборочными сведениями, имеет нормальный закон распределения или ген. средняя (математическое ожидание случайной величины) равно 5. Не располагая сведениями о всей ген. совокупности, высказанную гипотезу сопоставляют, по определённым правилам, с выборочными сведениями и делают вывод о том, можно принять гипотезу или нет. Процедура обоснованного сопоставления высказанной гипотезы с имеющимися в нашем распоряжении выборочными данными называется *статистической проверкой гипотез*. Правило, в соответствии с которым принимается или отклоняется данная гипотеза, называется *критерием статистическим.* Гипотеза называется параметрической, если в ней содержится некоторое утверждение о значении параметров распределения известного вида. Непараметрической называется гипотеза, в которой высказывается предположение обо всём распределении (о его виде или его общих свойствах). Гипотеза называется простой, если ей соответствует одно распределение наблюдений или одна точка пространства параметров; гипотеза называется сложной, если она сводится к выбору к.-л. распределения из целого множества или точке из интервала (конечного или бесконечного). Сложная Г.с. может быть представлена в виде множества простых Г.с. Напр., гипотеза о том, что математическое ожидание нормально распределенных величин – результатов наблюдения – равно некоторому конкретному значению a_0, является простой гипотезой ($a = a_0$). Гипотеза же $a \geq a_0$ будет сложной, составленной из простых гипотез

$$\{a = a^*, \quad a^* \in [a_0, \infty)\},$$

ей соответствует какое-либо значение параметра a из множества – полуинтервала $[a_0, \infty)$. Проверяемую гипотезу обычно называют нулевой (или основной) и обозначают H_0. Наряду с нулевой гипотезой H_0 рассматривают *альтернативную гипотезу*, или конкурирующую гипотезу H_1, являющуюся логическим отрицанием H_0. По своему прикладному содержанию гипотезы можно подразделить на несколько осн. видов. Гипотезы о типе закона распределения исследуемой случайной величины X – предположение о согласованности выборочного распределения $F(X)$ и некоторого гипотетического (модельного) $F^*(X)$. Гипотеза выглядит следующим образом:

$$H_0: \quad F(X) \equiv F^*(X),$$

где предполагаемая модельная функция может быть как заданной однозначно, т.е.

$$F^*(X) = F_0(X),$$

так и заданной с точностью до принадлежности к некоторому параметрическому семейству. В последнем случае

$$F^*(X) = F_0(X, \theta),$$

где – некоторый параметр (возможно, k - мерный), значения которого неизвестны, но могут быть определены по выборке с помощью специальных методов.

Гипотезы об однородности двух или нескольких выборок предполагают проверку утверждений о равенстве функций распределений, математических ожиданий, дисперсий или других характеристик k различных выборок:

$$H_0: \quad F_1(X) = F_2(X) = \ldots = F_k(X),$$
$$H_0: \qquad a_1 = a_2 = \ldots = a_k,$$
$$H_0: \qquad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2.$$

Если в результате проверки гипотеза H_0 не отвергается, то говорят, что соответствующие выборочные характеристики различаются статистически незначимо. В частном случае, когда число выборок $k=2$, и одна из выборок содержит малое количество наблюдений, (может быть всего одно) гипотеза первого из вышеперечисленных типов означает проверку аномальности одного или нескольких резко выде-

ляющихся наблюдений. Гипотезы о числовых значениях параметров исследуемой ген. совокупности состоят в том, что некоторый параметр θ распределения исследуемой совокупности, напр., среднее значение, дисперсия и т.п., имеет наперёд заданное значение θ_0 – проверяемая гипотеза

$$H_0: \quad \theta = \theta_0,$$

или множество значений Θ – проверяемая гипотеза

$$H_0: \quad \theta \in \Theta.$$

Гипотезы независимости двух и более выборок с одинаковыми распределениями и случайности выборочных значений. Эта гипотеза возникает в задачах, в которых элементами выборки являются измеренные значения к.-л. признака при изменяющихся условиях эксперимента, и требуется проверить гипотезу о влиянии на результаты измерений этих изменений. Применительно к двум выборкам объёма *n* нулевая гипотеза об их взаимной независимости соответствует соотношению

$$F(x_i, y_i) = F_X(x_i) \cdot F_Y(y_i) \quad (i = 1,2,\ldots,n),$$

где $F(x_i, y_i)$ – неизвестная функция распределения двумерной случайной величины, $F_X(x_i)$ и $F_Y(y_i)$ – некоторые одномерные функции распределения двух выборок. Нулевая гипотеза случайности выборки отвечает выражению

$$F(x_1, x_2, \ldots, x_n) = F(x_1) \cdot F(x_2) \cdots F(x_n)$$

и означает проверку того, что компоненты x_i независимы и одинаково распределены.

Гипотезы об общем виде модели, описывающей статистическую зависимость между признаками. В качестве гипотетических могут проверяться утверждения о линейном, квадратическом, экспоненциальном и т.п. типе зависимости.

См. также *Корреляционный анализ*, *Регрессионный анализ*.

ГИПОТЕЗА СТАТИСТИЧЕСКАЯ АЛЬТЕРНАТИВНАЯ

см. в ст. Альтернативная гипотеза.

ГИПОТЕЗА СТАТИСТИЧЕСКАЯ НУЛЕВАЯ (ОСНОВНАЯ)

см. в ст. Нулевая гипотеза

ГИПОТЕЗА СТАТИСТИЧЕСКАЯ ПРОСТАЯ

гипотеза, которой соответствует одно распределение наблюдений или одна точка пространства параметров. Напр., Г.с.п. непараметрическая полностью определяет теоретическую функцию распределения $F(x)$. Так, простыми будут гипотезы: «вероятность успеха в схеме Бернулли равна $1/2$», «теоретическая функция распределения является нормальной с нулевым средним и дисперсией, равной 2».

См. также *Гипотеза статистическая*.

ГИПОТЕЗА СТАТИСТИЧЕСКАЯ СЛОЖНАЯ

гипотеза, которая предполагает выбор какого-либо одного из целого множества распределений или точки из интервала (конечного или бесконечного). Примеры сложных гипотез: «вероятность успеха в схеме Бернулли заключена между 0,4 и 0,7», «теоретическая функция распределения является нормальной с нулевым средним и произвольной дисперсией», «теоретическая функция распределения не является нормальной». Непараметрическая Г.с.с. выделяет среди всех возможных функций распределения некоторое подмножество F_0 , содержащее более одной функции. Г.с.с параметрическая предполагает выделение среди некоторого параметрического семейства функций распределения $F(x, \theta)$ тех, у которых неизвестный параметр $\theta \in \Theta_0$, где Θ_0 – некоторое подмножество области Θ всех возможных значений неизвестного параметра θ. Г.с.с. может быть представлена в виде множества простых статистических гипотез.См. также *Гипотеза статистическая*.

ГИСТОГРАММА

один из видов графического представления эмпирического распределения. Г. используется

для изображения интервального *вариационного ряда*, и представляет собой столбчатую диаграмму. Для построения Г. все множество значений результатов наблюдений $X_1,\ldots,X_n$ делится на k интервалов группировки точками $x_0,\ldots,x_k$ (обычно интервалы выбирают равными), затем подсчитывается частоты n_i попадания наблюдений в каждый из интервалов $[x_{i-1},x_i)$ и относительные частоты $\hat{p}_i = n_i/n$.

На оси абсцисс отмечаются точки $x_0,\ldots,x_k$ и стоятся прямоугольники с основаниями, равными отрезкам $[x_{i-1},x_i]$, $i=1,2,\ldots,k$, и высотами, равными $h_i = \hat{p}_i/\Delta x_i$, где $\Delta x_i = x_i - x_{i-1}$ – длина соответствующего интервала группировки. В случае равных интервалов $[x_{i-1},x_i)$ группировки данных высоты прямоугольников иногда принимают равными либо $\hat{p}_i$, либо n_i. Пусть, напр., измерение веса у 50 новорожденных дало результаты (см. рис. 1):

вес, кг	[3 - 3,3)	[3,3 - 3,6)	[3,6 – 3,9)	[3,9 – 4,2)	[4,2 – 4,5)	[4,5 – 4,8)
n_i	5	11	17	11	4	2
h_i	0,33	0,73	1,13	0,73	0,26	0,13

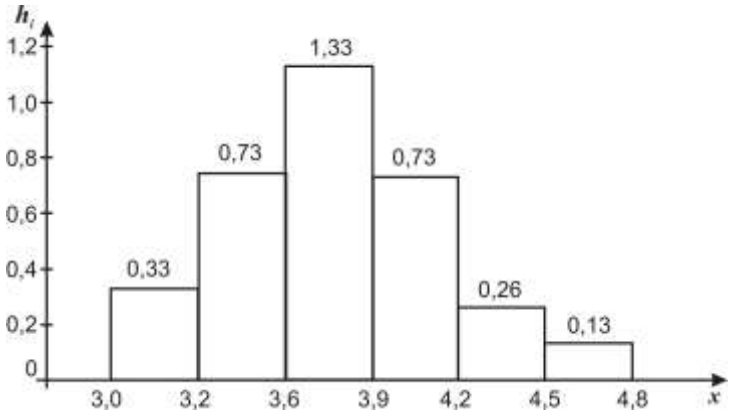


Рис. 1

Д

ДЕСКРИПТИВНАЯ МОДЕЛЬ

описательная модель, применяемая для установления статистических закономерностей исследуемых процессов, изучения вероятных путей их развития в неизменных условиях или в отсутствие внешних воздействий. Дескриптивная модель базируется на эмпирических данных и позволяет установить статистические закономерности поведения системы в целом и ее элементов. Дескриптивный характер модели определяется не только её математической структурой, но и характером ее использования. Дескриптивные модели только объясняют наблюдаемые факты или дают вероятный прогноз в отличие от нормативных моделей, предполагающих целенаправленную деятельность. К дескриптивным моделям относятся имитационные модели поведения тех или иных элементов исследуемой системы, напр., имитационные модели развития предприятия, а также прогнозные модели для различных частей экономики.

ДЕСКРИПТИВНАЯ СТАТИСТИКА

общее название статистических процедур, цель которых – получение обобщённой информации о выборке исследования. В отличие от индуктивной статистики Д.с. не делает выводов о ген. совокупности на основании результатов исследования частных случаев. Осн. задача Д.с. заключается в том, чтобы дать сжатую концентрированную характеристику конкретного изучаемого явления. Д.с. включает в себя обработку эмпирических данных, их систематизацию (группировку), наглядное представление в форме графиков и табл., а также их количественное описание посредством расчёта осн. статистических показателей.

Статистическое изучение явлений начинается с этапа статистического наблюдения, в ходе которого, в соответствии с познавательными целями и задачами, формируется массив исходных данных об изучаемом объекте, т.е. формируется информационная база исследования. Статистическое наблюдение – научно организованный, планомерный и систематический процесс сбора массовых сведений о социально-экономических и иных явлениях и процессах путём регистрации заранее намеченных существенных признаков. Процесс проведения статистического наблюдения включает этапы: программно-методологическая подготовка проведения наблюдения; организационная подготовка проведения наблюдения; выбор формы, способа и вида статистического наблюдения; проведение статистического наблюдения, сбор данных наблюдения, накапливание статистической информации; синтаксический, логический и арифметический контроль данных статистического наблюдения; выработка выводов и предложений по проведению статистического наблюдения. Научная организация проведения статистического наблюдения обеспечивает достоверность и высокое качество конечных результатов статистического исследования в целом.

Статистическая сводка – следующий после статистического наблюдения этап статистического исследования социально-экономических явлений. Статистической сводкой называется первичная обработка статистических данных с целью получения обобщённых характеристик изучаемого явления по ряду существенных для него признаков для выявления типичных черт и закономерностей, присущих явлению в целом. По глубине и точности обработки материала различают сводку простую и сложную. Простой статистической сводкой называется операция по подсчёту итоговых данных по совокупности единиц наблюдения и оформление этого материала в виде табл. Сложной статистической сводкой называется комплекс операций, включающих распределение единиц наблюдения изучаемого явления на группы, составление системы показателей для характеристики выделенных групп и подгрупп изучаемой совокупности явлений, подсчёт итогов в каждой группе и подгруппе, оформление результатов работы в виде статистических табл.

Статистической группировкой называется разбиение общей совокупности единиц объекта наблюдения по одному или нескольким признакам на однородные группы, различающиеся между собой в качественном и количественном отношении и позволяющие выделить типы явлений, изучить структуру совокупности или проанализировать взаимосвязи и взаимозависимости между признаками. По характеру решаемых задач статистические группировки бывают: типологические; структурные; аналитические. Типологическая группировка – разбиение разнородной совокупности единиц наблюдения на качественно однородные группы, классы, типы явлений. Структурная группировка – разбиение однородной в качественном отношении совокупности единиц по определенным признакам на группы, характеризующие её состав и структуру. Структурные группировки применяются в изучении практически всех социально-экономических явлений. Аналитическая группировка выявляет взаимосвязи и взаимозависимости между изучаемыми явлениями и признаками, их характеризующими. По способу построения группировки бывают простые и комбинационные. Простой называется группировка, в которой группы образованы по одному признаку. Комбинационной называется группировка, в которой разбиение совокупности на группы производится по двум и более группировочным признакам, взятым в сочетании (комбинации) друг с другом. Сначала группы формируются по одному признаку, затем группы делятся на подгруппы по

другому признаку, а эти, в свою очередь, делятся по третьему и т.д. Т.о., комбинационные группировки дают возможность изучить единицы совокупности одновременно по нескольким признакам. Построение статистических группировок проходит этапы: выбора группировочного признака; определения необходимого числа групп, на которые необходимо разбить изучаемую совокупность; установления границ интервалов группировки; установления для каждой группировки показателей или их системы, которыми должны характеризоваться выделенные группы. Результатом сводки материалов статистического наблюдения могут выступать данные, характеризующие количественное распределение единиц совокупности по тем или иным признакам. В этом случае речь идет о рядах распределения, основная задача анализа которых заключается в выявлении характера и закономерности распределения. Рядом распределения называется упорядоченное распределение единиц совокупности по определенному варьирующему признаку (атрибутивному или вариационному) на однородные группы. В зависимости от признака, положенного в основание построения ряда распределения, различают атрибутивные и вариационные ряды распределения. Атрибутивным называется ряд распределения, построенный по качественным признакам, не имеющим числового выражения и характеризующим свойство, качество изучаемого социально-экономического явления. Вариационным называется ряд распределения, построенный по количественному признаку, т.е. признаку, имеющему числовое выражение.

Описательная статистика использует три осн. метода агрегирования данных: табличное представление, графическое изображение и расчёт статистических показателей. Данные статистической сводки и группировки отражают в статистических табл. По своему виду табл. различаются на простые, групповые и комбинационные. Одно из осн. требований, предъявляемых к табл. – их выразительность и наглядность. Они не должны быть громоздкими и перегруженными излишними подробностями и деталями. Лучше вместо одной всеобъемлющей табл. привести несколько меньших по объёму, но таких, данные которых говорили бы сами за себя. Для получения более полного и наглядного представления об изучаемых явлениях и процессах по данным статистических табл. строят графики, диаграммы и т.д. Количественная определённость – объективное свойство предмета, познаваемое статистикой. Количественную характеристику статистика выражает через определённого рода числа, которые называются статистическими показателями. Статистический показатель – количественная характеристика явлений и процессов, непосредственно связанная с внутренним содержанием изучаемого процесса, его сущностью. Осн. статистические показатели можно разделить на две группы: меры среднего уровня и меры рассеяния. Меры среднего уровня дают усредненную характеристику совокупности объектов по определённому признаку. К ним относятся различные средние (арифметическая, гармоническая, геометрическая и др.), а также структурные средние – мода и медиана. К мерам рассеяния относятся: размах вариации, дисперсия, среднее квадратическое отклонение, интерквантильный размах.

См. также *Вариационный ряд*, *Визуализация данных*, *Выборочные характеристики*.

ДИСПЕРСИЯ

мера рассеяния или разброса значений случайной величины относительно её математического ожидания.

Д. случайной величины X определяется как её второй центральный момент и обозначается через $D(X)$ или σ^2 (теоретическая д.) и s^2: (эмпирическая Д.):

$$\begin{cases} \sigma^2 = D(X) = \int\limits_{-\infty}^{+\infty} [x\text{-}M(X)]^2 f(x)dx, \text{если } X - \text{непрерывна;} \\ \sigma^2 = D(X) = M[X - M(X)]^2 = \sum\limits_{i=1}^{n} [x_i - M(X)]^2 p_i, \text{если } X - \text{дискретна,} \end{cases}$$

где $M(X)$ – математическое ожидание случайной величины X, а p_i – вероятность x_i-*го* значения случайной величины *(i=1,2,…,n)*.

Свойства теоретической Д.:

1) $D(C) = 0$, где C – некоторая неслучайная величина;
2) $D(CX) = C^2D(X)$;
3) $D(C+KX) = K^2D(X)$, где C и K – некоторые неслучайные величины;
4) Если X и Y – независимые случайные величины, то $D(X+Y) = D(X)+D(Y)$ и $D(X-Y) = D(X) + D(Y)$.

Эмпирическую (выборочную) д. s^2 можно рассматривать как точечную оценку теоретической д., найденную по выборке $x_1, x_2, \ldots, x_n$ объёмом n:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \bar{x^2} - \bar{x}^2,$$

где $\bar{x}$ – средняя арифметическая (выборочная средняя), а x_i – результат i-го наблюдения i=1,2,..,n.

Свойства эмпирической Д.:

1) $s_c^2 = 0$, где c – некоторая неслучайная величина;
2) $s_{kx}^2 = k^2 s_x^2$, где k - некоторая неслучайная величина;
3) $s_{x+c}^2 = s_x^2 = s^2$;
4) $s^2 = \bar{x^2} - \bar{x}^2$;

несмещенной оценкой теоретической д. σ^2 является исправленная дисперсия s^2 вида:

$$\bar{s^2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{n}{n-1}s^2$$

Многомерным аналогом Д. является величина определителя ковариационной матрицы, называемая *обобщённой дисперсией* p-мерной случайной величины $X = (x_1, x_2, \ldots, x_p)$:

$$D_{об.} = det\Sigma = det\begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix}.$$

Диагональные элементы ковариационной матрицы Σ определяют *частные дисперсии* одномерных случайных величин x_i:

$\sigma_{ii} = M[x_i - M(x_i)]^2 = D(x_i)$.

ДИСПЕРСИЯ ВЫБОРОЧНАЯ

см. в ст. Выборочная дисперсия

ДИСПЕРСИЯ ОСТАТОЧНАЯ

оценка дисперсии ошибки уравнения регрессии, часть вариации (дисперсии) результативного признака Y, которую не удалось объяснить с помощью уравнения регрессии

$$\hat{Y} = f(X, \beta),$$

где X – матрица объясняющих переменных. Функция $f(X, \beta)$ может иметь как линейный так и нелинейный вид. Если исходная модель

$$Y = f(X, \beta) + \varepsilon$$

(ε – ошибка модели, β оцениваемые коэффициенты модели), а оцененное по исходным данным значение результативного признака

$$\hat{Y} = f(X, b),$$

то для i-го наблюдения разность между реальным и предсказанным по модели значением объясняемой переменной (остаток) будет

$$e_i = y_i - \hat{y}_i,$$

а дисперсия вектора остатков – Д.о. – будет определяться соотношением:

где n – объём многомерной выборки. Данная величина $\hat{\sigma}_\varepsilon^2$ предполагается к

$$e^2 = \frac{1}{}\sum_{i=1}^{n}{}^2_i = \frac{1}{}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

использованию в качестве оценки истинной дисперсии ошибки модели σ_ε^2, но вычисления показывают, что $\hat{\sigma}_\varepsilon^2$ будет смещённой оценкой σ_ε^2.

В качестве несмещённой оценки σ_ε^2 рассматривается величина:

$$s_e^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

где p – число объясняющих переменных.

См. также *Регрессионный анализ.*

ДИСПЕРСИОННЫЙ АНАЛИЗ

статистический метод анализа структуры связи между результативным признаком и факторными признаками; предложен Р. Фишером. При этом, среди факторов могут быть как случайные, так и неслучайные величины, измеряемые в любой из шкал: интервальной, порядковой или номинальной.

Решение задачи измерения связи опирается на разложение суммы квадратов отклонений наблюдаемых значений результативного признака Y от общей средней на отдельные слагаемые, обусловливающие изменение Y.

В соответствии с предполагаемой структурой связи строят план или дисперсионный комплекс наблюдений (экспериментов). Осн. элементарным объектом и понятием плана или комплекса является ячейка или клетка. Получаемые данные изображаются в виде комбинационной табл. (или ряда табл.), на пересечении строк и столбцов которой помещаются данные, принадлежащие конкретной ячейке комплекса. Такие табл. служат исходными в Д.а. и предназначены для получения оценок параметров распределения результативного признака ген. совокупности в зависимости от факторных значений, а также статистических выводов об отсутствии или наличии влияния факторов на результативный признак.

Предполагается, что результативный признак Y в ген. совокупности распределён нормально. Наблюдения, попавшие в каждую ячейку, образуют однородную группу не коррелированных между собой (и, в силу нормальности, независимых) случайных величин, имеющих одинаковые *математические ожидания* и дисперсии. Часть дисперсии, обусловленная действием всех неконтролируемых факторов, объединяется в один общий (случайный) фактор, называется остаточной и обозначается.

В зависимости от характера контролируемых факторов рассматриваются различные модели Д.а. Если все контролируемые факторы имеют неслучайные, фиксированные уровни, то модель называется детерминированной (модель М1). Если все контролируемые факторы имеют случайные уровни, то модель называется случайной (модель М2). Модель называют смешанной, если в ней имеются факторы, как со случайными, так и с фиксированными уровнями. Выбор модели определяется практическими соображениями, в частности возможностью или необходимостью распространения статистических выводов на ген. совокупность (модель М2 и смешанная) по каким-нибудь факторам (или взаимодействиям) либо достаточностью выводов относительно включенных в наблюдение экспериментов уровней факторов (модель М1 и смешанная).

Осн. критерием проверки гипотезы об отсутствии влияния отдельного фактора или взаимодействия факторов является *критерий Фишера* (F-критерий). Наиболее полный Д.а. (без предварительных условий) структуры трехфакторного, четырехфакторного и т.д. комплексов возможно провести с помощью детерминированной модели (М1); вычисления значительно упрощаются при одинаковом числе наблюдений в каждой клетке.

Д.а. относится к полному плану эксперимента, когда на влияние факторов и их взаимодействия не накладывается никаких ограничений. Наличие априорной информации об отсутствии взаимодействий факторов позволяет существенно сократить объем наблюдений. Напр., в двухфакторном Д.а., в котором анализируется влияние на результативный признак Y двух факторов (A и B), каждый из которых имеет n уровней, требуется, как минимум, n^2 наблюдений. Предположение об отсутствии влияния на Y взаимодействия факторов позволяет в этом случае перейти к плану эксперимента типа «*латинский квадрат*» и в n раз сократить объём наблюдений.

ДОВЕРИТЕЛЬНАЯ ВЕРОЯТНОСТЬ

такая вероятность γ, что событие с вероятностью $\alpha = 1 - \gamma$ можно считать невозможным. Выбор Д.в. полностью зависит от исследователя, причём во внимание принимаются не только его личные наклонности, но и физическая суть рассматриваемого явления. Так, степень

доверия авиапассажира к надёжности самолета, несомненно, должна быть выше степени доверия покупателя к надёжности электрической лампочки. В математической статистике обычно используют значения Д.в. 0,9; 0.95; 0,99; реже 0,999; 0,9999 и т.д. Д.в. применяется при интервальном оценивании параметров неизвестного распределения: либо значение γ задается исследователем – при этом решается задача построения интервала, который с вероятностью γ «накроет» неизвестный параметр; либо по построенной к.-л. образом интервальной оценке определяется указанная вероятность γ.

ДОВЕРИТЕЛЬНАЯ ОБЛАСТЬ

вектора параметров $\theta = (\theta_1, \theta_2, \ldots \theta_k)^T$ *ген. совокупности* – случайная область в соответствующем k-мерном пространстве, полностью определяемая результатами наблюдений, которая с близкой к единице вероятностью (надёжностью) γ содержит неизвестное значение вектора θ. Существует бесконечное множество Д.о., соответствующих одному и тому же значению γ. Обычно стараются определить Д.о., имеющие миним. размеры при данной надежности γ. Часто этому условию удовлетворяют области, симметричные относительно вектора оценок $\hat{\theta}$ параметра θ. Осн. трудность в построении Д.о. представляет определение законов распределения подходящих статистик. В настоящее время эти вопросы хорошо разработаны только для нормального распределения наблюдаемых случайных величин. Рассмотрим построение Д.о. для k-мерного вектора μ ген. средних. Пусть имеется результаты n наблюдений из ген. совокупности X с k-мерным нормальным распределением

$N_k(\mu, \Sigma)$,

для которых найден вектор средних $\bar{x}$ и несмещённая оценка S ковариационной матрицы Σ. При построении Д.о. используют статистику T^2 Хоттелинга:

$T^2 = n(\bar{x} - \mu)^T S^{-1} (\bar{x} - \mu)$,

где S^{-1}- матрица, обратная ковариационной матице S. Так как T^2 и F – распределение Фишера связаны соотношением

$T^2_{\alpha,k,n-k} = F_{\alpha,k,n-k} k(n-1)/(n-k)$,

где $\alpha = 1 - \gamma$, $F_{\alpha,k,n-k}$ – точка F - распределения, соответствующая уровню значимости α и числам степеней свободы k и $n-k$, то уравнение поверхности, ограничивающей доверительную область k ген. средних с надёжностью γ будет:

$$(\bar{x} - \mu)^T S^{-1} (\bar{x} - \mu) = F_{\alpha,k,n-k} \frac{k(n-1)}{n(n-k)}.$$

Это уравнение определяет k-мерный эллипсоид (эллипс при $k = 2$) с центром $\bar{x}$, т.к. его левая часть представляет положительно определенную квадратичную форму относительно μ. При известной ковариационной матрице Σ используется статистика, имеющая χ^2-распределение с k степенями свободы,

$t^2 = n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu)$.

В этом случае с надежностью γ вектор μ накрывается Д.о., задаваемой неравенством :

$(x - \mu)^T S^{-1} (x - \mu) \le (\chi^2)^{-1} (1 - \gamma)$.

Для построения совместной доверительной области разноплановых параметров (напр., математического ожидания и дисперсии) используется подход, состоящий в определении таких интервалов

$I_1(\theta_1), I_2(\theta_2), \ldots, I_m(\theta_m)$

для координат $\theta_1, \theta_2, \ldots \theta_m$

вектора параметров θ, для которых вероятность одновременного накрытия всех

$\theta_1, \theta_2, \ldots \theta_m$

соответствующими интервалами была бы не меньше заданного значения γ. Для этого необходимо найти доверительные интервалы

$I_1(\theta_1), I_2(\theta_2), \ldots, I_m(\theta_m)$

для координат вектора θ, соответствующие надежности

$\gamma^\circ = 1 - \frac{1}{m}(1 - \gamma)$.

Искомая область будет прямоугольной k-мерной.

ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ

границы *доверительного интервала* (θ_1, θ_2) для неизвестного параметра θ распределения, принадлежащие множеству допустимых значений параметра, определяемые по результатам наблюдений, Значения θ_1 и θ_2 (для определённости $\theta_1 < \theta_2$), должны быть подобраны так, чтобы интервал (θ_1, θ_2) с наперёд заданной достаточно большой вероятностью γ накрывал неизвестное значение параметра θ:

$$P(\theta_1 < \theta < \theta_2) = \gamma = 1 - \alpha.$$

Числа θ_1 и θ_2 называются Д.г., при этом θ_1 – нижняя граница, а θ_2 – верхняя граница; γ называется *доверительной вероятностью* (надёжностью или коэффициентом доверия), α – уровнем значимости. Из осн. соотношения видно, что однозначность в выборе Д.г. θ_1 и θ_2 отсутствует. Для устранения неоднозначности для каждой из границ необходимо задать значения величин α_1 и α_2, для которых справедливы соотношения:

$$P(\theta_1 < \theta) = 1 - \alpha_1, \quad P(\theta_2 > \theta) = 1 - \alpha_2.$$

При этом должно выполняться условие

$$\alpha_1 + \alpha_2 = \alpha = 1 - \gamma.$$

Если величины α_1 и α_2 взять равными, т.е.

$$\alpha_1 = \alpha_2 = \alpha/2 = (1 - \gamma)/2,$$

то доверительный интервал (θ_1, θ_2) называется центральным. Следует заметить, что наличие центральных интервалов не означает, что границы θ_1 и θ_2 находятся на равном расстоянии от истинного значения оцениваемого параметра.

См. также *Интервальное оценивание*.

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ

статистическая оценка параметра вероятностного распределения θ, имеющая вид интервала (θ_1, θ_2), который с высокой вероятностью γ «накрывает» неизвестное значение оцениваемого параметра. Границы этого интервала являются функциями от результатов наблюдений

$$\mathbf{x} = (x_1, x_2, \ldots, x_n):\ \theta_1 = \varphi_1(x_1, x_2, \ldots, x_n)$$

и $\theta_2 = \varphi_2(x_1, x_2, \ldots, x_n),\ \theta_1 < \theta_2.$

Предположим, что $x_1, x_2, \ldots, x_n$ независимые случайные величины с распределением вероятностей F_θ, которое зависит от числового параметра θ, $\theta \in \Theta \subseteq \mathbb{R}$, где Θ – т.н. параметрическое множество. Д.и. для параметра θ при фиксированном значении γ, $0 < \gamma < 1$, называется интервал с границами θ_1 и θ_2 в параметрическом множестве Θ такой, что

$$P(\theta_1 \le \theta \le \theta_2) = \gamma.$$

Вероятность γ называется *доверительной вероятностью* или надёжностью, границы θ_1 и θ_2 называются *доверительными границами*. Существует несколько способов построения интервальных статистических оценок, наиболее распространенный – метод Д.и., предложенный Е.Нейманом в 1935. Данный метод отличается от других методов интервального оценивания логической простотой и независимостью от априорных предположений о параметре θ.

Для векторного параметра $\theta = (\theta_1, \ldots, \theta_k)$ понятие Д.и. заменяется на соответствующую *доверительную область* в *k*-мерном пространстве. См. также *Интервальное оценивание*.

И

ИНТЕРВАЛ ГРУППИРОВАНИЯ

интервал, определяющий границы значений варьирующего признака, лежащих в пределах определённой группы. Каждый интервал имеет свою длину (ширину), верхнюю и нижнюю границы или хотя бы одну из них. Нижней (верхней) границей интервала называется наименьшее (наибольшее) значение признака в интервале. Шириной интервала называется разность между верхней и нижней границами. И.г. в зависимости от их ширины бывают равные и неравные. Последние делятся на прогрессивно возрастающие, прогрессивно убывающие, произвольные и специализированные. Если вариация признака проявляется в сравнительно узких границах и распределение носит равномерный характер, то строят группировку с равными интервалами. Ширина равного интервала h определяется по формуле:

$$h = (x_{\max} - x_{\min})/k,$$

где $x_{\max}, x_{\min}$ – макс. и миним. значения признака в совокупности; k – число групп. Если размах вариации признака в совокупности велик и значения признака варьируют неравномерно, то надо использовать группировку с неравными интервалами. Это достигается либо объединением двух или нескольких малочисленных или «пустых» последовательных равных интервалов, либо применением прогрессивно возрастающих и прогрессивно убывающих интервалов, в основе построения которых лежит принцип арифметической или геометрической прогрессии. Величина интервалов, изменяющихся в арифметической прогрессии, определяется по формуле:

$$h_{i+1} = h_i + a,$$

где a – константа: для прогрессивно возрастающих интервалов положительная и отрицательная для прогрессивно убывающих; h_i – величина i-го интервала; а в геометрической прогрессии:

$$h_{i+1} = h_i \cdot q,$$

где q – константа: > 1 – для прогрессивно возрастающих и < 1 – для прогрессивно убывающих интервалов. Решение вопроса о выборе равных или неравных интервалов зависит от числа единиц совокупности, попавших в каждую выделенную группу, т.е. от степени заполнения интервалов. При определении границ интервалов статистических группировок иногда исходят из того, что изменение количественного признака приводит к появлению нового качества. В этом случае граница интервала устанавливается там, где происходит переход от одного качества к другому. В группировках, имеющих целью отобразить качественные особенности и специфику выделяемых групп единиц изучаемой совокупности по признаку, применяются специализированные интервалы. При изучении социально-экономических явлений на макроуровне часто применяют группировки, интервалы которых не будут ни прогрессивно возрастающими, ни прогрессивно убывающими. Такие интервалы называются произвольными и, как правило, используются при группировке пр-тий, напр., по уровню рентабельности. И.г. могут быть закрытыми и открытыми. Закрытыми называются интервалы, в которых указаны верхняя и нижняя границы. Открытыми называются интервалы, у которых указана только одна граница: верхняя – у первого, нижняя – у последнего.

ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

определение интервала приближённых значений оцениваемого параметра θ. Границы этого интервала, принадлежащие множеству допустимых значений параметра, определяются по результатам наблюдений. В соответствии со смыслом И.о. по сделанной выборке на основе некоторых правил должны быть найдены два значения θ_1 и θ_2 (для определённости $\theta_1 < \theta_2$), такие,

$$P(\theta_1 < \theta < \theta_2) = \gamma = 1 - \alpha.$$

чтобы интервал (θ_1, θ_2) с наперёд заданной вероятностью γ накрывал неизвестное значение параметра

θ: $\theta_1 < \theta < \theta_2$.

Числа θ_1 и θ_2 называются *доверительными границами*, при этом θ_1 – нижняя граница, а θ_2 – верхняя граница. Совокупность доверительных границ (θ_1, θ_2) – *доверительный интервал*. При этом величину $\Delta = \theta_2 - \theta_1$ называют шириной (длиной) доверительного интервала для параметра θ. Число γ называется *доверительной вероятностью* (надёжностью или коэффициентом доверия), α – уровнем значимости. Доверительный интервал по своей природе случаен, как по своему расположению, так и по ширине, в силу того, что величины $\theta_{1,2}$ и Δ строятся, как функции выборочных данных $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Ширина Δ доверительного интервала существенно зависит от объёма выборки n (уменьшается с ростом n) и от величины доверительной вероятности (увеличивается с приближением γ к единице). Выбор конкретного значения доверительной вероятности в значительной степени зависит от целей и условий проведения измерений исследуемого параметра. Обычно полагают, что доверительная вероятность достаточно велика (напр., $\gamma = 0{,}95;\ \ 0{,}99\ \ 0{,}999$ и т.д.), при этом практически достоверно, что интервал (θ_1, θ_2) накроет неизвестный параметр θ. Существует несколько способов построения интервальных

статистических оценок, т.е. определения границ θ_1 и θ_2. Наиболее распространённый – метод доверительных интервалов, который приводит к построению точных доверительных областей в случае, когда можно подобрать некоторую статистику

$$T(x_1, x_2, \ldots, x_n) = T(\mathbf{x})$$

– функцию от результатов наблюдений, для которой выполнены условия: закон распределения статистики $T(\mathbf{x})$ не зависит от оцениваемого параметра θ и описывается одним из стандартных затабулированных распределений (стандартным нормальным, распределением Фишера, Стьюдента, χ^2-распределением); из того факта, что

$$P(T_1 < T < T_2) = \gamma$$

– значения данной статистики заключены в определённых пределах с заданной вероятностью, возможно сделать вывод, что оцениваемый параметр тоже должен лежать между некоторыми границами с той же самой вероятностью $P(\theta_1 < \theta < \theta_2) = \gamma$. Чаще всего в выражении для $T(\mathbf{x})$ анализируемый параметр θ и его точечная оценка $\hat{\theta}$ участвуют в комбинациях разности $(\hat{\theta} - \theta)$ или отношения $(\hat{\theta}/\theta)$. В качестве примера рассмотрим задачу интервального оценивания параметра a нормальной ген. совокупности с плотностью распределения

$N(a, \sigma^2)$, г

де $\sigma > 0$ – известное число. Для построения доверительного интервала рассматривается статистическая оценка

$$\bar{x} = (x_1 + \ldots + x_n)/n$$

параметра a и статистика

$$T(\mathbf{x}) = \frac{\bar{x} - a}{\sigma/\sqrt{n}},$$

которая при любом значении a имеет нормальное стандартное распределение с функцией распределения

$$\Phi(\mathrm{t}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-z^2/2} dz.$$

Поэтому для любого $t > 0$ вероятность

$$P(|T(\mathbf{x})| \le t) = P(-t \le T(\mathbf{x}) \le t) = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1$$

не зависит от параметра a. Для заданной доверительной вероятности γ по табл. нормального распределения определяем значение t_γ из соотношения $2\Phi(t_\gamma) - 1 = \gamma$. Тогда

$$P(|T(\mathbf{x})| \le t_\gamma) = P\left(\left|\frac{\bar{x} - a}{\sigma/\sqrt{n}}\right| \le t_\gamma\right) = P\left(\bar{x} - \frac{t_\gamma \sigma}{\sqrt{n}} \le a \le \bar{x} + \frac{t_\gamma \sigma}{\sqrt{n}}\right) = \gamma$$

Это означает, что случайный доверительный интервал

$$\left(\bar{x} - \frac{t_\gamma \sigma}{\sqrt{n}}, \bar{x} + \frac{t_\gamma \sigma}{\sqrt{n}}\right)$$

накрывает неизвестное среднее значение a с заданной вероятностью γ. Пусть имеется выборка объёма $n = 25$ из нормально распределённой ген. совокупности с дисперсией $\sigma^2 = 1$, для которой оценено среднее значение $\bar{x} = 0$. Зададим доверительную вероятность $\gamma = 0{,}99$, и по табл. нормального распределения определим t_γ из условия $\Phi(t_\gamma) = 1{,}99/2 = 0{,}995$. Получим

$$t_\gamma = \Phi^{-1}(0{,}995) = 2{,}575.$$

В результате доверительный интервал для параметра a будет (-0,515; 0,515). При увеличении объёма выборки до $n = 100$ при прочих равных условиях ширина доверительного интервала уменьшится (-0,258; 0,258). При уменьшении доверительной вероятности, напр., $\gamma = 0{,}8$, величина

$$t_\gamma = \Phi^{-1}(0{,}9) = 1{,}282$$

и доверительный интервал для оцениваемого параметра a при объеме выборки $n = 25$ уменьшится (-0,256; 0,256).

При наличии априорной информации о распределении параметра строятся байесовские интервальные статистические оценки. При *байе-*

совском подходе к оцениванию полагают, что оцениваемый параметр θ является случайной величиной с известным априори распределением $f_{pr}(\theta)$. Если известна к тому же $f(\hat{\boldsymbol{\theta}}\,|\,\theta)$ – условная плотность вероятности оценок $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ при фиксированном значении неизвестного параметра θ (её называют функцией правдоподобия), то может быть оценена условная плотность вероятности (апостериорная вероятность) параметра θ при заданном наборе оценок $\hat{\boldsymbol{\theta}}$:

$$f(\theta\,|\,\hat{\boldsymbol{\theta}}) = f_{ps}(\theta\,|\,\hat{\boldsymbol{\theta}}) = \frac{f_{pr}(\theta)\cdot f(\hat{\boldsymbol{\theta}}\,|\,\theta)}{\int\limits_{\boldsymbol{\Theta}} f_{pr}(\theta)\cdot f(\hat{\boldsymbol{\theta}}\,|\,\theta)d\theta}.$$

Априорная вероятность попадания оцениваемого параметра θ в интервал от θ_1 до θ_2

$$P_{pr}(\theta_1 < \theta < \theta_2) = \int\limits_{\theta_1}^{\theta_2} f_{pr}(\theta)d\theta = \gamma_{pr}.$$

Апостериорная вероятность – условная вероятность того, что истинное значение оцениваемого параметра θ лежит в интервале от θ_1 до θ_2 при найденном на основе исходной выборки значении оценки $\hat{\theta}$ определяется по формуле:

$$P_{ps}(\theta_1 < \theta < \theta_2\,|\,\hat{\theta}) = \int\limits_{\theta_1}^{\theta_2} f_{ps}(\theta\,|\,\hat{\boldsymbol{\theta}})d\theta = \gamma = 1-\alpha.$$

При этом, естественно, $\gamma \geq \gamma_{pr}$. В ситуациях, когда априорное распределение $f_{pr}(\theta)$ исследуемого параметра неизвестно, в качестве такового используют равномерное на отрезке

$[\theta_{\min}, \theta_{\max}]$ распределение,

где $[\theta_{\min}, \theta_{\max}]$ – априорный диапазон варьирования возможных значений оцениваемого параметра θ. Рассмотрим байесовскую интервальную оценку среднего значения a нормальной выборки объёма n с известной дисперсией σ^2 на основании выборочного среднего $\bar{x}$. Выборочное среднее имеет нормальное распределение:

$$f(\bar{x}|a) = \frac{1}{\sqrt{2\pi\sigma^2/n}}\exp\left(-\frac{(\bar{x}-a)^2}{2\sigma^2/n}\right).$$

Пусть априорное распределение оцениваемого параметра a является равномерным на отрезке $[b,c]$:

$$f_{pr}(a) = \begin{cases} 1/(c-b), & b \leq a \leq c, \\ 0, & a < b, \quad a > c. \end{cases}$$

Тогда условная плотность вероятности оцениваемого параметра может быть представлена в виде:

$$f(a\,|\,\bar{x}) = \frac{\dfrac{1}{\sqrt{2\pi\sigma^2/n}}\exp\left(-\dfrac{(\bar{x}-a)^2}{2\sigma^2/n}\right)}{\Phi\left(\dfrac{c-\bar{x}}{\sigma/\sqrt{n}}\right) - \Phi\left(\dfrac{b-\bar{x}}{\sigma/\sqrt{n}}\right)}, \quad b \leq \bar{x} \leq c.$$

При заданной доверительной вероятности γ верхняя и нижняя границы центрального доверительного интервала равны $a_{2,1} = \bar{x} \pm x_\gamma$, где величина x_γ определяется из уравнения:

$$\Phi\left(\frac{x_\gamma}{\sigma/\sqrt{n}}\right) = \frac{\gamma}{2}\left[\Phi\left(\frac{c-\bar{x}}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{b-\bar{x}}{\sigma/\sqrt{n}}\right)\right] + \frac{1}{2}.$$

Вычислим границы центрального доверительного интервала для апостериорной доверительной вероятности $\gamma = 0{,}95$ применительно к выборке объёма $n = 25$ с известной дисперсией $\sigma^2 = 1$ и значениями параметров симметричного априорного распределения величины a, равными $c = -b = 1$. Пусть на основе экспериментальных данных вычислено выборочное среднее $\bar{x} = 0$. Подставляя указанные значения величин в последнее выражение, имеем

$$\Phi(t_\gamma = x_\gamma\sqrt{n}/\sigma) = 0{,}975.$$

По табл. нормального распределения находим $t_\gamma \approx 1{,}96$. Следовательно, истинное значение параметра a заключено в интервале $a_{2,1} = \pm x_\gamma = \pm 0{,}392$ с заданной апостериорной вероятностью $\gamma = 0{,}95$. При этом априорная доверительная вероятность $\gamma_{pr} = 0{,}392$. Приведенные соотношения позволяют решать «обратную» задачу: по полученным к.-л. способом границам доверительного интервала определять доверительную вероятность (надёжность) оцениваемого параметра.

См. также *Байесовская оценка.*

ИСХОДНЫЕ СТАТИСТИЧЕСКИЕ ДАННЫЕ

характеристики объекта или его поведения, какого-либо процесса или явления, измеряемые в номинальной, порядковой или ординальной шкале. Наиболее распространённая форма представления И.с.д. – матрица (или табл.) типа «объект-свойство». Она возникает в ситуации, когда на каждом из *n* объектов исследуемой совокупности регистрируются значения целого набора признаков, количество которых p, в N последовательные моменты времени $t_1, t_2, \ldots, t_N$. Т.о., И.с.д. могут быть представлены в виде матрицы размерности $[n \times p]$, называемой пространственно-временной выборкой:

$$X = \begin{pmatrix} x_{11}(t) & x_{12}(t) & \ldots & x_{1p}(t) \\ x_{21}(t) & x_{22}(t) & \ldots & x_{2p}(t) \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1}(t) & x_{n2}(t) & \ldots & x_{np}(t) \end{pmatrix}, \quad t = t_1, t_2, \ldots, t_N,$$

где $x_{ij}(t)$ – значение признака анализируемого признака *j*, характеризующего состояние объекта *i* в момент времени t_k . На самом деле приведённая запись статистических данных определяет целую последовательность, а именно, N штук матриц «объект-свойство».

Можно сказать, что данные представленного вида содержат *n* реализаций *p*-мерного временного ряда

$$\left(x_1(t), x_2(t), \ldots, x_p(t)\right).$$

Если исследователь располагает т.н. одномоментными наблюдениями, то это соответствует случаю $N = 1$, при этом для упрощения обозначений индекс времени t опускается, а получающаяся выборка называется пространственной статистической. Другой частный случай И.с.д. получается, если обследуется во времени единственный объект, т.е. в предложенной матрице положить $n = 1$. Тогда речь идет об анализе единственной траектории p-мерного временного ряда. Если же дополнительно принять условие $p = 1$, то И.с.д. будут представлять один временной ряд. Вторая форма представления И.с.д. возникает в ряде ситуаций, когда статистические данные получают с помощью специальных опросов, анкет, экспертных оценок. При этом возможны случаи, когда элементом первичного наблюдения является не состояние *i*-го объекта, а некоторая характеристика γ_{ij} парного сравнения двух объектов (или признаков), соответственно с номерами *i* и *j*. Характеристика γ_{ij} может выражать меру различия или сходства, меру связи или взаимодействия, отношения предпочтения (например, $\gamma_{ij} = 1$ если объект *i* не хуже объекта *j*, и $\gamma_{ij} = 0$ в противном случае), меру взаимной коррелированности и т.д. В этом случае исследователь располагает в качестве массива исходных статистических данных матрицей парных сравнений размера $[n \times n]$, если рассматриваются характеристики парных сравнений n - объектов, или размера $[p \times p]$, если рассматриваются характеристики парных сравнений p - признаков:

$$\Gamma = \begin{pmatrix} \gamma_{11}(t) & \gamma_{12}(t) & \ldots & \gamma_{1p}(t) \\ \gamma_{21}(t) & \gamma_{22}(t) & \ldots & \gamma_{2p}(t) \\ \ldots & \ldots & \ldots & \ldots \\ \gamma_{n1}(t) & \gamma_{n2}(t) & \ldots & \gamma_{np}(t) \end{pmatrix}, \quad \begin{pmatrix} m = n \quad \text{или} \quad m = p, \\ t = t_1, t_2, \ldots, t_N \end{pmatrix}.$$

В статическом варианте, т.е. при $N = 1$, исследователь располагает лишь одной матрицей сравнений

$$\Gamma = \{\gamma_{ij}\},$$

описывающую ситуацию в какой-то один фиксированный момент времени. От формы записи типа матрицы X при наличии заданной метрики в пространстве объектов или в пространстве признаков можно перейти к матрице Γ . Одно-

значный обратный переход от матрицы парных сравнений Г к матрице X без дополнительных предположений и специальных методов, в общем, невозможен. Можно выделить два подхода к интерпретации и анализу И.с.д. Первый подход – вероятностно-статистический – развивается в рамках классической математической статистики, т.е. в условиях хотя бы приблизительного выполнения требований статистического ансамбля: когда имеется практическая или хотя бы мысленно представимая возможность многократного воспроизведения осн. комплекса условий, при которых производились измерения анализируемых данных. Данный подход предусматривает возможность вероятностной интерпретации анализируемых данных и получаемых в результате этого анализа статистических выводов. В поле зрения исследователя при подобной вероятностной интерпретации одновременно находятся две совокупности: реально наблюдаемая, или выборка, и теоретически домысливаемая, или ген. совокупность. Осн. свойства и характеристики выборки могут быть вычислены по имеющимся статистическим данным. Осн. же свойства ген. совокупности не известны, но с помощью вероятностно-статистических методов для них может быть получено более или менее точное представление по соответствующим выборочным характеристикам. Успех статистического анализа и моделирования И.с.д. зависит от правильного выбора модели механизма генерации этих данных. Критерий качества оценки или степени адекватности модели определяется на основе принципа максимального правдоподобия имеющихся у нас наблюдений, который в свою очередь базируется на знании модели закона распределения вероятностей этих наблюдений. Второй подход – логико-алгебраический – возникает в ситуации, когда исследователь не располагает никакими априорными сведениями о вероятностной природе анализируемых данных, или если эти данные вообще не могут быть интерпретированы как выборка из генеральной совокупности. Тогда при выборе критерия качества метода оценивания или степени адекватности конструируемой модели исследователь вынужден опираться на соображения конкретно-содержательного плана: как именно получены анализируемые данные и какова конечная прикладная цель их анализа. Поскольку эти соображения основаны на обычной логике и реализуются в виде критерия некоторого алгебраического вида, то соответствующий поход принято называть логико-алгебраическим. В рамках этого подхода исследователь не может интерпретировать И.с.д. как выборку из некоторой ген. совокупности, использовать вероятностные модели и претендовать на вероятностную интерпретацию выводов. Гл. различие двух возможных подходов к статистическому анализу И.с.д. заключается в способе обоснования выбора критерия качества метода или степени адекватности модели, а также в интерпретации самого критерия и статистических выводов. После того, как выбор конкретного вида оптимизируемого критерия качества осуществлен, математические средства решения задач статистического анализа и моделирования И.с.д. оказываются общими для обоих подходов: исследователь использует методы решения экстремальных задач для оптимизации выбранного критерия качества. На заключительном же этапе – этапе осмысления и интерпретации полученных результатов – каждый из подходов снова имеет свою специфику.

См. также *Многомерное шкалирование*.

К

КОИНТЕГРАЦИЯ

причинно-следственная зависимость в уровнях двух (или более) временных рядов, которая выражается в совпадении или противоположной направленности их тенденций и случайной колеблемости. Формальное определение К. двух переменных разработано Энглом и Грейнджером в 1987.

В соответствии с теорией между двумя временными рядами К. существует в случае, если линейная комбинация рядов – это стационарный временной ряд, содержащий только случайную компоненту и имеющий постоянную дисперсию на длительном промежутке времени.

Если x_t – интегрируемый временной ряд порядка κ_1 и y_t – интегрируемый временной ряд по-

рядка κ_2, причём $\kappa_2 > \kappa_1$, то при любом значении параметра θ (в том числе и $\theta = \theta_{мнк}$, где $\theta_{мнк}$ – МНК – оценка коэффициента регрессии в модели парной регрессии y_t по x_t) случайный остаток $e_t = y_t - \theta x_t$ будет интегрируемым временным рядом порядка κ_2. Если же $\kappa_1 = \kappa_2 = \kappa$, то константа θ может быть подобрана так, что e_t будет стационарным с нулевым средним. Вектор $(1; -\theta)$ или любой другой, отличающийся сомножителем называется коинтегрирующим (вектором К.).

К. временных рядов значительно упрощает процедуры и методы, используемые в целях анализа, поскольку в этом случае можно строить уравнение регрессии и определять показатели корреляции, применяя в качестве исходных данных непосредственно уровни изучаемых временных рядов, учитывая тем самым информацию, содержащуюся в исходных данных, в полном объеме. Однако, поскольку К. означает совпадение динамики временных рядов в течение длительного промежутка времени, то сама концепция применима только к временным рядам, охватывающим сравнительно длительные промежутки времени.

КОЛИЧЕСТВЕННАЯ ШКАЛА

шкала интервалов, отношений, разностей, абсолютных значений, для которых характерно наличие единицы измерения, позволяющей определить, насколько один объект отличается от другого по изучаемому критерию. На шкале интервалов, кроме отношений тождества и порядка, определено отношение разности, т.е. для любой пары объектов можно определить, на сколько единиц измерения один объект больше или меньше, чем другой. Числа упорядочены по рангам, разделены определенными интервалами. Допустимыми преобразованиями в шкале интервалов являются линейные возрастающие преобразования, т.е. линейные функции. Шкалы интервалов отличаются от шкалы отношений тем, что нулевая точка выбирается произвольно. Результаты измерений по шкале интервалов можно обрабатывать всеми *математическими методами*, кроме вычисления отношений. Шкала отношений строго определяет положение нулевой точки, благодаря чему, данная шкала не накладывает никаких ограничений на математический аппарат, используемый для обработки результатов наблюдений. Допустимыми преобразованиями шкалы отношений являются линейные возрастающие преобразования без свободного члена. По шкалам отношений измеряются такие показатели, как рост, возраст, доходы, стаж работы и т.п. По шкале отношений измеряют и те величины, которые образуются как разности чисел, отсчитанных по шкале интервалов. Так, календарное время измеряется по шкале интервалов, а интервалы времени – по шкале отношений. Следует отметить, что в большинстве статистических процедур не делается различия между свойствами интервальных шкал и шкал отношения. Шкала разностей, в отличие от шкалы отношений, не имеет естественного нуля, но имеет естественную масштабную единицу измерения, ей соответствует аддитивная группа действительных чисел. Она сходна со шкалой интервалов. Разница лишь в том, что значения данной шкалы нельзя умножать (делить) на константу. Примером шкалы разностей является историческая хронология. Шкала абсолютных величин напрямую определяет величину чего-либо. Для абсолютной шкалы допустимым является только тождественное преобразование, например, непосредственно подсчитывается число бракованных изделий в партии товара, число квартир в доме, количество учащихся, присутствующих на занятии и т.д. При таких измерениях на шкале отмечаются абсолютные количественные значения изучаемого объекта. Шкала абсолютных значений обладает теми же свойствами, что и шкала отношений, но величины, обозначенные на этой шкале, имеют абсолютные значения. Данные, полученные с помощью абсолютной шкалы, не преобразуются, и для анализа используются любые статистические меры. Результаты измерений по шкале абсолютных величин имеют наибольшую достоверность, информативность и чувствительность к неточностям измерений. Примерами К.ш. являются шкала Лайкерта (интервальная шкала), на основе которой изучается степень согласия или несогласия респондентов с определенными

высказываниями; шкала расстояний и температурные шкалы Цельсия, Фаренгейта и Кельвина; шкала массы тел и др.

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

КОЛИЧЕСТВО ИНФОРМАЦИИ ФИШЕРА

величина, используемая для определения эффективности точечных статистических оценок. Пусть $f(x,\theta)$ – плотность распределения вероятностей признака x, если x – непрерывная случайная величина, или

$$f(x_i,\theta) = P(X = x_i),$$

если x – дискретна. К.и. Фишера $I(\theta)$ о параметре θ, содержащееся в единичном наблюдении, определяется выражением:

$$I(\theta) = M\left[\frac{d\ln f(x,\theta)}{d\theta}\right]^2,$$

где M – математическое ожидание. Так как $(\ln f(x,\theta))'_\theta = f(x,\theta)'_\theta / f(x,\theta)$, то в случае, когда x – непрерывная случайная величина, К.и. Фишера вычисляется по формуле:

$$I(\theta) = \int_{-\infty}^{\infty}\left[\frac{f'_\theta(x,\theta)}{f(x,\theta)}\right]^2 f(x,\theta)dx.$$

Для дискретной случайной величины интеграл заменяется суммой:

$$I(\theta) = \sum_{i=1}^{n}\left[\frac{f'_\theta(x_i,\theta)}{f(x_i,\theta)}\right]^2 f(x_i,\theta).$$

Вычисленное значение К.и. Фишера используется в неравенстве Рао-Крамера-Фреше и позволяет определить тот минимум, который должна иметь дисперсия $D(\hat\theta)$ оценки $\hat\theta$ параметра θ, для того, чтобы быть эффективной:

$$D_{\text{ýô}}(\hat\theta) = \min D(\hat\theta) = \frac{1}{nI(\theta)},$$

В качестве примера найдем нижнюю границу дисперсии оценки ген. средней a для повторной выборки нормально распределенной ген. совокупности. Функция плотности вероятностей и её логарифм имеют вид:

$$f(x,a,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad \ln f(x,a,\sigma) = -\ln\sqrt{2\pi\sigma^2} - \frac{(x-a)^2}{2\sigma^2}.$$

Тогда К.и. Фишера

$$I(a) = M\left[(\ln f(x,a,\sigma))'_a\right]^2 = M\left(\frac{x-a}{\sigma^2}\right)^2 = \frac{M[(x-a)^2]}{\sigma^4} = \frac{D(X)}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Минимально возможная дисперсия оценки ген. средней (математического ожидания) будет определяться выражением:

$$\min D(a) = \frac{1}{nI(a)} = \frac{\sigma^2}{n}.$$

КОЭФФИЦИЕНТ АСИММЕТРИИ

наиболее употребительная мера асимметрии распределения, определяемая отношением:

$$A = \frac{\mu_3}{\mu_2^{3/2}},$$

где μ_2 и μ_3 – 2-й и 3-й центральные моменты распределения соответственно. Асимметрия распределения – качественное свойство кривой распределения, указывающее на отличие от симметричного распределения. Если К.а. положителен (отрицателен), то асимметрия распределения положительна (отрицательна). При положительной (отрицательной) асимметрии распределения более «длинная» часть кривой плотности распределения лежит правее (левее) *моды*.

См. также *Моменты выборочные*.

КОЭФФИЦИЕНТ ВАРИАЦИИ

характеристика рассеяния распределения вероятностей случайной величины. Существуют разные способы определения К.в. Наиболее часто используется К.в., определённый для положительной случайной величины X с математическим ожиданием $M(X)=a$ и дисперсией $D(X)=\sigma^2$, который определяется формулой:

$$V=\frac{\sigma}{a}.$$

Статистическая оценка К.в. подобного вида предложена К.Пирсоном в 1895:

$$V=\frac{\hat{\sigma}}{\overline{X}}\cdot 100\%$$

По сравнению с обычной мерой рассеяния – средним квадратическим отклонением σ К.в. является безразмерной характеристикой. Чем больше величина К.в., тем больше разброс значений вокруг средней, тем менее однородна совокупность по своему составу и тем менее представительна средняя. К.в. важен и в тех случаях, кода надо сравнивать средние квадратические отклонения, выраженные в разных единицах измерения. В экономике К.в. используют для моделирования технико-экономических показателей.

КОЭФФИЦИЕНТ КОНКОРДАЦИИ

коэффициент согласованности рангов, определя-ющий меру статистической связи между несколькими последовательностями рангов, характеризующими совокупность объектов. К.к. применяется в случаях, когда необходимо оценить степень согласованности мнений различных экспертов или установить наличие связь между несколькими переменными, измеренными в ординальной (порядковой) шкале. М. Кендэлом для этих целей был предложен показатель, вычисляемый по формуле:

$$W=\frac{12\sum_{i=1}^{n}D_i^2}{m^2(n^3-n)},$$

где m – число анализируемых порядковых переменных; n – число статистически исследуемых объектов (объем выборки); R_{ij} – ранг, присвоенный i-му объекту по переменной j, $i=1,2,\ldots,n$, $j=1,2,\ldots,m$;

$$D_i=\sum_{j=1}^{m}R_{ij}-\frac{m(n+1)}{2}$$

– отклонение суммы рангов по всем переменным для i-го объекта от средней их суммы для всех объектов. При наличии связанных рангов, т.е. неразличимых по некоторой переменной объектов, которым приписан одинаковый средний ранг, К.к. имеет несколько иной вид:

$$W=\frac{12\sum_{i=1}^{n}D_i^2}{m^2(n^3-n)-mB},\qquad B=\sum_{k=1}^{s}(B_k^3-B_k),$$

где s – число связок рангов (число групп неразличимых рангов), B_k – число связанных рангов в k-й группе, $k=1,2,\ldots,s$. Коэффициент W принимает значения в интервале $0\le W\le 1$, причем $W=1$, когда все m анализируемых ранговых переменных совпадают, $W=0$, когда никакой объект не имеет двух одинаковых рангов. Для проверки значимости К.к. W используют статистику

$$\chi^2=m(n-1)W,$$

которая при $n>7$ и справедливости нулевой гипотезы $H_0: W=0$ имеет χ^2- распределение с $\nu=n-1$ степенями свободы. Поэтому W при заданном уровне значимости α признается отличным от нуля (т.е. гипотеза H_0 отклоняется), если

$$m(n-1)W>\chi^2_{кр}=\chi^2_{\alpha,n-1},$$

где $\chi^2_{\alpha,n-1}$ – α-кваниль χ^2-распределения с $n-1$ степенью свободы определяется по специальным статистическим табл.

КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ

служит для измерения тесноты связи между признаками, измеренными в порядковой шкале. К. Спирмэном в 1904 предложен показатель тесноты связи между рангами r и s переменных $X=(x_1,\ldots,x_n)$ и $Y=(y_1,\ldots,y_n)$:

$$\rho = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(r_i - s_i)^2,$$

где r_i и s_i – ранги i-го объекта по переменным X и Y, n – число пар наблюдений. Если ранги всех объектов равны ($r_i = s_i$, $i = 1,2,\ldots,n$), то $\rho = 1$ (полная прямая связь), при полной обратной связи, когда $r_i = n - s_i + 1$, - $\rho = -1$, во всех остальных случаях $|\rho| < 1$. Указанная формула пригодна лишь в случае отсутствия связанных рангов в обеих ранжировках, т.е. когда все объекты имеют разные ранги. В противном случае неразличимым объектам, приписываются одинаковые средние ранги. Для корректировки формулы на общий случай определим для каждой из ранговых переменных r и s величину:

$$T_r = \frac{1}{12}\sum_{j=1}^{k_r}(R_j^3 - R_j)^2, \quad T_s = \frac{1}{12}\sum_{j=1}^{k_s}(S_j^3 - S_j)^2,$$

где k_r, k_s – число групп неразличимых рангов у переменных X и Y, а R_i, S_i – число элементов (рангов), входящих в j-ю группу неразличимых рангов. Тогда Р.к.к. Спирмэна следует вычислять по формуле:

$$\rho = \frac{\frac{1}{6}(n^3 - n) - \sum_{i=1}^{n}(r_i - s_i)^2 - T_r - T_s}{\sqrt{\left[\frac{1}{6}(n^3 - n) - 2T_r\right]\cdot\left[\frac{1}{6}(n^3 - n) - 2T_s\right]}}.$$

Если T_r и T_s являются небольшими относительно

$$\frac{1}{6}(n^3 - n)$$

величинами, то можно воспользоваться приближенным соотношением:

$$\rho = 1 - \frac{\sum_{i=1}^{n}(r_i - s_i)^2}{\frac{1}{6}(n^3 - n) - (T_r + T_s)}.$$

Предположим, два эксперта оценили качество 10 продуктов, и присвоили им ранги:

$r = (1;2;3;4;5;6;7;8;9;10)$ – 1-й эксперт

и $s = (2;3;4;1;6;5;9;7;8;10)$ – 2-й эксперт.

Вычисляя по первой из приведенных формул К.р.к., получим:

$$\rho = 1 - \frac{6}{(10^3 - 10)}(1 + 1 + 2^2 + 0 + 1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 0) = 1 - \frac{6 \cdot 14}{990} = 0{,}915,$$

что свидетельствует о согласованности мнений экспертов.

КРИТЕРИЙ АСИММЕТРИИ И ЭКСЦЕССА

критерий статистический, используемый для проверки гипотезы о нормальном законе распределения ген. совокупности *X* по результатам выборочных наблюдений $x_1, x_2, \ldots, x_n$. Проверяется гипотеза H_0 о том, что коэффициент асимметрии *Ac* и коэффициент эксцесса *Ek* генеральной совокупности равны 0, что характеризует случайную величину, имеющую нормальный закон распределения. Распределения выборочных характеристик $\hat{A}c$ и $\hat{E}k$ асимптотически нормальны, при заданных объемах выборок *n* известны полностью и затабулированы для различных уровней значимости α. Гипотеза H_0 отвергается на заданном уровне значимости α, если не выполняется хотя бы одно из неравенств:

$$\left|\hat{A}c\right| < Ac_{1-\alpha}; \qquad \left|\hat{E}k\right| < Ek_{1-\alpha},$$

где $Ac_{1-\alpha}$ и $Ek_{1-\alpha}$ – найденные по табл. распределения выборочных характеристик $\hat{A}c$ и $\hat{E}k$ значения квантилей уровня $1-\alpha$, для которых $P(\hat{A}c < Ac_{1-\alpha}) = 1-\alpha$ и $P(\hat{E}k < Ek_{1-\alpha}) = 1-\alpha$;

$\hat{A}c$ и $\hat{E}k$ – значения выборочных коэффициентов асимметрии и эксцесса, определяемые на основе выборки $x_1, x_2, \ldots, x_n$ по формулам:

$$\hat{A}c = \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}} = \frac{\sum_{i=1}^{l}(x_i - \bar{x})^3}{\left(\sqrt{\sum_{i=1}^{l}(x_i - \bar{x})^2}\right)^3};$$

$$\hat{E}k = \frac{\hat{\mu}_4}{\hat{\mu}_2^{\,2}} - 3 = \frac{\sum_{i=1}^{l}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{l}(x_i - \bar{x})^2\right)^2} - 3,$$

где $\hat{\mu}_k$ – выборочные центральные моменты *k*-го порядка.

К.а. и э. используется при $n \geq 50$ для грубой, приближённой проверки гипотезы H_0 о нор-

мальном законе распределения ген. совокупности. Он служит гл. обр. не для проверки нормальности, а для выявления отклонения от нормального закона распределения. Более обоснованные суждения о соответствии закону распределения можно сделать с помощью *критериев согласия*.

КРИТЕРИЙ БАРТЛЕТТА

критерий статистический, предназначенный для проверки *нулевой гипотезы* об однородности дисперсий в нескольких (k>2) нормальных ген. совокупностях, т.е.

H_0: $\sigma_1^2=\sigma_2^2=\ldots=\sigma_k^2$

по k имеющимся из них независимым случайным выборкам

$x_{11}, x_{12}, \ldots, x_{1n_1}$,

$x_{21}, x_{22}, \ldots, x_{2n_2}$,…,

$x_{k1}, x_{k2}, \ldots, x_{kn_k}$

объёмом, соответственно, $n_1 \neq n_2 \neq \ldots \neq n_k$ наблюдений (выборки разного объёма). Конкурирующая с ней гипотеза – H_1: $\sigma_j^2 \neq \sigma_m^2$, где равенство не выполняется, по крайней мере, для одной пары индексов j и m; предложен английским статистиком М. С. Бартлеттом.

Критическая статистика К.Б. имеет вид:

$$\chi^2_{набл} = \frac{\nu \ln \hat{S}^2_{ср} - \sum_{i=1}^{k} \nu_i \cdot \ln \hat{S}_i^2}{1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k} \frac{1}{\nu_i} - \frac{1}{\nu}\right]},$$

где $\hat{S}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$;

$i = \overline{1,k}$ – исправленные выборочные дисперсии i-й выборки;

x_{ij} – результат j-го наблюдения i-й выборки;

$\nu_i = n_i - 1$ – число степеней свободы i -й выборки;

$\nu = \sum_{i=1}^{k} \nu_i$ – суммарное число степеней свободы;

$\hat{S}^2_{ср} = \frac{\sum_{i=1}^{k} \nu_i \cdot \hat{S}_i^2}{\nu}$ – усреднённая по выборкам исправленная дисперсия.

При выполнении нулевой гипотезы H_0 и при $\nu_i > 3$, $\chi^2_{набл}$ приближённо имеет *χ^2-распределение Пирсона* с k-1 степенями свободы.

Для проверки нулевой гипотезы строят правостороннюю *критическую область*, границу которой определяют по табл. распределения χ^2 для уровня значимости критерия α и числа степеней свободы k-1 из условия:

$P(\chi^2 > \chi^2_{кр}(\alpha; k-1)) = \alpha$.

Критерий проверки гипотезы заключается в следующем:

если выполняется условие $\chi^2_{набл} > \chi^2_{кр}(\alpha; k\text{-}1)$,

то гипотезу H_0 отвергают с вероятностью ошибки α, ген. дисперсии считают неоднородными хотя бы в каких-то ген. совокупностях.

Если же $\chi^2_{набл} \leq \chi^2_{кр}(\alpha; k\text{-}1)$,

то считают, что гипотеза не противоречит опытным данным, ген. дисперсии можно считать одинаковыми (вероятность справедливости такого утверждения – *мощность критерия* 1-β).

К.Б. весьма чувствителен к отклонениям законов распределений X_i для $i = \overline{1,k}$ от нормального закона. В случае принадлежности результатов измерений нормальному закону выводы остаются корректными и при очень малых объёмах анализируемых выборок. Поэтому перед применением К.Б. рекомендуется проверять исследуемые совокупности на нормальность. При отклонении же закона распределения наблюдаемого показателя от нормального распределения статистика К.Б существенно отличается от χ^2-распределения. При этом распределении статистики становятся более зависимыми от объёма выборки, чем в случае нормального закона. При плосковершинных по сравнению с нормальным законом распределения (с отрицательными значениями коэффициента эксцесса), классический К.Б. затушевывает разницу в дисперсиях, а в случае более островершинных (при положительных коэффициентах эксцесса) – находит различия в дисперсиях, когда их нет. Т.о., если наблюдаемый закон отличается от нормального, применение классического К.Б. недопустимо. В таких случаях можно рекомен-

довать воспользоваться методикой статистического моделирования и последующего компьютерного анализа полученной закономерности или воспользоваться критериями менее чувствительными к требованию нормальности закона распределения – напр., критериями Левена или Брауна-Форсайта.

В случае, когда данные представлены в виде выборок из нескольких нормальных ген. совокупностях одинакового объёма

$n_1 = n_2 = \ldots = n_k$,

для проверки нулевой гипотезы H_0:

$\sigma_1^2=\sigma_2^2=\ldots=\sigma_k^2$ используют *критерий Кохрана.*

КРИТЕРИЙ ВИЛКОКСОНА

или критерий Вилкоксона-Манна-Уитни – *критерий непараметрический*, предназначенный для проверки *нулевой гипотезы* об однородности двух *ген. совокупностей* (или однородности двух выборок); один из исторически первых *критериев ранговых*, предложенный в 1945 Фрэнком Вилкоксоном, в 1947 существенно переработанный и расширенный Х. Б. Манном и Д. Р. Уитни.

Нулевая гипотеза проверяется по двум имеющимся случайным выборкам

$x_{11}, x_{12}, ..., x_{1m}$ и $x_{21}, x_{22}, ..., x_{2n}$

объёмом, соответственно, m и n наблюдений, пронумерованных так, чтобы $m \le n$ (в противном случае их просто меняют местами). Предполагается, что выборки независимы, взяты из ген. совокупностей, имеющих непрерывные функции распределения. Для проверки нулевой гипотезы выборки объединяются и по объединенной выборке объёма m+n строится общий *вариационный ряд.* $R_i^{(1)}$ – порядковый номер (ранг), который получает при этом *i*-й член вариационного ряда, построенного только по первой выборке, в общем вариационном ряду, т.е. $R_i^{(1)}$ – ранг элемента $x_{1(i)}$ в общем вариационном ряду,

где $x_{1(1)}, x_{1(2)}, ..., x_{1(m)}$

– вариационный ряд, построенный только по первой выборке.

Критическая статистика К.В., т.н. «сумма рангов», имеет вид:

$$W = \sum_{i=1}^{m} R_i^{(1)} .$$

В условиях справедливости нулевой гипотезы, статистика критерия имеет асимптотически нормальное распределение $N(\mu,\sigma)$

(при $m \to \infty$, $\lim \frac{m}{n} = c > 0$) с параметрами:

$$\mu_W = M(W) = \frac{1}{2} m \left(m+n+1\right),$$

$$\sigma_W^2 = D(W) = \frac{1}{12} mn \left(m+n+1\right).$$

Затем вычисляется стандартное значение критической статистики:

$$W_{cm} = \frac{W - \mu_W}{\sigma_W} = \frac{W - \frac{1}{2} m \left(m+n+1\right)}{\sqrt{\frac{1}{12} mn \left(m+n+1\right)}} .$$

По табл. стандартного нормального закона распределения для заданного уровня значимости α находят критическое значение статистики критерия $t_{кр}$. Напр., если табл. нормального закона содержат квантили нормального распределения, находят квантиль уровня

$1 - \frac{\alpha}{2}$ (или $100\alpha/2$ %-ную точку)

стандартного нормального закона, определяющую критическое значение критерия

$t_{кр} = t_{1-\alpha/2}$.

Нулевая гипотеза отвергается с вероятностью ошибки α, если $|W_{cm}| > t_{кр}$, и не отвергается в противном случае.

Существуют табл. процентных точек критической статистики критерия (посчитанные в условиях справедливости нулевой гипотезы однородности) и для доасимптотического случая $n \le 8$ [1].

Критическая статистика Манна-Уитни определяется по формуле:

$$U = W - \frac{1}{2} m \left(m+1\right).$$

В условиях справедливости нулевой гипотезы, статистика Манна-Уитни также имеет асимпто-

тически нормальное распределение N(μ,σ) с параметрами:

$$\mu_U = M(U) = \frac{1}{2}mn;$$

$$\sigma_U^2 = D(U) = D(W) = \frac{1}{12}mn\,(m+n+1).$$

И далее критерий проверяется аналогично статистике Вилкоксона – стандартизацией статистики U и нахождением критического значения с помощью табл. нормального распределения.

Поскольку W и U линейно связаны, то часто говорят не о двух критериях – К.В. и критерии Манна-Уитни, а об одном – критерии Вилкоксона, Манна и Уитни.

Значение статистики критерия не меняется при любом монотонном преобразовании шкалы измерения (т.е. он пригоден для статистического анализа данных, измеренных не только в количественной, но и в порядковой шкале).

Существует несколько способов использования критерия и несколько вариантов табл. критических значений, соответствующих этим способам.

КРИТЕРИЙ ДАРБИНА-УОТСОНА

критерий проверки наличия *автокорреляции* остатков, получаемых при использовании регрессионных моделей, при построении моделей временных рядов. Статистика Дарбина-Уотсона имеет вид:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2},$$

где $e_i = y_i - \hat{y}_i$ – величина отклонения фактического значения i-ого наблюдения y_i от расчётного $\hat{y}_i$ ($i = 1,2,...,n$). Можно показать, что величина d примерно равна $d = 2\cdot(1-r_1)$, *где* r_1 – коэффициент автокорреляции первого порядка (т.е. *парный коэффициент корреляции* между двумя последовательностями остатков $e_1, e_2, ..., e_{n-1}$ и $e_2, e_3, ..., e_n$.). Из последней формулы видно, что близость значения статистики d к нулю означает наличие высокой положительной автокорреляции (коэффициент r_1 близок к единице); близость значения статистики d к четырем означает наличие высокой отрицательной автокорреляции (коэффициент r_1 близок к минус единице). Естественно, в случае отсутствия автокорреляции значение статистики d будет близким к двум (коэффициент r_1 не сильно отличается от нуля). Применение на практике К.Д.-У. основано на сравнении расчётного значения статистики d с пороговыми, граничными значениями d_u и d_ℓ ($u-$upper– индекс верхней границы, $\ell-$low – индекс нижней границы).

Граничные значения d_u и d_ℓ, зависящие от числа наблюдений n, количества объясняющих переменных в модели, уровня значимости α, находятся по табл. (авторами критерия составлены табл. для α=0,05 и α=0,01). Алгоритм выявления автокорреляции остатков на основе К.Д.-У : выдвигается гипотеза H_0 об отсутствии автокорреляции остатков. Пусть альтернативная гипотеза состоит в наличии в остатках положительной автокорреляции первого порядка. Тогда при сравнении расчётного значения статистики d (d<2) с d_u и d_ℓ возможны варианты: 1. если $d < d_\ell$, то гипотеза H_0 об отсутствии автокорреляции отвергается (с вероятностью ошибки, равной α) в пользу гипотезы о положительной автокорреляции; 2. если $d > d_u$, то гипотеза H_0 не отвергается; 3. если $d_\ell \le d \le d_u$, то нельзя сделать определенный вывод по имеющимся исходным данным (значение d попало в область неопределенности).

Если альтернативной является гипотеза о наличии в остатках отрицательной автокорреляции первого порядка, то с пороговыми, граничными значениями d_u и d_ℓ сравнивается величина $4-d$ (при d>2). При этом возможны варианты: 1. если $4-d < d_\ell$, то гипотеза H_0 об отсутствии автокорреляции отвергается (с вероятностью ошибки, равной α) в пользу гипотезы об отрицательной автокорреляции; 2. если $4-d > d_u$, то гипотеза H_0 не отвергается; 3. если $d_\ell \le 4-d \le d_u$, то нельзя сделать определённый вывод по имеющимся исходным данным.

Данный критерий нельзя использовать, если среди объясняющих переменных содержатся

лагированные значения результативного показателя (напр., он не применим к моделям авторегрессии).

КРИТЕРИЙ ЗНАКОВ

критерий непараметрический, который используется для проверки гипотезы об однородности наблюдений двух зависимых выборок на основе измерений, сделанных по шкале не ниже ранговой. Критерий предназначен для установления общего направления сдвига между исследуемыми признаками. Он позволяет установить, изменяются ли показатели в сторону улучшения, повышения или усиления или, наоборот, в сторону ухудшения, понижения или ослабления.

Основу К.з. составляют наблюдения над случайными переменными x и y, полученные при рассмотрении двух зависимых выборок объема n. Расчёт наблюдаемого значения статистики критерия знаков включает следующие этапы: составляется n пар вида (x_i, y_i), где x_i, y_i – результаты наблюдений над случайными переменными x и y для объекта i; определяется направление сдвига в сравниваемых наблюдениях: элементам каждой пары (x_i, y_i) ставится в соответствие величина $z_i = y_i - x_i$, и паре присваивается знак « + », если $z_i > 0$, знак « – », если $z_i < 0$, « 0 », если $z_i = 0$. относительно z_i предполагается, что они во-первых взаимно независимы. Отметим, что при этом независимости между элементами x_i и y_i с одинаковым номером i не требуется. Это весьма важно на практике, когда наблюдения делаются для одного объекта и тем самым могут быть зависимы. И во-вторых, все z_i имеют равные нулю медианы, т.е. $p\ (z_i < 0) = p\ (z_i > 0) = 0{,}5$. подчеркнем, что законы распределения разных z_i могут не совпадать; подсчитывается общее число парных наблюдений $\boldsymbol{n}$, имеющих различия, т.е. рассматриваются пары, которым присвоены знаки « + » или « – »; подсчитывается меньшее число однозначных результатов сравнения m. для этого среди оставшихся n пар подсчитывается число пар со знаком « – » и число пар со знаком « + ». m равно мин. из этих чисел. если объём выборки $n \le 25$, наблюдаемое значение статистики совпадает с m: $z_{набл} = m$, при больших объемах выборки $n > 25$ – вычисляется по формуле:

$$Z_{\text{набл}} = \frac{(M + 0{,}5) - (n/2)}{\sqrt{n}/2}.$$

Нулевая гипотеза Hо состоит в том, что в состоянии изучаемых переменных нет значимых различий, т.е. преобладание направления сдвига является случайным. Альтернативная гипотеза H_1 предполагает, что для одной и той же совокупности объектов законы распределения величин X и Y различны, т. е. преобладание направления сдвига не является случайным. Для проверки нулевой гипотезы H_0 на уровне значимости α строят левостороннюю критическую область. Если объём выборки $n \le 25$, границу критической области $G_{кр}$ находят по специальным табл. из условия $P(G > G_{кр}(\alpha; n)) = \alpha$; при больших объемах выборки $n > 25$ – по табл. нормального закона распределения. Если наблюдаемое значение статистики $Z_{набл}$ не превосходит критического, то гипотезу отвергают, в противном случае считают, что гипотеза не противоречит опытным данным. К.з. предъявляет к тестируемой выборке только одно требование: шкала измерений должна быть порядковой, интервальной или относительной (т.е. тест нельзя применять к номинальным переменным). Других ограничений (в т.ч. и на форму распределения) нет. С одной стороны, это делает тест широко применимым, а с другой – снижает его мощность, поскольку тест не может опираться в своей работе на какие-либо предположения о свойствах анализируемого распределения. Невысокая мощность К.з. особенно сильно проявляется на небольших выборках. К.з. также может применяться и для проверки гипотез о значении медианы или о значении доли признака в ген. совокупности X при незначительных изменениях процедуры определения наблюдаемого значения статистики критерия. При проверке гипотезы о значении медианы H_0: $Me = Me_0$ под z_i понимается разность между i–м наблюдением выборки и предполагаемым значением параметра распределения:

$z_i = x_i - Me_0$.

При проверке гипотезы о значении доли признака H_0: $p = p_0$ для выборки объема n под M

понимают мин. из значений m и $n - m$, где m – число наблюдений выборки, обладающих рассматриваемым признаком:

$M = min\ \{m;\ n - m\}$.

КРИТЕРИЙ КОХРАНА

критерий *статистической проверки гипотезы* об однородности *дисперсий* нескольких независимых выборочных совокупностей, распределенных в соответствии с нормальным законом

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_l^2$,

выборки из которых имеют одинаковый объём. При проверке гипотезы рассматриваются нормально распределённые *ген. совокупности*

$X_1, X_2, \ldots, X_l$

с неизвестными дисперсиями $\sigma_1^2, \sigma_2^2, \ldots, \sigma_l^2$,

из которых взяты независимые случайные выборки одинакового объёма

$n_1 = n_2 = \ldots = n_l$,

на основе которых получены оценки выборочных дисперсий соответствующих совокупностей $S_1^2, S_2^2, \ldots, S_l^2$.

К.К. основан на статистике:

$$G_{набл} = \frac{S^2_{\max}}{S_1^2 + S_2^2 + \ldots + S_l^2},$$

где S^2_{max}, наибольшая из выборочных дисперсий. Наблюдаемое значение статистики можно рассчитать и на основе несмещённых оценок дисперсий ген. совокупностей по формуле:

$$G_{набл} = \frac{\hat{S}^2_{\max}}{\hat{S}_1^2 + \hat{S}_2^2 + \ldots + \hat{S}_l^2}.$$

При истинности нулевой гипотезы статистика $G_{набл}$ имеет G – распределение со степенями свободы $\nu_1 = n - 1$ числителя и $\nu_2 = l$ знаменателя.

Для проверки нулевой гипотезы H_0 на уровне значимости α строят правостороннюю критическую область. Границу критической области $G_{кр}$ находят по таблицам G-распределения из условия

$P(G > G_{кр}(\alpha; n - 1; l)) = \alpha$.

Критерий проверки гипотезы заключается в следующем: если выполняется условие

$G_{набл} > G_{кр}(\alpha; n-1; l)$,

то гипотезу отвергают, если же

$G_{набл} \leq G_{кр}(\alpha; n-1; l)$,

то считают, что гипотеза не противоречит опытным данным.

КРИТЕРИЙ ЛОГАРИФМА ОТНОШЕНИЯ ПРАВДОПОДОБИЯ

статистический критерий, основанный на статистике, представляющей собой логарифм отношения функций правдоподобия. Построение критерия рассмотрим на примере простой осн. гипотезы. Предположим, что наблюдаемые значения $x_1, \ldots, x_n$ можно рассматривать как независимую выборку из распределения, принадлежащего к классу распределений $F(x, \Theta)$, где Θ – k-мерный параметр. Проверяемая гипотеза $H_0 : \Theta = \Theta_0$. В качестве альтернативной гипотезы рассматривается гипотеза

$H_1 : \Theta = \Theta_{МП}$,

где $\Theta_{МП}$ – оценка макс. правдоподобия параметра Θ по выборке $x_1, \ldots, x_n$. Рассмотрим статистику:

$$\lambda = \lambda(x_1, \ldots, x_n) = -2 \ln \left[\frac{L(x_1, \ldots, x_n; \Theta_0)}{L\ (x_1, \ldots, x_n; \Theta_{МП})} \right].$$

Если семейство функций $F(x, \Theta)$ удовлетворяет условиям регулярности, то оценки макс. правдоподобия являются асимптотически наилучшими. В этом случае доказано, что введённая статистика λ имеет асимптотически, при $n \to \infty$, χ^2-распределение с k степенями свободы. Т.о., для нахождения критической точки $\chi^2_{\alpha,k}$ при заданном уровне значимости α достаточно обратиться к стандартным статистическим табл. χ^2 - распределения. Гипотеза H_0 отклоняется, если вычисленное значение статистики λ попадает в критическую область (в зависимости от соотношения между Θ_0 и $\Theta_{МП}$ выбирается правосторонняя или левосторонняя критическая область).

КРИТЕРИЙ НЕСМЕЩЁННЫЙ

критерий проверки статистических гипотез, обладающий свойством несмещённости. Любой статистический критерий характеризуется уровнем значимости α – вероятностью отклонить верную нулевую гипотезу (*ошибка первого рода*), и вероятностью ошибки второго рода β - вероятностью отклонить верную конкурирующую гипотезу. При этом величина

$$W = 1 - \beta,$$

называемая *мощностью критерия*, определяет вероятность принять верную конкурирующую гипотезу. Хороший критерий проверки гипотез наряду с макс. мощностью должен обладать одним важным свойством: вероятность отвергнуть правильную осн. гипотезу α должна быть меньше или равна вероятности W принять правильную конкурирующую гипотезу. Критерий проверки гипотез называется несмещённым, если для заданных гипотез мощность критерия не меньше заданного уровня значимости, т.е. для мощности критерия выполняется условие:

$$W = 1 - \beta \geq \alpha.$$

Другими словами, если критерий S предназначен для проверки одной (осн.) гипотезы, то он должен давать правильный ответ с большей вероятностью (чаще) при справедливости этой гипотезы, чем при справедливости других (т.е. правильности любых конкури-рующих). К.н., обеспечивающий наибольшую мощность среди других возможных несмещённых критериев, называется наиболее мощным несмещенным критерием. При этом следует оговорить, что понятие несмещённости критерия в задачах проверки гипотез не связано с определением несмещённости при точечных оценках параметров распределений.

КРИТЕРИЙ ОДНОРОДНОСТИ ДИСПЕРСИЙ

критерий, который служит для проверки *гипотез статистических* о равенстве дисперсий выборок. Наиболее употребительны критерии для нормальных *ген. совокупностей*. Для проверки однородности дисперсий двух нормальных ген. совокупностей используется т.н. F -критерий. Пусть имеются две ген. совокупности, дисперсии которых равны σ_1^2 и σ_2^2. Проверяемая нулевая гипотеза

$$H_0 : \sigma_1^2 = \sigma_2^2,$$

конкурирующая – $H_1 : \sigma_1^2 \neq \sigma_2^2$, $H_1 : \sigma_1^2 > \sigma_2^2$ или $H_1 : \sigma_1^2 < \sigma_2^2$.

F -критерий основан на использовании статистики:

$$F = \frac{s_1^2}{s_2^2}, \quad s_j^2 = \frac{n_j}{n_j - 1}\hat{\sigma}_j^2, \quad j = 1,2,$$

где n_1 и n_2 – объёмы независимых выборок из исследуемых ген. совокупностей, $\hat{\sigma}_j^2$ – выборочные дисперсии, оцененные по выборкам, s_j^2 – «исправленные» выборочные дисперсии. Данная статистика при истинности гипотезы H_0 имеет F -распределение Фишера-Снедекора с числами степеней свободы числителя и знаменателя, равными соответственно

$$\nu_1 = n_1 - 1 \text{ и } \nu_2 = n_2 - 1.$$

При формировании критерия принятия гипотезы H_0 следует учесть, что распределение статистики F является несимметричным. Поэтому гипотеза H_0 при заданном уровне значимости α принимается,

если $F < F_{кр} = F_{\alpha,\nu_1,\nu_2}$

в случае правосторонней критической области (для альтернативной гипотезы $H_1 : \sigma_1^2 > \sigma_2^2$),

либо если $F > F_{кр} = F_{1-\alpha,\nu_1,\nu_2}$

в случае левосторонней критической области (для альтернативной гипотезы $H_1 : \sigma_1^2 < \sigma_2^2$),

либо если $F_{1-\alpha/2,\nu_1,\nu_2} < F < F_{\alpha/2,\nu_1,\nu_2}$

для двусторон-ней критической области

($H_1 : \sigma_1^2 \neq \sigma_2^2$).

Здесь F_{α,ν_1,ν_2} – α -квантиль распределения Фишера-Снедекора, определяемая по известным статистическим табл. Предположим, имеются две выборки объёмов $n_1 = 15$ и $n_2 = 18$, для которых рассчитаны выборочные дисперсии

$$\hat{\sigma}_1^2 = 8{,}5 \text{ и } \hat{\sigma}_2^2 = 6{,}3.$$

Требуется проверить на уровне значимости $\alpha = 0{,}05$ гипотезу о равенстве дисперсий при альтернативной гипотезе $H_1 : \sigma_1^2 > \sigma_2^2$. Вычислим значение статистики F .

$$F=\frac{s_1^2}{s_2^2}=\frac{(15/14)8{,}5}{(18/17)6{,}3}=1{,}37.$$

Критическое значение F -критерия при $\alpha=0{,}05$, $\nu_1=n_1-1=14$ и $\nu_2=n_2-1=17$ равно $F_{кр}=F_{0{,}05;14;17}=2{,}23$.

Т.к. $F<F_{кр}=F_{0{,}05;14;17}$, то гипотеза H_0 принимается. Для проверки гипотезы однородности нескольких (с том числе более чем двух) дисперсий используют *критерий Бартлетта*, статистика которого имеет вид:

$$B=\frac{1}{q}\sum_{j=1}^{k}(n_j-1)\ln\left(\frac{\overline{s^2}}{s_j^2}\right),\quad q=1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_j-1}-\frac{1}{n_1+\ldots+n_k-k}\right),$$

Здесь k – число исследуемых ген. совокупностей, s_j^2 – «исправленные» выборочные дисперсии j -й выборки ($j=1,\ldots k$), $\overline{s^2}$ – оценка средней арифметической дисперсий

$$\overline{s^2}=\frac{15\cdot 8{,}5+18\cdot 6{,}3+25\cdot 9{,}3+32\cdot 5{,}8}{15+18+25+32-4}=\frac{659}{86}=7{,}66,$$

$$\overline{s^2}=\frac{1}{n_1+\ldots+n_k-k}\sum_{j=1}^{k}n_j\hat{\sigma}_j^2,\ \hat{\sigma}_j^2$$

– выборочные дисперсии. Проверяемая гипотеза $H_0:\sigma_1^2=\ldots=\sigma_k^2$; конкурирующая гипотеза H_1: некоторые дисперсии могут быть различными. Указанная статистика при $\min(n_1,\ldots,n_k)>3$ в условиях справедливости гипотезы H_0 и имеет χ^2– распределение с $\nu=k-1$ степенями свободы. Поэтому для заданного уровня значимости α при

$$\chi^2>\chi_{кр}^2=\chi_{\alpha,k-1}^2$$

гипотезу об однородности дисперсий следует отвергнуть. Здесь

$\chi_{\alpha,k-1}^2$ – α -квантиль

χ^2 -распределения с $(k-1)$ степенями свободы.

Если все выборки имеют одинаковый объём $n_1=\ldots=n_k=n$, то для проверки гипотезы H_0 против гипотезы H_1 можно воспользоваться менее мощным, но зато более простым *критерием Кохрана*. Статистика критерия Кохрана задается формулой:

$$G=\frac{s_{\max}^2}{s_1^2+\ldots+s_k^2},\quad s_{\max}^2=\max_{1\le j\le k}s_j^2,$$

и при справедливости гипотезы H_0 имеет G -распределение с числами степеней свободы $\nu_1=k$ и $\nu_2=n-1$. Критерий Кохрана предписывает принять гипотезу H_0 на уровне значимости α, если

$G<G_{кр}=G_{\alpha,k,n-1}$,

где $G_{\alpha,k,n-1}$ – α -квантиль

G -распределения. Гипотезы о дисперсиях возникают достаточно часто, так как дисперсия характеризует такие исключительно важные показатели, кик точность машин и приборов, технологических процессов, степень однородности совокупностей, финансовые риски (отклонение доходности активов от ожидаемого уровня) и др.

КРИТЕРИЙ ω^2

статистический *критерий согласия*, проверяющий, согласуется ли эмпирические данные с некоторым гипотетическим предположением относительно теоретической функции распределения. Рассмотрим выборку $x_1,\ldots,x_n$, произведенную из *ген. совокупности* с неизвестной теоретической функцией распределения, относительно которой имеются две непараметрические гипотезы: простая осн.

$H_0:F(x)=F_0(x)$

и сложная конкурирующая

$H_1:\ \ F(x)\neq F_0(x)$,

где $F_0(x)$ – предполагаемая известная функция распределения. *Критерий согласия Колмогорова* хорошо разделяет выборки из ген. совокупностей с теоретическими функциями распределения $F_0(x)$ и $F_1(x)$, если $|F_0(x)-F_1(x)|$

достаточно велико хотя бы на малом интервале изменения x. Встречается и обратная ситуация, когда $|F_0(x)-F_1(x)|$ мало, но постоянно на достаточно большом интервале изменения x.

В этом случае для разделения гипотез H_0 и H_1 естественно пользоваться каким-либо интегральным расстоянием. Статистика ω^2 критерия ω^2 задается выражением:

$$\omega^2 = \omega^2(x_1,\ldots,x_n) = n\int_{-\infty}^{\infty}[F^*(x)-F_0(x)]^2\,p_0(x)dx,$$

где $p_0(x)$ – плотность вероятностей гипотетической функции $F_0(x)$, $F^*(x)$ – эмпирическая функция распределения, определённая по имеющимся данным. Критическая область $\Omega_{кр}$ состоит из всех точек $(x_1,\ldots,x_n)$, для которых $\omega^2 > C$, где C – критическое значение критерия. Используя исходные статистические данные, статистику ω^2 можно записать в более удобном для практических расчётов виде:

$$\omega^2 = \sum_{k=1}^{n}\left[F_0(x_k)-\frac{2k-1}{2n}\right]^2+\frac{1}{12n}.$$

Распределение статистики ω^2 при условии справедливости гипотезы H_0 не зависит от гипотетической функции распределения $F_0(x)$ и при увеличении объёма выборки (уже при $n\ge 40$) сходится к ω^2-распределению. Поэтому уровень значимости критерия определяется по критическому значению C приближённой формулой $\alpha \approx 1-A(C)$, где $A(C)$ – функция ω^2-распределения. Если же задан уровень значимости α критерия, то критическое значение C практически совпадает с $(1-\alpha)$-квантилью $a_{1-\alpha}$ ω^2-распределения. Практическая реализация критерия ω^2 осуществляется по стандартной схеме: по выборке находятся значения $F_0(x_k)$ и вычисляется значение статистики ω^2, которое сравнивается с критическим значением C: если $\omega^2 < C$, то гипотеза H_0 не противоречит эмпирическим данным и принимается; в противном случае предпочтение отдается *альтернативной гипотезе* H_1.

КРИТЕРИЙ ОТНОШЕНИЯ ПРАВДОПОДОБИЯ

статистический критерий, основанный на статистике, представляющей собой отношение функций правдоподобия, построенных при условиях справедливости *альтернативной* и осн. *гипотез*. Рассмотрим случай построения статистики К.о.п. для различия двух простых гипотез. Пусть выборка $x_1,\ldots,x_n$ произведена из *ген. совокупности* с теоретической функцией распределения $F(x)$, относительно которой имеются две простые гипотезы: осн.

$$H_0 : F(x) = F_0(x)$$

и конкурирующая

$$H_1 : F(x) = F_1(x),$$

где $F_0(x)$ и $F_1(x)$ – известные функции распределения. Введём статистику:

$$\Lambda = \Lambda(x_1,\ldots,x_n) = \frac{L_1(x_1,\ldots,x_n)}{L_0(x_1,\ldots,x_n)},$$

где $L_j = \prod_{i=1}^{n} f_j(x_i)$ – функция правдоподобия при условии, что истинной является гипотеза $H_j, j=0,1$; $f_j(x)$ – соответствующие плотности распределений $F_j(x), j=0,1$.Статистика Λ носит название отношения правдоподобия. Естественно предположить, что чем больше отношение правдоподобия, тем большее предпочтение мы должны оказать гипотезе H_1. Т.о. критическая область К.о.п. состоит из всех точек $(x_1,\ldots,x_n)$, для которых $\Lambda(x_1,\ldots,x_n)$ больше критического значения $\Lambda_{кр}$. Если известна плотность распределения $f_\Lambda(u)$ статистики Λ при условии справедливости гипотезы H_0, то построение критерия проверки справедливости гипотезы H_0 при заданном уровне значимости α сводится к определению критического значения $\Lambda_{кр}$ из условия:

$$\int_{\Lambda_{кр}}^{\infty} f_\Lambda(u)du = \alpha.$$

Если $\Lambda > \Lambda_{кр}$, то гипотеза H_0 отвергается с вероятность ошибки α, если $\Lambda < \Lambda_{кр}$, то гипотеза H_0 считается не противоречащей исходным данным и принимается. Предпочтительность использования данного критерия диктуется очень важным его свойством, доказанным в виде леммы Неймана-Пирсона, кото-

рая утверждает, что для случая сравнения двух простых гипотез К.о.п. является наиболее мощным среди всех других критериев. Другими словами, для К.о.п. вероятность ошибки второго рода (отклонение гипотезы H_1, когда она на самом деле верна) минимальна.

КРИТЕРИЙ ПОСЛЕДОВАТЕЛЬНЫЙ

критерий проверки *гипотез статистических*, основанный на последовательной схеме наблюдений, при которой число необходимых наблюдений определяется в процессе эксперимента, т.е. не фиксируется заранее. В этих условиях число наблюдений (объём выборки) – случайная величина, зависящая от результатов наблюдений. Впервые К.п. был использован при выборочном статистическом контроле качества продукции. Логическая схема построения К.п. та же, что и для критериев с фиксированным числом наблюдений, с одним отличием: последовательно для каждого фиксированного объёма выборки $n = 1,2,\ldots k, k+1,\ldots$ область Ω возможных значений статистики критерия $T(x_1,\ldots,x_n) = T_n(x)$ разбивается на три непересекающиеся части: область $\Omega_Д$ правдоподобных, область $\Omega_{кр}$ неправдоподобных и область Ω_* сомнительных в условиях справедливости проверяемой гипотезы H_0 значений. Т.о. область Ω_* ограничена некоторыми константами A_0 и A_1. Наблюдения производятся до тех пор, пока не будет впервые нарушено какое-нибудь из неравенств: $A_0 < T_n(x) < A_1$. Если в момент прекращения испытаний $T_n(x) \le A_0$ (т.е. $T_n(x) \in \Omega_Д$), то принимается гипотеза H_0; если

$T_n(x) \ge A_1$ (т.е. $T_n(x) \in \Omega_{кр}$),,

то проверяемая гипотеза H_0 отвергается, или принимается некоторая альтернатива H_1; при

$A_0 < T_n(x) < A_1$ (т.е. $T_n(x) \in \Omega_*$),

производится следующее $(n+1)$ наблюдение. Эта процедура характеризуется вероятностями *ошибок первого* и второго *рода*

$\alpha = P(H_1 \mid H_0)$ и $\beta = P(H_0 \mid H_1)$

и средним числом

$N_j(n) = E(\nu \mid H_j)$

наблюдений n до момента остановки процедуры $(j = 0;1)$. Если вероятности ошибок α и β заданы, то любой критерий с такими ошибками называется критерием силы (α, β). В классе критериев данной силы предпочтение отдается тому, для которого требуется меньше наблюдений. Критерий, оптимизирующий одновременно как $N_0(n)$, так и $N_1(n)$, называется оптимальным.

Т.о. для применения К.п. надо указать: а) проверяемую гипотезу; б) способ построения критической статистики $T_n(x)$; в) способ построения областей $\Omega_Д, \Omega_{кр}$ и Ω_*.

В качестве конкретного примера последовательного критерия рассмотрим известный *критерий отношения правдоподобия* Вальда (1947), предназначенный для различия двух гипотез по результатам независимых наблюдений $X_1, X_2 \ldots$. Гипотеза H_0 заключается в том, что случайные величины X_i имеют распределение вероятностей с плотностью $f_1(x)$, а гипотеза H_1 – в том, что X_i имеют плотность $f_2(x)$. Обозначим через $L_{jn} = \prod_{i=1}^{n} f_j(x_i)$ функцию правдоподобия для первых n испытаний при условии, что истинной является гипотеза $H_j, j = 0,1$. Статистика критерия Вальда определяется отношением правдоподобия: $\lambda_n = L_{1n} / L_{0n}$, а границы области неопределённости A_0 и A_1 критерия Вальда силы (α, β) удовлетворяют неравенствам:

$$A_0 \ge A_0^* = \beta/(1-\alpha), \quad A_1 \le A_1^* = (1-\beta)/\alpha.$$

При этом, если границы A_0 и A_1 заменить их оценками A_0^* и A_1^*, то сила полученного критерия будет

(α^*, β^*),

где $\alpha^* \le \alpha/(1-\beta), \quad \beta^* \le \beta/(1-\alpha)$

и $\alpha^* + \beta^* = \alpha + \beta$.

Для практических целей удобнее рассматривать не величины λ_n, а их логарифмы. Тогда прекращение проверки и принятие одной из гипотез будет при первом n, при котором нарушается какое-нибудь неравенство:

$$b_0 = \ln A_0^* < \ln \lambda_n < \ln A_1^* = b_1, \quad \text{где} \quad b_0 < 0, \quad b_1 > 0, \quad \ln \lambda_n = \sum_{i=1}^{n} \ln \frac{f_1(x_i)}{f_0(x_i)}.$$

Пусть, напр., X_i имеют нормальное распределение с плотностью

$$f(x,a,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-a)^2/2\sigma^2},$$

где $\sigma = 1$ – известная величина. Гипотеза H_0 состоит в том, что $a = a_0 = 0,$ а гипотеза H_1 – в том, что $a = a_1 = 0{,}5,$ и пусть $\alpha = 0{,}01, \quad \beta = 0{,}04.$ Используя приведённые соотношения, найдём:

$$A_0^* = \frac{\beta}{(1-\alpha)} = \frac{4}{99}, \quad A_1^* = \frac{1-\beta}{\alpha} = 96, \quad b_0 = \ln A_0^* = -3{,}21, \quad b_1 = \ln A_1^* = 4{,}56,$$

$$\ln \lambda_n = 0{,}5 \sum_{i=1}^{n} x_i - 0{,}25n.$$

Поэтому неравенства $\lambda_n < 4/99$ и $\lambda_n > 96$ равносильны неравенствам

$\sum_{i=1}^{n} x_i < 0{,}25n - 3{,}21$ и $\sum_{i=1}^{n} x_i > 0{,}25n + 4{,}56$,

соответственно. Критерию можно дать наглядную геометрическую интерпретацию (см. рис. 1). На плоскости (xOy) наносятся две прямые $y = 0{,}25n - 3{,}21$ и $y = 0{,}25n + 4{,}56$, и ломаная линия с вершинами в точках

$(n, \sum_{i=1}^{n} x_i), \ n = 1,2,\ldots .$

Если ломаная впервые выходит из полосы, ограниченной этими прямыми, через верхнюю границу, принимается гипотеза H_1, если – через нижнюю, принимается H_0 .Для К.п. Вальда доказано, он заканчивается за конечное число шагов. Т.к. границы b_0 и b_1 зависят только от α и β, а отношение $\lambda_n = L_{1n}/L_{0n}$ можно вычислить на основе *исходных статистических данных*, то не возникает вопроса о нахождении распределения статистики критерия при нулевой и конкурирующей гипотезах. Необходимость иметь информацию о распределении статистики критерия возникает только при нахождении среднего числа наблюдений до принятия решений:

$$N_0(n) \approx \frac{(1-\alpha)b_0 + \alpha b_1}{M(\ln \lambda_n \mid H_0)}, \quad N_1(n) \approx \frac{\beta b_0 + (1-\alpha)b_1}{M(\ln \lambda_n \mid H_1)},$$

$$\text{где} \quad M(\ln \lambda_n \mid H_0) = \int_{-\infty}^{+\infty} \ln \frac{f_1(x_i)}{f_0(x_i)} f_0(x_i)dx, \quad M(\ln \lambda_n \mid H_1) = \int_{-\infty}^{+\infty} \ln \frac{f_1(x_i)}{f_0(x_i)} f_1(x_i)dx.$$

Для дискретной случайной величины интеграл заменяется суммой, а функции плотности вероятности $f_j(x)$ на вероятности $p_j(x_i)$. Для рассматриваемого примера:

$$M(\ln \lambda_n \mid H_j) = \begin{cases} -\dfrac{(a_1 - a_0)^2}{2\sigma^2} = -0{,}08, & \textit{при} \quad j = 0, \\ \dfrac{(a_1 - a_0)^2}{2\sigma^2} = 0{,}08, & \textit{при} \quad j = 1. \end{cases}$$

Тогда среднее число наблюдений, необходимых для принятия гипотез: $N_0 \approx 26$ и $N_1 \approx 34$, в то время как для различения указанных гипотез по выборкам фиксированного объема потребовалось бы более 60 наблюдений:

$$n^* = \left[\sigma^2 (u_\alpha + u_\beta)^2 / (a_1 - a_0)^2\right] + 1 = \left[(2{,}33 + 1.75)^2 / 0{,}5^2\right] + 1 = [60{,}84] + 1 = 61,$$

где σ^2 означает целую часть числа, u_α и u_β – соответствующие квантили нормального стандартного распределения $N(0,1)$, т.е. решения уравнений

$$\Phi(u_\alpha) = \alpha, \quad \Phi(u_\beta) = \beta.$$

Исследования показывают, что К.п. примерно в два-четыре раза выгоднее (по затратам на наблюдения), чем наилучший из классических критериев – критерий Неймана-Пирсона.

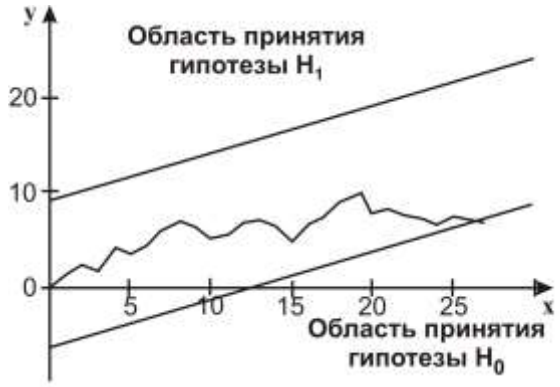


Рис. 1

КРИТЕРИЙ НЕПАРАМЕТРИЧЕСКИЙ

критерий *статистической проверки гипотезы*, который не рассматривают анализируемое статистическое распределение как функцию, его применение не предполагает предварительного вычисления параметров распределения. Т.о. К.н. основываются на более слабых допущениях в отношении анализируемых данных в сравнении со стандартными параметрическими процедурами. *Методы статистики непараметрические* разработаны для случаев, когда исследователь ничего не знает о параметрах исследуемой совокупности: отсюда и название методов. Они не основываются на оценке параметров (таких как *математическое ожидание* или *дисперсия*) при описании выборочного распределения интересующей величины. Поэтому эти методы иногда также называются свободными от параметров или свободно распределенными.

Гл. достоинством применения непараметрических методов является возможность отойти от допущений, необходимых для использования параметрических процедур. Дополнительным соображением в пользу выбора непараметрических методов служит присущая некоторым таким критериям легкость применения и простота вычислений. Кроме этого, они могут быть использованы и для случайных величин, наблюдения над которыми представлены в номинальной и порядковой шкале.

Однако эти преимущества непараметрических методов реализуются за счёт снижения их качественных характеристик. Слабое место К.н. состоит в их относительно низкой статистической мощности по сравнению со стандартными параметрическими процедурами. К.н. обычно требуют больших объёмов выборки, чтобы сравняться по статистической мощности с параметрическими критериями. В сравнении с параметрическими, К.н. менее точны, что зачастую приводит к ложному принятию нулевой гипотезы, так как для того, чтобы её отвергнуть необходимо, чтобы наблюдения выборки характеризовались более значительными отклонениями. И наконец, К.н. менее информативны, напр., позволяют определить направление сдвига в данных, но не указывают его величину.

КРИТЕРИЙ ПИРСОНА X^2

критерий проверки гипотезы о том, что изучаемая случайная величина подчиняется заданному закону распределения: H_0: $F(x) = F_0(x)$. Наблюдаемое значение статистики критерия рассчитывается на основе данных, представленных в виде вариационного ряда по формуле:

$$T_n = \chi^2_{набл} = \sum_{i=1}^{l} \frac{(m_i/n - p_i)^2}{p_i} n,$$

где m_i – частота i-го значения или интервала (число наблюдений выборки, равных i-му значению x_i, или попадающих в i-й интервал $(a_i; b_i)$, $i = 1, \ldots, l$; p_i – вероятность принятия случайной величиной i-го значения или вероятность попадания в i-й интервал; n – объём выборки $n = \Sigma\, m_i$.

Часто для расчётов вводят понятие "теоретической частоты" $m_i^T = np_i$, что позволяет преобразовать формулу наблюдаемого значения статистики критерия к виду:

$$\chi^2_{набл} = \sum_{i=1}^{l} \frac{(m_i - m_i^T)^2}{m_i^T}.$$

По теореме Пирсона при истинности гипотезы H_0 и n $\to \infty$, распределение статистики $\chi^2_{набл}$ сходится к χ^2-распределению с $v = l - r - 1$ степенями свободы, где r – число параметров предполагаемого теоретического закона, использованных для вычисления теоретических частот и оцениваемых по выборке. Для проверки нулевой гипотезы H_0 на уровне значимости α строят правостороннюю критическую область. Границу критической области $\chi^2_{кр}$ находят по табл. χ^2-распределения из условия $P(\chi^2 > \chi^2_{кр}(\alpha; l - r - 1)) = \alpha$.

Гипотеза отвергается на уровне значимости α, если вычисленное значение $\chi^2_{набл}$ окажется больше критического $\chi^2_{кр}(\alpha,v)$, найденного по табл. распределения χ^2 для уровня значимости α и числа степеней свободы $v = l - r - 1$. В противном случае гипотеза не отвергается.

По теоретическим соображениям при расчёте $\chi^2_{набл}$ не следует исходить из слишком малых значений m_i^T. Поэтому рекомендуется перед расчетами объединить соседние интервалы (варианты) т.о., чтобы $m_i^T \geq 5 \div 10$ для объединенных интервалов. Кроме того, объём выборки должен быть достаточно велик ($n \geq 50$).

Теоретические законы, как правило, определяются для всех действительных значений случайной величины. Это обстоятельство следует учитывать при получении вероятностей p_i, т.е. учитывать, если это необходимо, расширенные интервалы $(-\infty; b_1)$ и $(a_l; \infty)$. При расчёте теоретических частот иногда производят округление до целых чисел, при этом следует вычислять вероятности с такой точностью, чтобы погрешность округления была наименьшей и выполнялось равенство

$$\sum_{i=1}^{l} m_i^T = n.$$

КРИТЕРИЙ РАНГОВЫЙ

критерий, в котором вместо исходных данных наблюдаемых значений выборки используются их ранги (номера наблюдения в выборке, упорядоченной по возрастанию значений). К.р. широко используются для проверки гипотез об однородности двух выборок, т.е. гипотез о совпадении законов распределения ген. совокупностей, из которых эти выборки взяты.

Проверку гипотезы об однородности при исследовании зависимых выборок из генеральных совокупностей X и Y объёма N можно осуществить на основе *критерия знаков*, реализуемого по схеме: 1. составляется N пар вида (x_i, y_i), где x_i, y_i – результаты наблюдений над случайными переменными X и Y для объекта i; 2. определяется направление сдвига в сравниваемых наблюдениях: элементам каждой пары (x_i, y_i) ставится в соответствие величина $z_i = y_i - x_i$, и паре присваивается знак « + », если $z_i > 0$, знак « – », если $z_i < 0$, « 0 », если $z_i = 0$.; 3. 3. подсчитывается общее число парных наблюдений n, имеющих различия, т.е. рассматриваются пары, которым присвоены знаки « + » или « – »; 4. каждой разности z_i присваивается ранг. Для этого, полученные разности z_i упорядочиваются по абсолютной величине (без учёта знаков) и в упорядоченной выборке каждому z_i присваивается порядковый номер (от 1 до n). Ранг разности z_i равен ее порядковому номеру, если значение z_i встречается в выборке один раз, и среднему значению порядковых номеров, если значение z_i встречается в выборке неоднократно; 5. подсчитывается меньшая сумма рангов однозначных результатов сравнения M. Для этого среди оставшихся n пар суммируются ранги пар со знаком « – » и ранги пар со знаком « + ». M равно мин. из этих чисел. Если объём выборки $n \leq 30$, наблюдаемое значение статистики совпадает с M: $Z_{набл}$ = M, при больших

объёмах выборки $n > 30$ – вычисляется по формуле:

$$Z_{\text{набл}} = \frac{M - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}};$$

6. для проверки нулевой гипотезы H_0 на уровне значимости α строят левостороннюю критическую область. Если объём выборки $n \le 30$, границу критической области $G_{кр}$ находят по специальным таблицам из условия

$P(G > G_{кр}(\alpha; n)) = \alpha$,

при больших объёмах выборки $n > 30$ – по таблице нормального закона распределения. Если наблюдаемое значение статистики $Z_{набл}$ не превосходит критическое, то гипотезу отвергают, в противном случае считают, что гипотеза не противоречит опытным данным.

Проверку гипотезы об однородности при исследовании независимых выборок из генеральных совокупностей X и Y объёмов n_x и n_y можно осуществить на основе *критерия Вилкоксона*, реализуемого по схеме: 1. каждому наблюдению из обеих выборок присваивается ранг. Для этого, наблюдения выборок перемешиваются и упорядочиваются, и в упорядоченной выборке каждому наблюдению присваивается порядковый номер (от 1 до $n_x + n_y$). Ранг наблюдения равен её порядковому номеру, если это наблюдение встречается в выборке один раз, и среднему значению порядковых номеров, если оно встречается в выборке неоднократно; 2. подсчитывается сумма рангов первой и второй выборок. Если объём выборки $n \le 10$, наблюдаемое значение статистики совпадает с суммой рангов первой выборки: $Z_{набл} = R_x$, при объёмах выборки $n > 10$ – вычисляется по формуле:

$$Z_{\text{набл}} = \frac{R_x - \frac{n_x(n_x + n_y + 1)}{2}}{\sqrt{\frac{n_x n_y (n_x + n_y + 1)}{12}}};$$

3. для проверки нулевой гипотезы H_0 на уровне значимости α строят левостороннюю критическую область. Если объем выборки $n \le 10$, границу критической области $G_{кр}$ находят по специальным таблицам из условия

$P(G > G_{кр}(\alpha; n)) = \alpha$;

при больших объемах выборки $n > 10$ – по таблице нормального закона распределения. Если наблюдаемое значение статистики $Z_{набл}$ попадает в критическую область, то гипотезу отвергают, в противном случае считают, что гипотеза не противоречит опытным данным.

КРИТЕРИЙ СОГЛАСИЯ

статистическое правило, в соответствии с которым проверяется *гипотеза статистическая* об аналитическом виде закона распределения вероятностей анализируемой *ген. совокупности*, причём гипотеза может как однозначно задавать закон распределения, так и определять лишь его тип. Закон распределения случайной величины $\hat{F}(x)$, построенный на основе наблюдений, имеющихся в распоряжении исследователя, называется эмпирическим. Закон распределения $F_n(x)$, на соответствие которому проверяется эмпирическое распределение, называется гипотетическим. Задача критерия – проверить согласие эмпирического и гипотетического законов распределения.

При проверке рассматриваются две гипотезы. Нулевая гипотеза H_o предполагает отсутствие значимых отличий, т.е. ряд наблюдений $x_1, x_2, \ldots, x_n$ образует случайную выборку, извлеченную из ген. совокупности X с функцией распределения $F(x)=F(x;\theta_1,\theta_2,\ldots,\theta_k)$, где общий вид функции F(x) считается известным, а параметры $\theta_1,\theta_2,\ldots,\theta_k$ могут быть как известными, так и неизвестными. Альтернативная гипотеза H_0 утверждает, что различие между эмпирическим и гипотетическим распределением значимо.

К.с. основан на использовании различных мер расстояний между анализируемой эмпирической функцией $\hat{F}_n(x)$ распределения, определённой по выборке, и гипотетической функцией распределения F(x) ген. совокупности X. К.с. состоит в том, что выбирается некоторая случайная величина (статистика) T_n, являющаяся мерой расхождения (рассогласования) между рядом наблюдений и гипотетическим распределением. Случайная величина T_n, есть функ-

ция наблюдаемых относительных частот, и в зависимости от вида этой функции распределение T_n будет задавать соответствующий критерий согласия.

Для заданного уровня значимости α на основании закона распределения T_n, определяют критическое значение

$T_{n\,кр}$ так, что $P(T_n > T_{n\,кр}) = \alpha$.

Для выборки $x_1, x_2, \ldots, x_n$ вычисляется наблюдаемая величина $T_{n\,набл}$. Если $T_{n\,набл} > T_{n\,кр}$, то нулевая гипотеза отвергается. Если же $T_{n\,набл} \leq T_{n\,кр}$, то нулевая гипотеза не отвергается: в этом случае отклонения от гипотетического закона распределения считаются незначимыми, т.е. данные наблюдения не противоречат гипотезе о виде распределения.

Можно осуществлять проверку гипотезы о виде распределения с помощью критерия согласия и в другом порядке. По наблюдаемому значению ($T_{n\,набл}$) определить вероятность

$\alpha_{набл} = P(T_n > T_{n\,набл})$.

Если $\alpha_{набл} \leq \alpha$, то отклонения значимы, т.е. гипотеза отвергается, если же $\alpha_{набл} > \alpha$, то гипотеза не отвергается. Важно отметить, что значения $\alpha_{набл}$, достаточно близкие к единице, указывают на нерепрезентативность выборки; выборку следует повторить, соблюдая принцип случайности отбора.

Задача проверки соответствия эмпирических распределений гипотетическим часто является предварительным этапом более сложных статистических процедур, напр. исследования взаимосвязи между показателями. Выдвижение гипотезы о соответствии эмпирического распределения анализируемой случайной величины некоторому известному закону, эквивалентно выбору модели порождения данных исследуемого процесса, которая является отправной точкой любого статистического исследования, т.к. определяет методы доступные для анализа. Напр., многие статистические методы анализа исходят из предположения о нормальности распределения исследуемых величин (линейная и логистическая регрессия, байесовская классификация и т.д.), так что для их применения необходимо предварительно обосновать согласованность закона распределения показателей с нормальным.

КРИТЕРИЙ СОГЛАСИЯ КОЛМОГОРОВА

критерий проверки гипотезы о соответствии полученного распределения предполагаемой модели, или гипотезы, что два эмпирических распределения подчиняются одному закону.

Наблюдаемое значение статистики критерия рассчитывается на основе оценки эмпирической функции распределения $F_n(x)$, которая строится по выборке $x = (x_1, x_2, \ldots, x_n)$ и имеет вид:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{x_i \leq x},$$

где $I_{xi \leq x}$ указывает, попало ли наблюдение x_i в область $(-\infty, x]$:

$$I_{x_i \leq x} = \begin{cases} 1, при x_i \leq x \\ 0, при x_i > x \end{cases}.$$

При рассмотрении одной выборки, проверяется гипотеза H_0: $F(x) = F_0(x)$, которая состоит в том, что выборка подчиняется заданному распределению, против альтернативной гипотезы H_1: $F(x) \neq F_0(x)$. Статистика критерия для эмпирической функции распределения $F_n(x)$ определяется по формуле:

$$T_{n\,набл} = \sqrt{n} D_n = \sqrt{n} \cdot \sup_x \left| F_n(x) - F_0(x) \right|,$$

где n – объём выборки, $F_0(x)$ – теоретическая функция распределения, sup S – точная верхняя грань множества S.

В случае, когда рассматриваются две выборки, проверяется гипотеза

H_0: $F_1(x) = F_2(x)$

о том, что данные двух выборок подчиняются одному и тому же закону распределению, против альтернативной гипотезы

H_1: $F_1(x) \neq F_2(x)$.

Статистика критерия для эмпирической функции распределения $F_n(x)$ определяется следующим образом:

$$T_{n\,набл} = \sqrt{\frac{nm}{n+m}} D_{n,m} = \sqrt{\frac{nm}{n+m}} \cdot \sup_x \left| F_{1,n}(x) - F_{2,m}(x) \right|,$$

где n, m – объёмы двух выборок,

$F_{1,n}(x)$, $F_{2,m}(x)$

– эмпирические функции распределения, построенные по независимым выборкам объёмом n и m.

Для проверки гипотезы строят правостороннюю критическую область, границу которой определяют в соответствии с законом распределения Колмогорова:

$$K(\lambda)=\begin{cases}0 & при\ \lambda\le 0\\ \sum_{j=-\infty}^{\infty}(-1)^j\exp(-2j^2\lambda^2) & при\ \lambda>0\end{cases}.$$

Если статистика $T_{n\ набл}$ превышает квантиль распределения Колмогорова K_α заданного уровня значимости α, то нулевая гипотеза H_0 (о соответствии закону $F_0(x)$ или об однородности выборок) отвергается. Иначе гипотеза принимается на уровне α. Если n достаточно велико, то K_α можно приблизительно рассчитать по формуле:

$$K_\alpha\approx\sqrt{-\frac{\ln\alpha/2}{2}}.$$

Можно осуществлять проверку гипотезы с помощью К.с.К. и в другом порядке. По наблюдаемому значению $T_{n\ набл}$ определить, пользуясь законом распределения Колмогорова, вероятность

$\alpha_{набл} = P(T > T_{n\ набл}) = 1 - K(T_{n\ набл})$.

Если $\alpha_{набл} \le \alpha$, то отклонения значимы, т.е. гипотеза отвергается, если же $\alpha_{набл} > \alpha$, то гипотеза не отвергается.

КРИТЕРИЙ СОГЛАСИЯ КОЛМОГОРОВА-СМИРНОВА

см. в ст. Критерий согласия Колмогорова

КРИТЕРИЙ СОСТОЯТЕЛЬНЫЙ

критерий статистический, обладающий асимптотическим свойством: при $n\to\infty$ мощность критерия $W\to 1$. Состоятельность критерия означает, что с ростом числа наблюдений он позволяет с вероятностью, близкой к 1, «улавливать» любые отклонения от осн. гипотезы. В частности, К.с. является асимптотически несмещённым.

КРИТЕРИЙ СТАТИСТИЧЕСКИЙ

правило принятия решения на основе выборочных наблюдений в задаче *статистической проверки гипотез*, т.е. процедура обоснованного сопоставления проверяемой гипотезы с имеющимися в распоряжении исследователя выборочными данными.

Исходя из вида проверяемой гипотезы H_0, составляется специальная функция (статистика) от выборочных данных, которая при выполнении проверяемой гипотезы подчиняется известному закону распределения. Затем выбирается конкурирующая (альтернативная) гипотеза H_1 и уровень значимости α, на основе которых весь диапазон возможных значений статистики разбивается на два непересекающихся множества: область допустимых значений и критическую область (область малоправдоподобных значений статистики в условиях справедливости проверяемой гипотезы).

При решении практических задач границы критической области выбираются из соответствующей табл. закона распределения, которому подчиняется статистика. Если наблюдаемое значение статистики (вычисленное по выборочным данным) попало в критическую область, то делается вывод, что проверяемая гипотеза H_0 – не согласуется с результатами наблюдений имеющейся выборки и отвергается (признается ошибочной). Поэтому более правдоподобной считается конкурирующая гипотеза H_1. Принимая такое решение, исследователь может совершить ошибку первого рода (отвергнуть верную, в действительности, гипотезу) с вероятностью α.

Если наблюдаемое значение статистики не попало в критическую область, то делается вывод, что проверяемая гипотеза H_0 подтверждается эмпирическими данными имеющейся выборки и не отвергается (т.е. признается справедливой). Тогда ошибочной признается альтернативная гипотеза H_1. Принимая такое решение, исследователь может с вероятностью β совершить ошибку второго рода (принять неверную, в действительности, гипотезу H_0, когда справедлива H_1). Вероятность $\mathbf{1}\text{-}\beta$ принятия

гипотезы H_1, когда она, в действительности, верна, называется мощностью критерия.

Т.о., применение процедуры проверки гипотез сопряжено с возможностью совершения ошибок двух родов – отвергнуть гипотезу, когда она верна (ошибка первого рода), или принять гипотезу, когда она неверна (ошибка второго рода). При проверке гипотез желательно добиваться минимизации значений ошибок обоих родов. Но в большинстве задач невозможно одновременно минимизировать обе ошибки. Стремление минимизировать одну из них приводит к возрастанию другой. При многократном повторении процедуры принятия решений доля неверно принимаемых решений характеризуется значениями α и β. Если вероятность совершения определенной ошибки становится очень малой, можно говорить о практической невозможности совершить эту ошибку.

Если эмпирические данные согласуются с предполагаемой гипотезой, это не исключает возможности согласования тех же данных с другой гипотезой. Принятие гипотезы не означает, что она является единственно верной (или даже самой правильной). Поэтому принятие гипотезы – не более, чем достаточно правдоподобное, не противоречащее опыту, предположение. Строго доказать справедливость проверяемой гипотезы – невозможно. Допустимо лишь утверждать, что гипотеза не противоречит опытным данным.

Наиболее часто решаемые задачи: проверка гипотез о виде закона распределения или о значениях его параметров, об однородности выборок, о свойствах параметров связи, о наличии аномальных наблюдений и т.д. Значительная часть одномерных задач распространяется и на многомерный случай.

КРИТЕРИЙ СТАТИСТИЧЕСКИЙ НАИБОЛЕЕ МОЩНЫЙ

критерий, используемый для проверки простых *гипотез статистических*, имеющий наибольшую мощность среди всех других критериев. Если обозначить уровень значимости критерия α (вероятность *ошибки первого рода*), через β – вероятность ошибки второго рода, тогда мощность критерия будет определяться формулой

$$W = 1 - \beta .$$

В конкретных задачах для фиксированного значения ошибки первого рода α может быть предложен ряд критериев (правил выбора решения), каждому из которых будет соответствовать свое значение *мощности критерия*. К.с.н.м. будет определяться соотношением:

$$W = 1 - \beta \to \max .$$

Напр., критерий Неймана-Пирсона – *критерий отношения правдоподобия*, обеспечивает макс. мощность решения в пользу *нулевой гипотезы*, а потому в классе простых гипотез является наиболее мощным. Вычисление функции мощности критерия – достаточно сложная задача, т.к. для этого требуется знать распределение статистики критерия не только при нулевой гипотезе, но и при альтернативах.

КРИТЕРИЙ СТЬЮДЕНТА (Т-КРИТЕРИЙ)

критерий статистический, основанный на применении *распределения Стьюдента* и используемый для проверки гипотез о средних значениях нормальных распределений; впервые применён в 1908 У. Госсетом, известным под псевдонимом Стьюдент. Пусть результаты наблюдений X_1, ..., X_n – взаимно независимые нормально распределённые случайные величины с неизвестными параметрами a и σ^2. Для проверки гипотезы H_0: $a=a_0$ в соответствии с К.с. используется статистика:

$$t_{выб.} = \frac{\overline{X} - a_0}{s}\sqrt{n-1}$$

где $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k, \quad s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2.$

При условии, что гипотеза $a=a_0$ справедлива, статистика $t_{выб}$ имеет распределение Стьюдента *с* $\nu = n-1$ степенями свободы. В зависимости от формулировки альтернативной гипотезы H_1 критическая область может быть двусторонней (при $H_1 : a \neq a_0$) или односторонней (при $H_1 : a \leq a_0$ или $H_1 : a \geq a_0$). При *альтернативной гипотезе* $H_1 : a \neq a_0$ исходная гипотеза

H_0: $a=a_0$ принимается при заданном уровне значимости α, если $|t_{выб}| \le t_{кр}(\alpha,\nu)$, где $t_{кр}(\alpha,\nu)$ находится из соотношения:

$$\int_{-t_{кр}}^{t_{кр}} S_\nu(t)\ dt = 1-\alpha,$$

по плотности $S_\nu(t)$ распределения Стьюдента (или по специальным табл.); в противном случае исходная гипотеза отклоняется в пользу альтернативной: $a \ne a_0$.

Если заранее известно, что $a \ge a_0$, то гипотеза H_0 будет отклоняться в пользу гипотезы $a > a_0$ при $|t_{выб}| > t_{кр}(\alpha,\nu)$,
где $t_{кр}(\alpha,\nu)$ находится по формуле:

$$\int_{0}^{t_{кр}} S_\nu(t)\ dt = 1-\alpha,$$

и приниматься в противоположном случае. При этом К.с. будет равномерно наиболее мощным критерием уровня α среди всех критериев проверки гипотезы $a=a_0$ относительно альтернативных гипотез $a > a_0$.

К.с. используется также для критерия однородности средних значений в двух нормальных ген. совокупностях, имеющих одинаковую, но неизвестную дисперсию σ^2. Рассмотрим выборки $X=(X_1,\ldots,X_n)$ и $Y=(Y_1,\ldots,Y_m)$, причём X и Y – взаимно независимы, величины X_i имеют нормальное распределение с параметрами a_x и σ^2, а Y_j – нормальное распределение с параметрами a_y и σ^2, где дисперсия σ^2 неизвестна. Лучшая оценка для σ^2 :

$$s^2 = \frac{ns_x^2 + ms_y^2}{n+m-2},$$

где $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$, $s_y^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i-\overline{Y})^2$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$, $\overline{Y} = \frac{1}{m}\sum_{j=1}^{m}Y_j$

– *выборочные дисперсии* и выборочные средние, вычисленные по выборкам X и Y. Гипотеза однородности формулируется как гипотеза равенства средних значений $a_x = a_y$. Для проверки данной гипотезы используется статистика:

$$t_{выб} = \frac{\overline{X}-\overline{Y}}{s\sqrt{\frac{1}{n}+\frac{1}{m}}},$$

которая имеет распределение Стьюдента с $\nu = n+m-2$ степенями свободы в предположении, что гипотеза $a_x = a_y$ верна. Соответствующий критерий строится стандартным способом.

К.с. используется для оценки значимости коэффициентов корреляции (в т.ч. и частных коэффициентов корреляции; для оценки значимости коэффициентов уравнения регрессии.

См. также *Корреляционный анализ.*

КРИТЕРИЙ ФИШЕРА (F-КРИТЕРИЙ)

критерий статистический, предназначенный для проверки *нулевой гипотезы* об однородности дисперсий в двух нормальных ген. совокупностях

$N(\mu_1,\sigma_1)$ и $N(\mu_2,\sigma_2)$,

т.е. H_0: $\sigma_1^2 = \sigma_2^2$

по двум имеющимся из них независимым случайным выборкам

$x_{11}, x_{12}, \ldots, x_{1n_1}$ и $x_{21}, x_{22}, \ldots, x_{2n_2}$

объёмом, соответственно, n_1 и n_2 наблюдений. Имя критерию дал Снедекор в честь Рональда Фишера, выдающегося англ. статистика, впервые в 1920 предложившего статистику критерия как отношение дисперсий.

Критическая статистика К.Ф. имеет вид (т.н. дисперсионное отношение):

$$F_{набл} = \frac{\hat{S}_1^2}{\hat{S}_2^2}; \qquad \hat{S}_1^2 > \hat{S}_2^2,$$

где $\hat{S}_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 = \frac{n_i}{n_i-1}S_i^2$; $i=1,2$;

$\hat{S}_i^2$ и S_i^2 – соответственно, несмещённые (исправленные) и смещённые выборочные дисперсии.

Статистика обычно строится т.о., чтобы в числителе стояла бóльшая выборочная дисперсия (это обусловлено спецификой построения табл. процентных точек F-распределения, и, если это не так, следует просто изменить нумерацию выборок и в числитель поставить выборку с бóльшей исправленной дисперсией). Этим обусловлена правосторонняя *критическая область* критерия и *альтернативная* (конкурирующая) *гипотеза* H_1:$\sigma_1^2>\sigma_2^2$.

Критическая статистика, при условии выполнения нулевой гипотезы, имеет *F*- распределение Фишера-Снедекора с числом степеней свободы числителя $\nu_1=n_1-1$ и числом степеней свободы знаменателя $\nu_2=n_2-1$. Критическое значение статистики критерия находится для заданного уровня значимости α по табл. распределения Фишера-Снедекора:

$F_{кр}(\alpha; \nu_1=n_1-1; \nu_2=n_2-1)$.

Нулевая гипотеза отвергается с вероятностью ошибки α, если $F_{набл}> F_{кр}$, и принимается в противном случае.

Если нулевая гипотеза не отвергается, то различие выборочных исправленных дисперсий не является статистически значимым, его можно объяснить влиянием каких-либо случайных факторов (вероятность справедливости таких утверждений – *мощность критерия* $1-\beta$). Если же нулевая гипотеза отвергается, это говорит о статистически значимом различии между генеральными дисперсиями, т.е. о существенном отличии вариабельности изучаемого признака в двух рассматриваемых генеральных совокупностях (с вероятностью ошибки в такого рода утверждениях α).

Часто К.Ф применяется как первый шаг при проверке однородности двух независимых нормальных совокупностей $N(\mu_1,\sigma_1)$ и $N(\mu_2,\sigma_2)$: для проверки гипотезы о равенстве ген. средних H_0: $\mu_1=\mu_2$ с помощью *критерия Стьюдента* в случае неизвестных ген. дисперсий требуется проверка равенства ген. дисперсий $\sigma_1^2=\sigma_2^2$ изучаемых нормальных случайных величин – условия применимости критерия Стьюдента. Это осуществляется с помощью К.Ф. и, если он не отвергает нулевую гипотезу о равенстве ген. дисперсий, можно применять критерий Стьюдента и проверять равенство ген. средних.

Реже в К.Ф. проверяется более сложный вариант нулевой гипотезы:

H_0: $\sigma_1^2/\sigma_2^2=k$ (k –заданное число).

Тогда в качестве критического значения выбирают $k\cdot F_{кр}(\alpha; \nu_1=n_1-1; \nu_2=n_2-1)$. Нулевая гипотеза отвергается с вероятностью ошибки α, если

$F_{набл}> k\cdot F_{кр}$

и принимается в противном случае.

Встречается более общее понимание К.Ф. как любого статистического критерия, критическая статистика которого имеет распределение Фишера-Снедекора в случае выполнения нулевой гипотезы. Наибольшее распространение такие критерии имеют в дисперсионном, корреляционном и регрессионном анализе, где статистики критериев представляют собой отношения сумм квадратов отклонений.

КРИТИЧЕСКАЯ ОБЛАСТЬ

область «малоправдоподобных», в условиях справедливости проверяемой *нулевой гипотезы* H_0 (при *альтернативной гипотезе* H_1), значений *критической статистики* $\Theta_n^*(x_1, x_2, \ldots, x_n)$. При попадании значений статистики критерия в критическую область проверяемая гипотеза H_0 отвергается с вероятностью ошибки α.

Пользуясь табл. известного распределения статистики критерия, находят для заданного уровня значимости критерия α критические значения, которые разбивают все множество возможных значений статистики $\theta_n^*(x_1, x_2, \ldots, x_n)$ на два непересекающихся подмножества (области):

К.о. W (область отклонения гипотезы) и ей противоположная область принятия гипотезы.

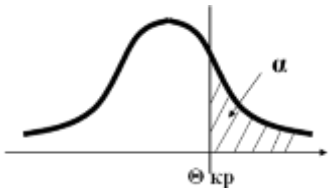
К.о., в зависимости от вида альтернативной (конкурирующей) гипотезы H_1, может быть трёх типов: правосторонней, левосторонней и двусторонней.

Границы К.о. при заданном уровне значимости α находят из табл. распределения критической статистики с помощью квантилей уровня $(1-\alpha)$,

α, α/2, (1- α/2), исходя из условий и соотношений:

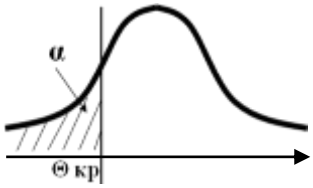
- для правосторонней К.о.

  $P(\theta_n^* > \theta_{кр}) = \alpha$ – если осн. опасность для проверяемой нулевой гипотезы H_0 представляют «слишком большие» значения критической статистики, т.е. альтернативная гипотеза H_1 провозглашает большие значения, чем нулевая;



- для левосторонней К.о.

  $P(\theta_n^* < \theta_{кр}) = \alpha$ – если проверяется нулевая гипотеза H_0 против «слишком малых» значений статистики критерия в альтернативной гипотезе H_1;

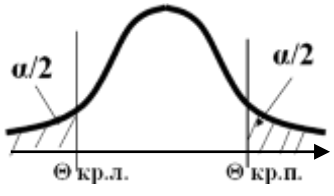


- для двусторонней симметрической области

  $P(\theta_n^* > \theta_{кр.пр}) = \alpha / 2$;

  $P(\theta_n^* < \theta_{кр.лев}) = \alpha / 2$; $\theta_{кр.лев} < \theta_{кр.пр}$.

  – если нулевая гипотеза проверяется как против «неправдоподобно больших», так и против «неправдоподобно малых» значений статистики критерия.



Если посчитанные по выборке наблюденные значения статистики критерия попадают в К.о., то нулевую гипотезу H_0 отвергают с вероятностью *ошибки первого рода* α в пользу альтернативной (конкурирующей) гипотезы H_1. В противном случае гипотеза H_0 не противоречит результатам наблюдений.

К.о. выбирается так, чтобы вероятность попадания в неё была бы миним. (равной α), если верна нулевая гипотеза H_0, и макс. в противоположном случае:

$$\begin{cases} P(\theta_n^* \in W | H_0) = \alpha \Rightarrow \min \\ P(\theta_n^* \in W | H_1) = 1 - \beta \Rightarrow \max \end{cases},$$

где $P(\theta_n^* \in W|H_i)$ – вероятность попадания статистики критерия θ_n^* в К.о. W, если верна гипотеза H_i; второе условие выражает требование максимума мощности критерия 1-β – вероятности правильного отклонения нулевой гипотезы H_0, когда она не верна, т.е. вероятности не совершить ошибку второго рода β.

КРИТИЧЕСКАЯ СТАТИСТИКА (СТАТИСТИКА КРИТЕРИЯ)

функция от результатов выборочных наблюдений $\Theta_n^*(x_1, x_2, \ldots, x_n)$, используемая при *статистической проверке гипотез.* На основании численного значения К.с. принимается решение об отклонении или принятии проверяемой *нулевой гипотезы* H_0. К.с. – основа любого *критерия статистического*.

К.с., как и всякая функция от результатов наблюдений, и сама является случайной величиной, поэтому, в предположении справедливости проверяемой нулевой гипотезы, точно или асимптотически подчинена некоторому, как правило, хорошо изученному и затабулированному закону распределения с плотностью $f_{\Theta n}$ (напр., нормальному закону распределения или закону распределения Стьюдента и т.п.).

Один из осн. принципов построения К.с.– принцип отношения правдоподобия. Общий содержательный смысл К.с., построенной по этому принципу, состоит в том, что ею определяется, как правило, мера расхождения имеющихся в нашем распоряжении выборочных данных с проверяемой нулевой гипотезой H_0. Так, в *гипотезах статистических* о типе закона распределения К.с. определяет меру различия между анализируемой эмпирической функцией распределения $\hat{F}_{э(n)}$ и гипотетической теоретической функцией F_T. В гипотезах об однородности нескольких выборок величина К.с. измеряет степень расхождения соответ-

ствующих выборочных характеристик в различных выборках. В гипотезах о числовых значениях параметров К.с. Θ_n^* определяет существенность отклонений выборочных характеристик от соответствующих гипотетических значений и т.д.

С помощью известного закона распределения К.с. по соответствующим табл. распределения находят для заданного уровня значимости критерия α критические значения, которые разбивают все множество возможных значений статистики $\theta_n^*(x_1, x_2, \ldots, x_n)$ на два непересекающихся подмножества (области): *критическая область* (область отклонения гипотезы) и ей противоположная область принятия гипотезы. Если посчитанные по выборке наблюденные значения статистики критерия попадают в критическую область, то нулевую гипотезу H_0 отвергают в пользу *альтернативной* (конкурирующей) *гипотезы* H_1 с вероятностью ошибки α. В противном случае гипотезу H_0 не отвергают.

Л

ЛОГИТ-МОДЕЛЬ БИНАРНОГО ВЫБОРА

модель бинарного выбора вида

$$\mathrm{P}\{y_t = 1|x_t\} = F(x_t^T\beta), \quad t = 1,2,\ldots,n,$$
$$y_t \epsilon \{0,1\}, \qquad x_t^T = (1, x_{t1}, x_{t2}, \ldots, x_{tp}), \qquad \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T,$$

в которой $F(\cdot)$ – *логистическая функция*:

$$F(\omega) = \Lambda(\omega) = \frac{e^{\omega}}{1 + e^{\omega}}.$$

Эта функция удовлетворяет условиям:

$$\left.\begin{array}{l} \Lambda(\omega) \text{ монотонно возрастает по } \omega, \\ 0 \le \Lambda(\omega) \le 1, \\ \Lambda(\omega) \longrightarrow 1 \text{ при } \omega \longrightarrow +\infty, \\ \Lambda(\omega) \longrightarrow 0 \text{ при } \omega \longrightarrow +\infty, \\ \Lambda(-\omega) = 1 - \Lambda(\omega). \end{array}\right\}$$

Предельный эффект каждого фактора X_k $(k = 1,2,\ldots,p)$ удовлетворяет соотношению:

$$\frac{\partial P\{y_t = 1|x_t\}}{\partial x_t} = \frac{\partial \Lambda(x_t^T\beta)}{\partial x_t}\,\beta = \Lambda(x_t^T\beta)\big(1 - \Lambda(x_t^T\beta)\big)\beta,$$

т.е. зависит от значений всех факторов, а направление его влияния на объясняемую величину определяется знаком соответствующего ему параметра β_k.

Удобно интерпретировать Л.-м.б.в. с использованием т.н. логита – логарифма шанса искомого события, поскольку она относительно него линейна:

$$logit(p_t) = ln\frac{p_t\{y_t = 1|x_t\}}{1 - p_t\{y_t = 1|x_t\}} = x_t^T\beta.$$

При изменении k-й объясняющей переменной на малую величину Δx_{kt} приводит (при неизменных значениях остальных объясняющих переменных) к изменению шанса события $\{y_t = 1\}$ приблизительно на

$x_t^T\beta \cdot 100\%$.

Оценивание параметров Л.-м.б.в. чаще всего производят *методом макс. правдоподобия*, причём условия первого порядка

$$\frac{\partial logL(\beta)}{\partial \beta} = \sum_{t=1}^{n}\left[y_t - \frac{e^{x_t^T\beta}}{1 + e^{x_t^T\beta}}\right]x_t = 0, \quad \text{где} \quad L(\beta) = \prod_{t=1}^{n}(\mathrm{P}\{y_t = 1|x_t\})^{y_t}\,(\mathrm{P}\{y_t = 0|x_t\})^{1-y_t},$$

означают (при наличии постоянного фактора X_0), что сумма оцениваемых вероятностей равна числу наблюдений в выборке, для которых $y_t = 1$. Анализ условий второго порядка метод максимального правдоподобия (ММП) показывает, что матрица производных второго порядка является отрицательно определённой, что обуславливает вогнутость функции $L(\beta)$ и быструю сходимость алгоритма ММП. В другом случае, когда данные, сгруппированы по значениям объясняющей переменной, Л.-м.б.в. преобразуют к виду линейной модели с *гетероскедастичными остатками*. Затем переходят от переменной y_t к переменной $logit(p_t)$ и используют взвешенный *метод наименьших квадратов*, решая задачу оптимизации:

$$\sum_{t=1}^{k} w_t[logit(p_t) - x_t^T\beta] \longrightarrow \min_{\beta},$$

$w_t = n_t\,\Lambda(x_t^T\beta)\big(1-\Lambda(x_t^T\beta)\big)$, n_t – число значений y_t в *t*-й группе, $\sum_{t=1}^{k} n_t$

ЛОГАРИФМИЧЕСКАЯ ФУНКЦИЯ ПРАВДОПОДОБИЯ

натуральный логарифм функции правдоподобия: $l(\theta) = \ln L(\theta)$. Поскольку логарифм – строго возрастающая функция, максимум Л.ф.п. достигается при том же значении параметра θ, что и максимум функции правдоподобия, а в силу специфики конкретных функций правдоподобия, часто представляющих собой произведения степенных и показательных функций, в *методе макс. правдоподобия* бывает удобнее вместо максимизации функции правдоподобия решать эквивалентную задачу максимизации её логарифма.

М

МАССИВ ИСХОДНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

см. в ст. Исходные статистические данные

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

раздел математики, объединяющий систему понятий, приёмов и математических методов, основанных на теоретико-вероятностных моделях, предназначенных для сбора, систематизации, обработки и интерпретации статистических данных с целью получения научных и практических выводов. В отличие от прикладной статистики, М.с. имеет дело лишь со статистическими данными, носящими случайный характер, т.е. с результатами наблюдений или измерений т.н. случайных величин, подчиняющихся некоторому закону распределения вероятностей.

Для эффективной эксплуатации математико-статистического способа рассуждения и соответствующего математического аппарата необходимо стационарное (не изменяющееся во времени) действие некоторого реального комплекса условий. Этот комплекс должен включать неизбежность «мешающего» влияния большого числа случайных (не поддающихся строгому учёту и контролю) факторов, которые, в свою очередь, не позволяют делать полностью достоверные выводы о том, произойдёт или не произойдёт определённое событие. При этом предполагается принципиальная возможность многократного повторения эксперимента или наблюдения в рамках того же самого реального комплекса условий. Такую ситуацию принято называть условиями действия статистического ансамбля или условиями соблюдения устойчивости однородности исследуемой совокупности.

Пример реальных ситуаций, подчиняющихся требованию статистической устойчивости (или укладывающихся в рамки статистического ансамбля) – область азартных игр. Подбрасывая монету, или бросая игральную кость и интересуясь вероятностью осуществления события, заключающегося соответственно в появлении «герба» или «шестёрки», полагаем: а) можно многократно повторить тот же самый эксперимент в тех же самых условиях; б) наличие большого числа случайных факторов, характеризующих условия проведения каждого такого эксперимента, не позволяет делать полностью определённого (детерминированного) заключения о том, произойдёт в результате данного эксперимента интересующее нас событие или не произойдёт; в) чем большее число однотип-

ных экспериментов мы произведём, тем ближе будут подсчитанные по результатам экспериментов относительные частоты появления интересующих нас событий к некоторым постоянным величинам, называемым вероятностями этих событий, а именно: относительная частота появления «герба» при подбрасывании симметричной монеты будет приближаться к 1/2, выпадения «шестёрки» при бросании симметричной кости – к 1/6.

Требования статистического ансамбля применительно к указанным двум типам экспериментов означают необходимость использования одной и той же (или совершенно идентичных) симметричной монеты, симметричной кости. Соблюдение условий статистического ансамбля в более серьёзных и сложных сферах человеческой деятельности – в экономике, социальных процессах, технике и пром-ти, медицине, в различных отраслях науки – вопрос, требующий специального рассмотрения в каждом конкретном случае. Оценивая специфику задач в различных областях человеческого знания с позиции соблюдения в них свойств а) – в) статистической устойчивости и принимая во внимание накопившийся опыт вероятностно-статистических приложений, можно условно разбить все возможные приложения на три категории.

К первой категории областей применения – категории высокой работоспособности математико-статистических методов – относятся ситуации, в которых свойства а) – в) статистической устойчивости исследуемой совокупности бесспорно имеют место либо нарушаются столь незначительно, что это практически не влияет на точность статистических выводов, полученных с использованием теоретико-вероятностных моделей. Сюда помимо «игровой» тематики могут быть отнесены отдельные разделы экономики и социологии и в первую очередь задачи, связанные с исследованием поведения объекта (индивида, семьи или другой социально-экономической или производственной единицы) как представителя большой однородной совокупности подобных же объектов.

Ко второй – категории допустимых вероятностно-статистических приложений – относятся ситуации, характеризующиеся нарушениями требования сохранения неизменными условий эксперимента (вторая половина требования а)) и вытекающими отсюда отклонениями от требования в). Характерная форма такого рода отклонений от условий статистического ансамбля – объединение в одном ряду наблюдений (подлежащих обработке) различных порций исходных данных, зарегистрированных в разных условиях (в разное время или в разных совокупностях). В эту же категорию входит определённый класс задач, связанных с анализом коротких временных рядов, зарегистрированных в условиях, практически исключающих возможность статистической фиксации сразу нескольких эмпирических реализаций исследуемого временного ряда на одном и том же временном интервале. Использование вероятностно-статистических методов обработки в этом случае допустимо, но должно сопровождаться пояснениями о несовершенстве и приближённом характере получаемых выводов и по возможности должно дополняться другими методами научного анализа.

К третьей категории задач статистической обработки исходных данных – категории недопустимых вероятностно-статистических приложений относятся ситуации, характеризующиеся либо принципиальным неприятием главной идеи понятия статистического ансамбля – массовости исследуемой совокупности, либо полной детерминированностью изучаемого явления, т.е. отсутствием «мешающего» влияния множества случайных факторов (нарушение требования б)). В подобных ситуациях исследователь должен пользоваться методами прикладной статистики или анализа данных и не должен претендовать на вероятностную интерпретацию обрабатываемых данных и получаемых в результате их обработки выводов.

Наиболее важные и распространённые классы задач М.с. связаны с анализом и интерпретацией данных $X=\{X_1, X_2, \ldots, X_n\}$, представляющих выборку объёма n из семейства распределений $F=\{\mathcal{P}_\theta, \theta \in \Theta\}$ где θ – некоторый (вообще говоря, векторный) параметр, значе-

ние которого неизвестно исследователю, а Θ – множество его возможных значений. Это значит, что X представляет собой последовательность независимых одинаково распределённых по какому-то из законов класса F случайных величин (наблюдений). Выборка реализует представление об *n* независимых повторениях в неизменных условиях эксперимента, связанного с измерением (наблюдением) некоторой величины. В совр. М.с. достаточно часто встречаются последовательности наблюдений $X_1,\ldots,X_n$, не являющихся независимыми, но представляющими реализации какого-либо случайного процесса: марковского, стационарного (в этом случае говорят о статистике временных рядов) и пр.

Задачи М.с. состоят в том, чтобы в рамках данной математической модели по реализации случайного элемента *X* (наблюдению) сделать то или иное заключение о параметре θ. Вероятностная природа и ограниченный объём статистических данных не позволяют делать абсолютно точные утверждения о значении θ. Цель статистической теории – выработка наиболее эффективных методов обработки данных, дающих «наиболее точные» утверждения с «наибольшей или заданной степенью достоверности».

Теоретический базис М.с, на котором основаны методы и приёмы решения её разнообразных типовых задач (регрессии, классификации, дисперсионного анализа, анализа временных рядов и т. п.), состоит из двух разделов: теории статистического оценивания и теории статистической проверки гипотез.

Теория статистического оценивания (неизвестных значений параметров или функций) разрабатывает математические приёмы и методы, с помощью которых на основании исходных статистических данных можно вычислить как можно более точные приближённые значения (статистические оценки) для одного или нескольких числовых параметров или функций, характеризующих функционирование исследуемой системы.

Принципиальная возможность получения работоспособных приближений такого рода на основании статистического обследования лишь части анализируемой ген. совокупности (т.е. на основании ограниченного ряда наблюдений, или выборки) обеспечивается замечательным свойством статистической устойчивости выборочных характеристик.

Статистическая оценка строится в виде функции от результатов наблюдений, а потому сама – величина случайная. При повторении выборки из той же самой ген. совокупности и при подстановке новых выборочных значений в ту же самую «функцию-оценку» получается другое число в качестве приближённого значения интересующего нас параметра, т.е. имеется неконтролируемый разброс в значениях оценки при повторениях эксперимента.

В качестве осн. меры точности статистической оценки неизвестного параметра θ чаще всего используется средний квадрат её отклонения от оцениваемого значения $E(\hat{\theta}-\theta)^2$, а в многомерном случае – ковариационная матрица компонент векторной оценки $\hat{\theta}$. Чем меньше эта величина (или обобщённая дисперсия оценки $\hat{\theta}$ в многомерном случае), тем точнее (эффективнее) оценка. Для широкого класса ген. совокупностей существует неравенство Рао-Крамера-Фреше, задающее тот минимум $\Delta^2_{\min}$ (по всем возможным оценкам) среднего квадрата $E(\hat{\theta}-\theta)^2$, улучшить который невозможно. Естественно использовать этот минимум $\Delta^2_{\min}$ в качестве начальной точки отсчёта меры эффективности оценки, определив эффективность $e(\hat{\theta})$ любой оценки $\hat{\theta}$ параметра θ в виде отношения:

$$e(\hat{\theta}) = \frac{\Delta^2_{\min}}{E(\hat{\theta}-\theta)^2}.$$

Свойство состоятельности оценки $\hat{\theta}$ обеспечивает её статистическую устойчивость, т.е. её сходимость (по вероятности) к истинному значению оцениваемого параметра θ по мере роста объёма выборки, на основании которой эта оценка строится. Свойство несмещённости оценки $\hat{\theta}$ заключается в том, что результат усреднения всевозможных значений этой оценки, полученных по различным выборкам заданного объёма (из одной и той же генеральной совокупности), даёт в точности истинное

значение оцениваемого параметра, т.е. $E\hat{\theta} = \theta$. Далеко не всегда следует настаивать на необходимом соблюдении свойства несмещённости оценки: несущественное само по себе уже при умеренно больших объёмах выборки, оно может чрезмерно обеднить класс оценок, в рамках которого решается задача построения наилучшей оценки.

С учётом случайной природы каждого конкретного оценочного значения $\hat{\theta}$ неизвестного параметра θ представляет интерес построение целых интервалов оценочных значений $\Delta_P\hat{\theta}$, а в многомерном случае – целых областей, которые с наперёд заданной (и близкой к единице) вероятностью P накрывали бы истинное значение оцениваемого параметра θ, т.е.

$\mathcal{P}\{\theta \in \Delta_P\ \hat{\theta}\} = P$.

Эти интервалы (области) принято называть доверительными или интервальными оценками. Существует два подхода к построению интервальных оценок: точный (конструктивно реализуемый лишь в сравнительно узком классе ситуаций) и асимптотически-приближённый (наиболее распространённый в практике статистических приложений). Осн. методы построения статистических оценок: метод макс. правдоподобия; метод моментов; обобщённый метод моментов; метод наименьших квадратов; метод, использующий «взвешивание» наблюдений: цензурирование, урезание, порядковые статистики. Различные варианты метода, использующего «взвешивание» наблюдений, находят всё большее распространение в связи с устойчивостью получаемых при этом статистических выводов по отношению к возможным отклонениям реального распределения исследуемой ген. совокупности от постулируемого модельного.

Наличие априорной информации об оцениваемом параметре, позволяющей сопоставить с каждым возможным значением неизвестного параметра некую вероятностную меру его достоверности, позволяет существенно уточнить оценки, полученные традиционными методами (методом максимального правдоподобия, методом моментов и т.п.) в условиях отсутствия такой информации. Построение таких оценок осуществляется с помощью т.н. байесовского подхода, а сами оценки называются байесовскими.

Теория статистической проверки гипотез разрабатывает процедуры обоснованного сопоставления высказанной гипотезы (о значениях параметров или о природе анализируемой модели) с имеющимися исходными статистическими данными $X_1, X_2, \ldots, X_n$. Такие процедуры называют статистическими критериями).

Результат подобного сопоставления может быть либо отрицательным (данные наблюдения противоречат высказанной гипотезе, а потому от этой гипотезы следует отказаться), либо неотрицательным (данные наблюдения не противоречат высказанной гипотезе, а потому её можно принять в качестве одного из естественных и допустимых решений). При этом неотрицательный результат статистической проверки гипотезы H не означает, что высказанное предположительное утверждение наилучшее: просто оно не противоречит имеющимся выборочным данным, однако таким же свойством могут наряду с H обладать и другие гипотезы.

По своему прикладному содержанию гипотезы, высказываемые в ходе статистической обработки данных, можно подразделить на несколько осн. типов.

1. Гипотезы о типе закона распределения исследуемой случайной величины. При обработке ряда наблюдений $X_1, X_2, \ldots, X_n$ исследуемой случайной величины ξ важно понять механизм формирования выборочных значений X_i, т.е. подобрать и обосновать некоторую модельную функцию распределения $F_{\text{mod}}(X)$, с помощью которой можно адекватно описать исследуемую функцию распределения $F_\xi(X)$. На определённой стадии исследования это приводит к необходимости проверки гипотез типа $H : F_\xi(X) \equiv F_{\text{mod}}(X)$, где гипотетическая модельная функция может быть как заданной однозначно (тогда

$F_{\text{mod}}(X) = F_0(X),$,

где $F_0(X)$ – полностью известная функция), так и заданной с точностью до принадлежности к некоторому параметрическому семейству, тогда

$F_{\text{mod}}(X) = F(X;\theta)$,

где θ – некоторый *k*-мерный параметр, значения которого неизвестны, но могут быть оценены по выборке *X* с помощью методов статистического оценивания.

Проверка гипотез этого типа осуществляется с помощью т.н. критериев согласия и опирается на ту или иную меру различия между анализируемой эмпирической функцией распределения $F_{\xi}^{(n)}(X)$ и гипотетическим модельным законом $F_{\text{mod}}(X)$.

2. Гипотезы об однородности двух или нескольких обрабатываемых выборок или некоторых характеристик анализируемых совокупностей. Наиболее типичные задачи такого рода характеризуются следующей общей ситуацией: пусть имеется несколько «порций» выборочных данных:

1-я: $X_{11}, X_{12} \ldots, X_{1n_1}$;
2-я : $X_{21}, X_{22} \ldots, X_{2n_2}$;
……
l-я: $X_{l1}, X_{l2} \ldots, X_{ln_l}$.

Обозначая функцию распределения, описывающую вероятностный закон, которому подчиняются наблюдения *j*-й выборки, с помощью $F_j(X)$, и снабжая тем же индексом все интересующие нас эмпирические и теоретические характеристики этого закона (средние значения $\hat{a}_j(n_j)$ и a_j; дисперсии $\hat{\sigma}_j^2(n_j)$ и σ_j^2 и т.д.), осн. гипотезы однородности можно записать в виде:

$H_a: a_1 = a_2 = \ldots = a_l$;

В случае неотрицательного результата проверки этих гипотез говорят, что соответствующие выборочные характеристики (напр., $\hat{a}_1(n_1), \ldots, \hat{a}_l(n_l)$) различаются статистически незначимо. Частный случай гипотез такого типа: число выборок *l*= 2, а одна из выборок содержит малое количество наблюдений (в частном случае – одно). Тогда проверка гипотезы означает проверку аномальности одного или нескольких резко выделяющихся наблюдений.

3. Гипотезы о числовых значениях параметров исследуемой ген. совокупности. Пусть, напр., ряд наблюдений даёт нам значения некоторого параметра изделий, измеренные на *n* изделиях, случайно отобранных из массовой продукции определённого станка автоматической линии, и пусть a_0 – заданное номинальное значение этого параметра. Каждое отдельное значение X_i может как-то отклоняться от заданного номинала. Чтобы проверить правильность настройки станка, надо убедиться в том, что среднее значение параметра у производимых на нём изделий будет соответствовать номиналу, т.е. проверить гипотезу типа

$H: E\xi = a_0$.

К гипотезе аналогичного типа приводит попытка проверить статистическую незначимость отличия от нуля выборочного коэффициента корреляции $\hat{r}(x^{(1)}, x^{(2)})$, построенного по совокупности двумерных наблюдений $X_i = (x_i^{(1)}, x_i^{(2)}),\ i = 1, 2, \ldots, n,$ что применительно к соответствующей теоретической характеристике может быть записано как предположительное утверждение:

$H: r(x^{(1)}, x^{(2)}) = 0.$

В общем случае гипотезы подобного типа имеют вид:

$H_0: \theta = \Delta_0,$

где θ – некоторый параметр, от которого зависит исследуемое распределение, а Δ_0 – область его конкретных гипотетических значений, которая может состоять всего из одной точки. К этому типу сводятся, напр., гипотезы независимости и стационарности обрабатываемого ряда наблюдений, гипотеза симметричности анализируемого распределения вероятностей и др.

4. Гипотезы о типе зависимости между компонентами исследуемого многомерного признака. Подобно тому, как при исследовании закона распределения обрабатываемых наблюдений бывает важно правильно подобрать соответствующий модельный закон, так при исследовании статистической зависимости, например, компоненты $x^{(2)}$ от компоненты $x^{(1)}$ анализируемого двумерного признака $X = (x^{(1)}, x^{(2)})$ бывает важно проверить гипотезу об общем виде этой зависимости. Напр., гипотезу о том, что $x^{(2)}$ и $x^{(1)}$ связаны линейной регрессионной связью, т.е.:

$H: E(x^{(2)} \mid x^{(1)} = x) = b_0 + b_1 \cdot x,$

где b_0 и b_1 – некоторые неизвестные параметры модели.

Статистические критерии, с помощью которых проверяются гипотезы различных типов по своему назначению и характеру решаемых задач чрезвычайно разнообразны, однако их объединяет общность логической схемы, по которой они строятся. Коротко эту логическую схему можно описать: а) выдвигается гипотеза; б) задаются величиной т.н. уровня значимости критерия α. Всякое статистическое решение, т.е. решение, принимаемое на основании ограниченного ряда наблюдений, неизбежно сопровождается некоторой, хотя может и очень малой, вероятностью ошибочного заключения как в ту, так и в другую сторону. Скажем, в какой-то небольшой доле случаев α гипотеза H_0 может оказаться отвергнутой, в то время как на самом деле она справедлива, или, наоборот, в какой-то небольшой доле случаев β мы можем принять гипотезу, в то время как на самом деле она ошибочна, а справедливым оказывается некоторое конкурирующее с ней предположение – альтернатива H_1. При фиксированном объёме выборочных данных величину вероятности одной из этих ошибок можно выбирать по своему усмотрению. Если же объём выборки можно как угодно увеличивать, то имеется принципиальная возможность добиваться как угодно малых вероятностей обеих ошибок α и β при любом фиксированном конкурирующем предположительном утверждении H_1. В частности, при фиксированном объёме выборки обычно задаются величиной α вероятности ошибочного отвержения проверяемой гипотезы H_0, которую часто называют «осн.» или «нулевой». Эту вероятность ошибочного отклонения «нулевой» гипотезы принято называть уровнем значимости или размером критерия. Выбор величины уровня значимости α зависит от сопоставления потерь, понесённых в случае ошибочных заключений в ту или иную сторону: чем весомее потери от ошибочного отвержения высказанной гипотезы H_0, тем меньшей выбирается величина α. Однако поскольку такое сопоставление в большинстве практических задач оказывается затруднительным, пользуются некоторыми стандартными значениями уровня значимости:

$$\alpha = 0,1;\ 0,05;\ 0,025;\ 0,01;\ 0,005;\ 0,001.$$

Особенно распространённой является величина $\alpha = 0,05,$ означающая, что в среднем в пяти случаях из 100 будет ошибочно отвергаться высказанная гипотеза при пользовании данным статистическим критерием; в) задаются некоторой функцией от результатов наблюдения (критической статистикой)

$$\gamma^{(n)} = \gamma(X_1, X_2, \ldots, X_n).$$

Критической статистикой определяется мера расхождения имеющихся выборочных данных с высказанной (и проверяемой) гипотезой H_0. Эта критическая статистика $\gamma^{(n)}$, как и всякая функция от результатов наблюдения, сама является случайной величиной и в предположении справедливости гипотезы H_0 подчинена некоторому хорошо изученному (затабулированному) закону распределения с плотностью $f_{\gamma^{(n)}}(u)$. Для построения критической статистики, как правило, используется т.н. принцип отношения правдоподобия; г) из табл. распределения

$f_{\gamma^{(n)}}(u)$ находятся $100(1-\frac{\alpha}{2})\%$-ная точка $\gamma_{\alpha/2}^{(\min)}$ и $100\frac{\alpha}{2}\%$-ная точка $\gamma_{\alpha/2}^{(\max)}$

разделяющие всю область мыслимых значений случайной величины $\gamma^{(n)}$ на три части: область неправдоподобно малых (1), неправдоподобно больших (Ш) и естественных, или правдоподобных (в условиях справедливости гипотезы H_0) значений (II). В тех случаях, когда основную опасность для утверждения представляют только односторонние отклонения, т. е. только «слишком маленькие» или только «слишком большие» значения критической статистики $\gamma^{(n)}$ находят лишь одну процентную точку: либо 100 $(1-\alpha)\%$-ную точку $\gamma_{\alpha}^{(\min)}$ которая будет разделять весь диапазон значений $\gamma^{(n)}$ на две части: область неправдоподобно малых и область правдоподобных значений; либо $100\alpha\%$-ную точку $\gamma_{\alpha}^{(\max)}$ она будет разделять весь диапазон значений $\gamma^{(n)}$ на область неправдоподобно больших и область правдоподобных значений;

д) наконец, в функцию $\gamma^{(n)}$ подставляют имеющиеся конкретные выборочные данные $X_1,\ldots,X_n$ и подсчитывают численную величину $\gamma^{(n)}$. Если окажется, что вычисленное значение принадлежит области правдоподобных значений $\gamma^{(n)}$, то гипотеза H_0 считается не противоречащей выборочным данным. Если же $\gamma^{(n)}$ слишком мала или слишком велика, делается вывод, что $\gamma^{(n)}$ на самом деле не подчиняется закону $f_{\gamma^{(n)}}(u)$ (этот вывод сопровождается вероятностью ошибки, равной α) и высказанное предположение H_0 ошибочно. Т.о., решение, принимаемое на основании статистического критерия, может оказаться ошибочным как в случае отклонения проверяемой гипотезы H_0 (с вероятностью α), так и в случае её принятия (с вероятностью β). Вероятности α и β ошибочных решений называют также ошибками соответственно первого и второго рода, а величину $1-\beta$ — *мощностью критерия*. Очевидно, из двух критериев, характеризующихся одной и той же вероятностью α отвергнуть в действительности правильную гипотезу H_0, следует предпочесть тот, который сопровождается меньшей ошибкой второго рода (или большей мощностью).

Если проверяемое предположительное утверждение сводится к гипотезе о том, что значение некоторого параметра θ в точности равно заданной величине θ_0, то эта гипотеза называется простой; в других случаях гипотеза называется сложной.

По прикладной направленности в инструментарии М.с. можно выделить определённые разделы: разведочный (предварительный) анализ (включая описательную статистику, целенаправленное проецирование многомерных данных, различные приёмы визуализации и т. п.), регрессионный, дисперсионный, ковариационный и факторный анализы, методы классификации (включая дискриминантный анализ и расщепление смесей распределений), главных компонент анализ, анализ временных рядов, анализ таблиц сопряжённости.

МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

один из наиболее распространённых методов построения *точечных оценок* неизвестных параметров закона *распределения вероятностей*, состоящий в том, что в качестве оценок выбираются такие значения параметров, при которых данные результаты наблюдений наиболее вероятны; в совр. виде метод сформулирован Р. Фишером в 1912 .

Пусть θ – некоторый неизвестный параметр закона распределения (скалярный или векторный). Считая, что известен общий вид $p(x;\theta)$ закона распределения случайной величины X, соответствующей *ген. совокупности*, из которой извлечена выборка $x_1, x_2, \ldots, x_n$ (для дискретной случайной величины

$$p(x_i;\theta) = \mathrm{P}\{X = x_i\}$$

– это функция вероятности, а для непрерывной случайной величины – *функция плотности вероятности*

$$p(x;\theta) = f_X(x)),$$

по выборке $x_1, x_2, \ldots, x_n$ составляется функция правдоподобия

$$L(\theta) = = p(x_1;\theta)\,p(x_2;\theta)\cdots p(x_n;\theta),$$

равная в случае *случайной величины дискретной* вероятности получения именно набора $x_1, x_2, \ldots, x_n$ при извлечении выборки объёмом n из ген. совокупности, а в случае *случайной величины непрерывной* – функции плотности этой вероятности. Чем больше $L(\theta)$, тем вероятнее (или правдоподобнее) получить при наблюдениях именно данную конкретную выборку $x_1, x_2, \ldots, x_n$. Оценкой макс. правдоподобия параметра θ называют при этом такое значение $\theta_{\text{мп}}$, при котором функция правдоподобия $L(\theta)$ принимает макс. значение: $\theta_{\text{мп}} = \arg\max_{\theta} L(\theta)$. Во многих случаях вместо задачи определения макс. функции правдоподобия удобнее решать эквивалентную задачу максимизации *логарифмической функции правдоподобия*

$$l(\theta) = \ln L(\theta):\ \theta_{\text{мп}} = \arg\max_{\theta} l(\theta),$$

так как натуральный логарифм является монотонно возрастающей функцией. Макс. функции правдоподобия или её логарифма в большин-

ство случаев определяется путём приравнивания нулю первых производных (и проверке выполнимости достаточных условий максимума второго порядка). Так, при оценивании неизвестного параметра λ закона распределения Пуассона

$$p(x;\lambda) = \mathrm{P}\{X = x\} = \lambda^x \mathrm{e}^{-\lambda}/\,x!\ \text{с}$$

помощью М.м.п., функция правдоподобия имеет вид

$$L(\lambda) = \ = p(x_1;\lambda)\,p(x_2;\lambda)\cdots p(x_n;\lambda) \ =$$

$$\left(\lambda^{x_1}\,\mathrm{e}^{-\lambda}/\,x_1!\right)\left(\lambda^{x_2}\,\mathrm{e}^{-\lambda}/\,x_2!\right)\cdots\left(\lambda^{x_n}\,\mathrm{e}^{-\lambda}/\,x_n!\right) = \lambda^{\sum_{i=1}^{n} x_i}\,\mathrm{e}^{-n\lambda}/(x_1!\,x_2!\cdots x_n!)\,.$$

Отсюда логарифмическая функция правдоподобия

$$l(\lambda) = \ln L(\lambda) = \ln\left(\lambda^{\sum_{i=1}^{n} x_i}\,\mathrm{e}^{-n\lambda}/(x_1!\,x_2!\cdots x_n!)\right) =$$

$$= \sum_{i=1}^{n} x_i \ln\lambda - n\lambda - \ln(x_1!\,x_2!\cdots x_n!)\,.$$

Чтобы функция $l(\lambda)$ достигала макс. значения, её производная должна обратиться в нуль: $l'(\lambda) = 0$, а вторая производная должна быть отрицательной:

$l''(\lambda) < 0$.

Поскольку

$$l'(\lambda) = \sum_{i=1}^{n} x_i \Big/ \lambda - n\,,\ l''(\lambda) = -\sum_{i=1}^{n} x_i \Big/ \lambda^2\,,$$

то оценка макс. правдоподобия равна

$$\hat{\lambda}_{\text{мп}} = \sum_{i=1}^{n} x_i \Big/ n = \overline{x}$$

(при этом вторая производная отрицательна, так как в числителе дроби стоит положительное число, ввиду того, что ген. совокупность распределена по закону Пуассона и её элементы не могут принимать отрицательных значений, в знаменателе дроби также стоит положительное число $\hat{\lambda}^2_{\text{мп}}$, а перед дробью стоит знак «минус»). В ряде случаев производные функции правдоподобия могут быть не определены или нигде не обращаться в нуль – в таких ситуациях оценку макс. правдоподобия следует искать на границах области определения. Напр., если случайная величина X распределена по равномерному закону на отрезке $[a;b]$ (границы которого являются неизвестными параметрами), то $p(x;\theta) = f_X(x) = 1/(b-a)$ при $x \in [a;b]$ и $p(x;\theta) = f_X(x) = 0$ при $x \notin [a;b]$. Функция правдоподобия

$$L(a;b) = 1/(b-a)^n\,,$$

если все наблюдения $x_1, x_2, \ldots, x_n$ принадлежат отрезку $[a;b]$ и обращается в нуль в противном случае, производные

$\partial L/\partial a = n/(b-a)^{n+1} \neq 0$

и $\partial L/\partial b = -n/(b-a)^{n+1} \neq 0$,

но оценки макс. правдоподобия существуют: так как

a ,, $\min\{x_1, x_2, \ldots, x_n\}$,

а b . . .$\max\{x_1, x_2, \ldots, x_n\}$,

то максимум функции правдоподобия достигается при

$\hat{a}_{\text{мп}} = \min\{x_1, x_2, \ldots, x_n\}$

и $\hat{b}_{\text{мп}} = \max\{x_1, x_2, \ldots, x_n\}$.

При весьма общих условиях регулярности (накладываемых на изучаемый закон распределения) оценки макс. правдоподобия являются состоятельными, асимптотически несмещёнными (при $n \to \infty$), асимптотически эффективными и асимптотически нормальными. Если существует состоятельная и эффективная оценка θ параметра θ, то такая оценка – оценка макс. правдоподобия $\theta_{\text{мп}}$. Однако оценки макс. правдоподобия не всегда являются наилучшими. Во-первых, состоятельность оценок, асимптотическая несмещённость оценок, асимптотическая эффективность оценок и их асимптотическая нормальность могут начать проявляться только при очень больших объёмах выборки, которыми не располагают на практике. Во-вторых, для применения М.м.п. необходимо точно знать тип анализируемого закона распределения вероятностей, что не всегда выполнимо, и в результате оценка может резко потерять свои хорошие свойства при отклонении реального закона распределения исследуемой случайной величины от предполагаемого. Есте-

ственным образом М.м.п. распространяется на оценивание параметров случайных величин многомерных (векторных), случайных процессов и др. М.м.п. основан на идеях К. Ламберта и Д. Бернулли (18 в.), в частных случаях М.м.п. использовался К. Гауссом в 19 в.

См. также *Асимптотические свойства оценок.*

МЕТОД МОМЕНТОВ

один из наиболее распространенных методов построения *точечных оценок* неизвестных параметров закона *распределения вероятностей*, состоящий в приравнивании теоретических моментов случайной величины соответствующим *моментам выборочным*, и выражения из этой системы уравнений неизвестных параметров закона распределения вероятностей через элементы выборки. Предполагается, что известен общий вид $p(x;\theta)$ закона распределения вероятностей случайной величины X, соответствующей *ген. совокупности*, из которой извлечена *выборка* $x_1, x_2, \ldots, x_n$ (т. е. известна функция вероятности $p(x_i;\theta) = \mathrm{P}\{X = x_i\}$ для *случайной величины дискретной* или плотность распределения

$$p(x;\theta) = f_X(x)$$

для *случайной величины непрерывной*), где q – некоторый неизвестный параметр закона распределения (скалярный или векторный). Определяется зависимость

$$\theta = g(\nu_1, \nu_2, \ldots, \nu_k, \mu_1, \mu_2, \ldots, \mu_l)$$

параметра q от начальных моментов с первого по *k*-й и центральных моментов с первого по *l*-й, после чего для вычисления оценки $\theta_{\text{м}}$ параметра θ по методу моментов в эту зависимость вместо неизвестных теоретических моментов подставляют соответствующие *моменты выборочные*

$\hat{\nu}_i$ и $\hat{\mu}_j$: $\theta_{\text{мм}} = g(\hat{\nu}_1, \hat{\nu}_2, \ldots, \hat{\nu}_k, \hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_l)$.

Очевидным достоинством метода моментов является его простота, при некоторых довольно общих условиях метод моментов позволяет найти оценки, которые при $n \to \infty$ являются асимптотически нормальными и смещенными не более чем на величину порядка $1/n$, однако качество оценок, полученных с помощью этого метода, не всегда бывает высоким, особенно при небольших объемах выборки. Примером, когда метод моментов приводит к такому же результату, как и *метод макс. правдоподобия*, а именно, к состоятельной, несмещённой и эффективной оценке параметра, является закон распределения Пуассона: начальный момент первого порядка равен

$$\nu_1 = \mathrm{M}X = \lambda,$$

откуда $\lambda = \nu_1 = \mathrm{M}X$.

Поэтому оценкой метода моментов для параметра λ будет

$$\hat{\lambda}_{\text{м}} = \hat{\nu}_1 = \overline{x}.$$

Однако можно было выразить неизвестный параметр λ через центральный момент второго порядка $\mu_2 = \mathrm{D}X = \lambda$ и получить альтернативную оценку параметра λ по методу моментов

$$\hat{\lambda}^{(2)}_{\text{м}} = \hat{\mu}_2 = \hat{\sigma}^2 \neq \hat{\lambda}_{\text{м}}.$$

В случае равномерного закона распределения с помощью метода моментов вовсе не представляется возможным получить хорошие оценки параметров. Рассмотрим, напр., первый начальный и второй центральный моменты:

$$\nu_1(X) = \mathrm{M}X = (a+b)/2,$$
$$\mu_2(X) = \mathrm{D}X = (b-a)^2/12.$$

Выражая отсюда a и b (учитывая, что $a < b$), и заменяя теоретические моменты их выборочными аналогами, получаем оценки по методу моментов

$$\hat{a}_{\text{м}} = \hat{\nu}_1 - \sqrt{3\hat{\mu}_2} = \overline{x} - \sqrt{3\hat{\sigma}^2_X},$$
$$\hat{b}_{\text{м}} = \hat{\nu}_1 + \sqrt{3\hat{\mu}_2} = \overline{x} + \sqrt{3\hat{\sigma}^2_X},$$

не совпадающие с оценками по методу макс. правдоподобия

$$\hat{a}_{\text{мп}} = \min\{x_1, x_2, \ldots, x_n\}$$

и $\hat{b}_{\text{мп}} = \max\{x_1, x_2, \ldots, x_n\}$.

МЕТОД МОНТЕ-КАРЛО

метод статистических испытаний, метод исследования поведения систем управления со случайными воздействиями, состоящий в компьютерной имитации изучаемых процессов при помощи моделирования случайных величин. Идея

М.М.-К. состоит в том, что строится вероятностная модель реального процесса, затем с помощью компьютерно реализованных датчиков случайных чисел проводят большое число испытаний, в которых моделируется поведение исследуемой системы. По результатам большого числа испытаний на основании *закона больших чисел* делают выводы о поведении реального процесса. Название метода связано с городом Монте-Карло, известного своими казино, и в частности, игрой в рулетку, которая представляет собой один из простейших примеров датчиков случайных чисел. Применение М.М.-К. часто дает существенный экономический эффект, позволяя вместо проведения дорогостоящих (или принципиально невозможных) экспериментов с реальным объектом «проиграть», имитировать его поведение на компьютере. Процесс получения значений случайной величины с заданным законом распределения с помощью датчиков случайных чисел называется моделированием (или разыгрыванием) этой случайной величины. В «Анализ данных» пакета Microsoft Excel входит программа «им случайных чисел», позволяющая получить последовательности значений различных случайных величин: *случайной величины дискретной*, заданной своим рядом распределения; альтернативной (т.е. принимающей значение 1 с вероятностью p и значение 0 с вероятностью $1-p$, в программе «Генерация случайных чисел» данный закон распределения назван законом Бернулли); биномиальной случайной величины с заданным числом испытаний n и вероятностью успеха в единичном испытании p; случайной величины, распределённой по закону Пуассона с параметром λ; случайной величины, распределённой по равномерному закону на отрезке $[a; b]$; случайной величины, распределённой по нормальному закону с заданным *математическим ожиданием* a и средним квадратичным отклонением σ. На самом деле для реализации М.М.-К. достаточно иметь датчик случайных чисел, являющихся реализациями случайной величины, распределённой по равномерному закону на отрезке [0; 1]. Напр., требуется смоделировать систему массового обслуживания, состоящую из продавца и покупателей в магазине. Предполагая, что покупатели образуют простейший поток событий, можно считать, что количество покупателей за время τ имеет закон распределения Пуассона с параметром $\mu_1\tau$ (где интенсивность поступления заказов μ_1 – число покупателей в единицу времени – может быть определена по выборочным данным), а время между двумя последовательно пришедшими покупателями распределено по показательному закону с параметром μ_1. Время обслуживания покупателя продавцом можно считать показательной случайной величиной с параметром μ_2, равным интенсивности обслуживания, которая также может быть определена по статистическим данным. Теперь можно смоделировать рабочий день магазина и выяснить значения важных параметров: среднего времени дневного простоя продавцов, средней длины очереди и т. п. Для этого нужно провести расчёты по следующей схеме. В начале рабочего дня продавец свободен. Он будет свободен в течение времени, пока не появится первый покупатель, которое получим как реализацию случайной величины, распределённой по показательному закону с параметром μ_1. После прихода покупателя продавец будет занят его обслуживанием в течение времени, которое представляет собой случайную величину, распределенную по показательному закону с параметром μ_2. Реализацию этой случайной величины также получим с помощью датчика случайных чисел. Время прихода следующего покупателя вновь определим как реализацию экспоненциальной случайной величины с параметром μ_1. Если в момент прихода нового покупателя продавец еще занят обслуживанием предыдущего, новый покупатель встает в очередь, и её длина увеличивается на единицу. В момент окончания обслуживания очередного покупателя продавец начинает обслуживание нового покупателя (если есть очередь к этому моменту) либо начинает простаивать. Проведя достаточно много испытаний, можно быть уверенным, что в силу *закона больших чисел* средние значения параметров реального процесса будут близки к средним значениям, полученным в ходе моделирования. Моделирование произвольной случайной величины X с помощью датчика случайных чисел, распределенных по равномерному закону на от-

резке [0; 1], основано на том факте, что случайная величина

$Y = F_X(X)$,

где $F_X(x)$ – функция распределения вероятностей случайной величины X, имеет равномерный закон распределения на отрезке [0; 1]. Действительно, при $y \le 0$ значение функции распределения вероятностей случайной величины $Y = F_X(X)$ равно $F_Y(y) = \mathrm{P}\{Y < y\} = 0$, при $y > 1$ значение $F_Y(y) = \mathrm{P}\{Y < y\} = 1$, поскольку $F_X(X) \in [0;1]$.

При $0 < y \le 1$ значение

$$F_Y(y) = \mathrm{P}\{Y < y\} = \mathrm{P}\{F_X(X) < y\} = \mathrm{P}\left\{X < F_X^{-1}(y)\right\} = F_X\left(F_X^{-1}(y)\right) = y.$$

[здесь $F_X^{-1}(y)$ обозначает функцию, обратную к функции $F_X(x)$]. Отсюда следует, что если y представляет собой значение случайной величины, распределенной по равномерному закону на отрезке [0; 1], полученное с помощью датчика случайных чисел, то для получения значения x случайной величины X достаточно решить уравнение $F_X(x) = y$. Напр., для моделирования времени прихода первого покупателя, которое распределено по показательному закону с параметром μ_1, достаточно взять реализацию y случайной величины, распределенной по равномерному закону (на отрезке [0; 1]), и определить x как решение уравнения

$1 - \mathrm{e}^{-\mu_1 x} = y \Leftrightarrow x = -\ln(1-y)/\mu_1$.

Вычисления можно сократить, если заметить, что если случайная величина Y распределена по равномерному закону на отрезке [0; 1], то точно такой же закон распределения будет и у случайной величины $1 - Y$, поэтому вместо формулы

$x = -\ln(1-y)/\mu_1$

для моделирования показательной случайной величины используют соотношение

$x = -\ln y/\mu_1$.

Аналогичным образом моделируют и другие случайные величины. М.М.-К. может быть применён и для решения задач, на первый взгляд далеких от вероятностно-статистических, напр., для вычисления определенного интеграла $\int_a^b g(t)dt$ от ограниченной функции (для определенности будем считать, что $0 \le g(t) \le b$) можно многократно смоделировать две случайные величины X и Y, распределенные по равномерному закону на отрезке $[0; a]$ и на отрезке $[0; b]$ соответственно, вычислить относительную частоту $\hat{p}_n$ наступления события $g(X) < Y$ и определить искомый интеграл из соотношения

$\hat{p}_n = \int\!\!\int g(t)dt \,/(ab)$,

представляющего собой формулу геометрической вероятности. Идеи М.М.-К. уходят корнями в далекое прошлое, когда классики теории вероятностей проверяли на практике теоретические результаты. В совр. виде и с таким названием М.М.-К. предложен в 1949 Н. Метрополисом и С. Уламом. Далее М.М.-К. получил развитие в работах Н.П. Бусленко, А.А. Вавилова, В.С. Владимирова, И.М. Гельфанда, В.М. Глушкова, С.М. Ермакова, И.Н. Коваленко, Г.И. Марчука, Г.А. Михайлова, Н.Н. Моисеева, Дж. фон Неймана, И.М. Соболя, В.В. Чавчанидзе, Ю.А. Шрейдера и Р. Экхардта. М.М.-К. применяется в различных областях математики, экономики, физики и техники. В качестве примеров помимо имитационных моделей отметим генетические алгоритмы и генетическое программирование.

МЕТОДЫ ОЦЕНКИ НЕПАРАМЕТРИЧЕСКИЕ

группа методов построения *оценок статистических* по результатам наблюдений. Название «непараметрические» подчеркивает их отличие от классических (параметрических) методов *математической статистики*, в которых считается, что неизвестный теоретический закон распределения *ген. совокупности* принадлежит некоторому семейству, зависящему от конечного числа параметров (напр., может предполагаться, что ген. совокупность распределена по нормальному закону с двумя неизвестными параметрами a и σ, далее по *выборке* можно построить *точечные оценки* этих неизвестных параметров, проверить некоторые *гипотезы статистические* и т.д.).

Особенность М.о.н. состоит в их независимости от неизвестного теоретического закона распределения ген. совокупности. Так, *среднее значение выборочное* вне зависимости от закона распределения является оценкой несмещённой, *состоятельной и эффективной математического ожидания* ген. совокупности; исправленная *выборочная дисперсия* является состоятельной и несмещённой оценкой *дисперсии* ген. совокупности. Приведём пример непараметрической оценки *функции плотности вероятности* и *функции распределения вероятностей* непрерывной ген. совокупности. Предположим, что по выборке $x_1, x_2, \ldots, x_n$ построен интервальный *вариационный ряд*, т.е. набор *интервалов группирования*

$[a_1; a_2)$, $[a_2; a_3)$, …, $[a_\nu; a_{\nu+1})$

с указанием интервальных относительных частот $\hat{p}_j = m_j / n$,

где m_j – интервальная частота, т.е. число элементов выборки, попавших в j-й интервал ($j = 1, 2, \ldots, \nu$). Предположим, что длины всех интервалов группирования равны одному и тому же числу $\Delta = a_2 - a_1 = a_3 - a_2 = \ldots = a_{\nu+1} - a_\nu$. Непараметрической оценкой функции плотности вероятности $f_X(x)$ случайной величины X служит эмпирическая (выборочная) функция плотности, рассчитываемая по интервальному вариационному ряду как

$\hat{f}_X(x) = \hat{p}_j / \Delta$

при $x \in (a_j; a_{j+1}]$, $j = 1, 2, \ldots, \nu$, и $\hat{f}_X(x) = 0$

при $x \notin (a_1; a_\nu]$. При этом ломаная с вершинами в точках

$\left(x'_j;\ \hat{f}_X(x'_j)\right)$,

где через $x'_j = (a_j + a_{j+1})/2$

обозначены середины интервалов группирования, называется *полигоном* частот, а фигура, состоящая из прямоугольников, в основании которых лежат интервалы группирования $(a_j; a_{j+1}]$, а высотами являются значения $\hat{f}_X(x'_j)$, называется *гистограммой*. Непараметрической оценкой функции распределения вероятностей $F_X(x)$ случайной величины X служит эмпирическая (выборочная) функция распределения, рассчитываемая по интервальному вариационному ряду следующим образом:

$\hat{F}_X(x) = 0$ при $x \le a_1$;

$\hat{F}_X(x) = 1$ при $x > a_{\nu+1}$;

$\hat{F}_X(x) = \sum_{k=1}^{i} \hat{p}_k$ при $x \in (a_j; a_{j+1}]$,

$j = 1, 2, \ldots, \nu$. При этом ломаная с вершинами в точках $\left(x'_j; \hat{F}_X(x'_j)\right)$ называется *кумулятой*. В пакете Microsoft Excel реализован автоматизированный расчёт интервальных частот, построение полигона, гистограммы и кумуляты. Эмпирическая (выборочная) функция распределения $\hat{F}_X(x)$ является для функции распределения вероятностей $F_X(x)$ случайной величины X в каждой точке x оценкой состоятельной и несмещённой. Более того, если обозначить

$D_n = \sup_{x \in \mathbb{R}} \left|\hat{F}_X(x) - F_X(x)\right|$,

то случайная величина $\sqrt{n} D_n$ имеет функцию распределения

$K_n(\lambda) = \mathrm{P}\left\{\sqrt{n} D_n < \lambda\right\}$,

не зависящую от $F_X(x)$ и при $n \to \infty$ стремящуюся к пределу

$K(\lambda) = \sum_{t=-\infty}^{+\infty} (-1)^t \, \mathrm{e}^{-2t^2\lambda^2}$.

Функция $K(\lambda)$ табулирована, напр.,

$K(1{,}36) = 0{,}95$, $K(1{,}63) = 0{,}99$.

На данном результате основан ряд *критериев непараметрических*, в частности, *критерий согласия Колмогорова*. К методам оценки непараметрическим относятся также методы оценивания медианы, квантилей, процентных точек, *моды*, а также различные ранговые методы, в частности, *ранговые коэффициенты корреляции*, *коэффициент конкордации* и др. Первые *методы оценки непараметрические* были использованы в 19 в. А. Лежандром, К. Гауссом и П. Лапласом. В 1930–50 существенный вклад в развитие М.о.н. внесли А.Н. Колмогоров, Р. Мизес, В.И. Гливенко, М. Розенблатт, Н.В. Смирнов и Н.Н. Ченцов. Развитие М.о.н привело к появлению совр. подходов к построению робастных и адаптивных оценок, развитию *бутстреп-моделирования*, метода складного ножа, теории рекуррентного оценивания и др.

МЕТОДЫ РОБАСТНЫЕ

статистические методы, которые позволяют получать достаточно надёжные оценки статистической совокупности с учётом неявного закона её распределения и наличия существенных отклонений в значениях данных. У истоков развития методов робастного оценивания стояли американский статистик Д. Тьюки и швейцарский математик П. Хьюбер.

При решении задач робастного оценивания выделяют два типа данных, засоряющих статистическую совокупность. К первому типу относят данные, несущественно отличающиеся от значений, которые наиболее часто встречаются в изучаемой совокупности. Эти данные не вызывают значительных искажений в аналитических результатах и могут обрабатываться обычными методами статистического оценивания.

Второй тип данных – резко выделяющиеся на фоне изучаемой совокупности, их называют «засорением» или «грубыми ошибками», они оказывают сильное искажающее воздействие на аналитические результаты. Эти данные должны подвергаться специальной обработке. В практике устойчивого оценивания различают следующие основные причины появления грубых ошибок: специфические особенности отдельных элементов изучаемой совокупности. Как правило, они приводят к появлению случайных, или «нормальных» («обычных») отклонений; неправильное причисление элементов к исследуемой совокупности, например, ошибки группировки, ошибки при организации наблюдения и т. п.; грубые ошибки при регистрации и обработке данных.

Если грубые ошибки – результат неправильных причислений элементов или ошибок регистрации, то их появление и уровни непредсказуемы, а распределение может значительно отклоняться от гипотетического распределения осн. массива статистических данных. При обработке «грубых» ошибок (засорений) можно выделить два осн. подхода. Первый ориентирован на устранение из выборочной совокупности ошибок и оценку параметров по оставшимся «истинным» значениям. Второй подход предполагает в каждом случае с грубой ошибкой выделение истинных значений признака и собственно ошибки

$$x = x_{ист} + \xi;$$

при этом осуществляется модификация данных т.о., чтобы искажающий элемент ξ получил нормальное распределение с нулевым математическим ожиданием. Тогда для некоторого множества грубых ошибок вариативной величины x сумма $\sum \xi$ приближается к нулю, а оценки $\hat{x}$ – к истинным значениям параметров выборочной совокупности.

МОДА

числовая характеристика закона *распределения вероятностей* случайной величины, соответствующая её наиболее часто встречающемуся значению. М. дискретной случайной величины X называется значение этой случайной величины $x_{\text{mod}} = x_i$, соответствующее наибольшей вероятности $p_i = \max_j p_j$. М. абсолютно непрерывной случайной величины X называется точка локального максимума плотности распределения:

$$f_X(x_{\text{mod}}) = \max_{x \in \square} f_X(x).$$

В общем случае случайная величина может иметь более одной моды. Распределения, имеющие единственную моду, называются унимодальными и играют наиболее важную роль в *теории вероятностей*, *математической* и прикладной *статистике.* Для непрерывных случайных величин, плотность распределения которых симметрична относительно оси ординат, *математическое ожидание*, М. и *медиана* совпадают. В случае же, когда распределение случайной величины не является симметричным, М. и медиана являются достаточно важными характеристиками. Проиллюстрируем это на примере. Предположим, что производитель одежды собирается выпустить новую женскую юбку, для чего проводит опрос женщин относительно идеальной, с их точки зрения, длины юбки. Пусть X – длина, названная случайно отобранной женщиной. Если выпустить юбку длины MX, то, скорее всего, ни одна из женщин не будет полностью удовлетворена, поскольку почти наверняка MX не совпадет ни с одним из

названных женщинами значений. Если выпустить юбку длины x_{med}, то окажется, что ровно половина женщин предпочитает юбки более короткие, чем эта, а другая половина – более длинные. А вот если выпустить юбку длины x_{mod}, то такая длина удовлетворит макс. количество женщин. По *выборке* можно вычислить *точечную оценку* М. Для дискретной ген. совокупности выборочная М. – наиболее часто встречающийся элемент выборки, а для непрерывной ген. совокупности выборочная М. вычисляется по интервальному *вариационному ряду*:

$$\hat{x}_{\text{mod}} = a_m + \Delta\left(\hat{p}_m - \hat{p}_{m-1}\right)/\left(2\hat{p}_m - \hat{p}_{m-1} - \hat{p}_{m+1}\right).$$

Здесь Δ – ширина интервала группирования, a_m – начало модального интервала, т.е. такого интервала группирования $(a_m; a_{m+1})$, что

$$\hat{p}_m = \max_{i=1,2,\ldots,\nu} \hat{p}_i .$$

МОДЕЛЬ ЭКОНОМИКО-СТАТИСТИЧЕСКАЯ

система математических соотношений, которая описывает некоторый экономический объект, и параметры которой определяются с помощью статистических методов на основании фактических данных. Структура и конкретный вид М.э.-с. определяются спецификой моделируемого объекта, теоретическими представлениями исследователя об объекте, целями исследования, доступной информацией, используемыми статистическими методами и информационными технологиями. Процесс моделирования состоит из нескольких этапов: на первом этапе определяется общий вид модели, входящие в нее экзогенные переменные и эндогенные переменные, соотношения между ними; на втором этапе производится статистическое оценивание неизвестных параметров модели на основе данных наблюдений; далее с помощью модели решают задачи анализа состояния моделируемого объекта, *прогнозирования* его развития.

См. также *Вероятностная модель*, *Вероятностно-статистическая модель*, *Модель имитационная*, *Модель эконометрическая* *Модель, экономико-математическая*.

МОМЕНТЫ ВЫБОРОЧНЫЕ

средние значения выборочные произведений степеней отклонения *случайных величин* от некоторых чисел, служащие оценками *математических ожиданий* этих произведений. В общем случае смешанным М.в. порядка

$$k_1 + k_2 + \cdots + k_m$$

случайной величины многомерной (векторной) X относительно неслучайного вектора

$\mathbf{c} = (c_1, c_2, \ldots, c_m)$ называется величина

$$\hat{M}_{k_1, k_2, \ldots, k_m} = = \overline{(X^{(1)} - c_1)^{k_1} (X^{(2)} - c_2)^{k_2} \cdots (X^{(m)} - c_m)^{k_m}} .$$

Если все c_i равны нулю $(i = 1, 2, \ldots, m)$, то М.в. называются начальными, а если

$c_i = \overline{X^{(i)}}$ $(i = 1, 2, \ldots, m)$,

то центральными. По аналогии с теоретическими начальными моментами случайной величины X порядка

k $(k = 0, 1, 2, \ldots)$,

т.е. величинами

$\nu_k(X) = \mathrm{M}(X^k)$,

и теоретическими центральными моментами порядка k $(k = 0, 1, 2, \ldots)$ этой случайной величины [т. е. величинами

$\mu_k(X) = \mathrm{M}(X - \mathrm{M}X)^k$]

начальные и центральные моменты выборочные порядка

k $(k = 0, 1, 2, \ldots)$ – статистики $\hat{\nu}_k = \overline{X^k}$

и $\hat{\mu}_k = \overline{(X - \bar{X})^k}$

соответственно. Начальный М.в. первого порядка случайной величины X – её среднее значение выборочное

$\hat{\nu}_1(X) = \bar{X}$,

а центральный М.в. второго порядка – её выборочная дисперсия

$\hat{\mu}_1(X) = \hat{\sigma}_X^2$.

Смешанный центральный М.в. второго порядка двумерной случайной величины

$$\hat{\mu}_{1,1}=\overline{\left(X^{(1)}-\overline{X^{(1)}}\right)\left(X^{(2)}-\overline{X^{(2)}}\right)}$$

называется выборочной ковариацией, её компонент $X^{(1)}$ и $X^{(2)}$, обозначается

$$\operatorname{cov}\left(X^{(1)},X^{(2)}\right)$$

и служит одной из осн. выборочных характеристик степени близости связи между случайными величинами $X^{(1)}$ и $X^{(2)}$ к линейной функциональной. Свойства М.в. аналогичны свойствам соответствующих теоретических моментов:

$$\hat{\nu}_0(X)=\hat{\mu}_0(X)=1;\ \hat{\mu}_1(X)=0;$$

$$\hat{\mu}_2(X)=\hat{\sigma}_X^2=\hat{\nu}_2(X)-\hat{\nu}_1^2(X);$$

$$\hat{\mu}_3(X)=\hat{\nu}_3(X)-3\hat{\nu}_1(X)\hat{\nu}_2(X)+2\hat{\nu}_1^3(X);\ \hat{\mu}_4(X)=$$

$$=\hat{\nu}_4(X)-4\hat{\nu}_1(X)\hat{\nu}_3(X)+6\hat{\nu}_1^2(X)\hat{\nu}_2(X)-3\hat{\nu}_1^4(X);\ \hat{\mu}_{1,1}=\hat{\nu}_{1,1}-\hat{\nu}_1^{(1)}\hat{\nu}_1^{(2)}=$$

$$=\overline{X^{(1)}X^{(2)}}-\overline{X^{(1)}}\cdot\overline{X^{(2)}}.$$

См. также *Ковариационная матрица*, *Корреляционное отношение*, *Коэффициент корреляции*.

МОЩНОСТЬ КРИТЕРИЯ

характеристика качества *критерия статистического*, вероятность не совершить ошибку второго рода. Если проверяется *гипотеза статистическая* простая $H_0:\theta=\theta_0$ при простой альтернативной гипотезе $H_1:\theta=\theta_1$, то вероятность ошибки второго рода равна

$$\beta=\mathbf{P}\{H_0|H_1\}=\mathbf{P}\{\psi\notin S_{\text{кр}}|H_1\},$$

а мощность критерия –

$$1-\beta=\mathbf{P}\{H_1|H_1\}=\mathbf{P}\{\psi\in S_{\text{кр}}|H_1\},$$

где ψ – критическая статистика, $S_{\text{кр}}$ – критическая область. Т.о., М.к. равна вероятности попадания критической статистики критерия в критическую область при условии ложности гипотезы H_0. Если альтернативная гипотеза является сложной:

$$H_1:\theta\in\Theta_1,$$

то М.к. определяется как

$$1-\beta=\inf_{\theta\in\Theta_1}\mathrm{P}\{H_1|H_1\}=\inf_{\theta\in\Theta_1}\mathrm{P}\{\psi\in S_{\text{кр}}|H_1\}.$$

В теории *статистической проверки гипотез*, основанной в 1930-х гг. Ю. Нейманом и Э. Пирсоном, задача проверки гипотезы формулируется следующим образом: фиксируется достаточно малый уровень значимости критерия α (напр., $\alpha=0,05$ или $\alpha=0,01$), и строится критерий, который имеет наибольшую мощность при условии, что вероятность *ошибки первого рода* не превышает заданный уровень значимости α. При этом реальные значения М.к. в конкретных задачах проверки гипотез часто оказываются значительно меньше желаемого уровня, поэтому в таких ситуациях говорят не о том, чт.е. основания принять гипотезу H_0, а о том, что нет оснований её отвергнуть, ведь приняв H_0, можно с достаточно большой вероятностью β совершить ошибку второго рода.

См. также *Критерий статистический наиболее мощный*.

Н

НАКОПЛЕННАЯ ЧАСТОТА

выборочная характеристика, определяемая как число элементов *выборки* $x_1,x_2,\ldots,x_n$, меньших данного числа *x*. Пусть $x_1',x_2',\ldots,x_l'$ ($l\le n$) – варианты, т.е. различные элементы выборки, расположенные в порядке неубывания, $m_1,m_2,\ldots,m_l$ – соответствующие этим вариантам частоты (m_i – количество появлений в выборке числа x_i'). При этом объём выборки равен сумме всех частот:

$$n=\sum_{i=1}^{l}m_i.$$

Выборочной случайной величиной называется *случайная величина дискретная* $\hat{X}$, принимающая значения x_i' $(i=1,2,\ldots,l)$ с вероятно-

стями $\hat{p}_i = m_i / n$ – относительными частотами. Н.ч. (к числу x) – число

$$\sum_{k:\, x_i' < x} m_i ,$$

а Н.ч. относительная (накопленная частость) – число

$$\mathrm{P}\{\hat{X} < x\} = \sum_{k:\, x_i' < x} \hat{p}_i = \sum_{k:\, x_i' < x} m_i \Big/ n .$$

С помощью Н.ч. относительных строятся оценки функции распределения ген. совокупности.

См. также *Методы оценки непараметрические*, *Эмпирическая (выборочная) функция распределения*.

НЕЗАВИСИМОСТЬ НАБЛЮДЕНИЙ

Пусть задана последовательность случайных величин $x_1, x_2, ..., x_n$, i -й член которой (x_i) лишь обозначает результат наблюдения, который мы могли бы получить при i-м наблюдении n-кратного эксперимента, связанного с наблюдением исследуемой случайной величины ξ (какого-либо признака у единиц рассматриваемой совокупности). При этом переход от i-го наблюдения выборки к i+1-му не обязательно выполняется в хронологической последовательности: выбранные для наблюдения объекты могут образовывать так называемую «пространственную выборку» и наблюдаться, напр., одновременно.

Если условия эксперимента не меняются от наблюдения к наблюдению и если n-кратный эксперимент организован т.о., что результаты наблюдения на каждом (i-м) шаге никак не зависят от предыдущих и не влияют на будущие результаты наблюдений, то говорят, что наблюдения независимы. В этом случае очевидно, что вероятностные закономерности поведения i-го наблюдения рассматриваемой гипотетической выборки $x_1, x_2, ..., x_n$ остаются одними и теми же для всех $i = 1,2,..., n$ и полностью определяются законом распределения вероятностей наблюдаемой случайной величины, т.е.

$$P\{x_i < x\} = P\{\xi < x\} = F_{\xi}(x) .$$

При этом из взаимной независимости наблюдений выборки следует, что последовательность случайных величин $x_1, x_2, ..., x_n$ состоит из независимых компонент, т.е. их совместная функция распределения

$$F_{(x_1, x_2, ..., x_n)}(z_1, z_2, ..., z_n) = P\{x_1 < z_1, ..., x_n < z_n\}$$

может быть представлена в следующем виде:

$$F_{(x_1, x_2, ..., x_n)}(z_1, z_2, ..., z_n) = \prod_{i=1}^{n} P\{x_i < z_i\} = \prod_{i=1}^{n} F_{\xi}(z_i)$$

Если ряд наблюдений образует последовательность независимых и одинаково распределенных случайных величин, то выборка называется случайной. Число n наблюдений, образующих выборку, называют объёмом выборки.

НОМИНАЛЬНАЯ ШКАЛА

(классификационная или шкала наименований) – шкала, при которой все измеряемые объекты или значения измеряемых свойств представляются как множество непересекающихся и исчерпывающих всю совокупность классов. При этом каждому классу дается наименование или присваивается знак. Эти символы служат только для целей идентификации обозначаемых объектов или для их нумерации (в случае использования цифр), затем каждому элементу из одного класса объектов приписывается одно имя (знак, цифра). Порядок расположения отдельных значений номинальной шкалы может быть любым. В Н.ш. допустимыми являются все взаимнооднозначные преобразования. При использовании Н.ш. невозможно определить количество классов и упорядочить их, т.е. номер класса не отражает его количественного содержания. Н.ш. не допускает никаких математических операций со значениями. Например, числа нельзя складывать и вычитать, но можно

подсчитывать столько раз, сколько встречается то или иное число. Но возможно применение некоторых *статистических непараметрических критериев*, оценок и сравнений *распределений частот* различных классов в выборках, характеристик этих распределений (*моды, медианы*). Для величин, оцененных в Н.ш. неприменимы любые *параметрические меры*, такие как *среднее, дисперсия, коэффициенты корреляции* и даже некоторые *непараметрические статистики*. С помощью шкальных значений номинальной шкалы можно установить только, относятся ли (или не относятся) два данных объекта к одному и тому же классу. В шкале наименований измерены, например, номера паспортов, автомашин, страховых свидетельств государственного пенсионного страхования, медицинского страхования и т.д.. Частным случаем Н.ш. является дихотомическая шкала, фиксирующая наличие или отсутствие у объекта некоторого свойства. Признак измеренный по дихотомической шкале называется альтернативным. Так наличие качества принято обозначать числом "1", его отсутствие – числом "0". Н.ш. являются наиболее слабыми среди шкал с точки зрения измеримости. К Н.ш. относятся атлас цветов (шкала цветов) или шкала (классификация) растений Карла Линнея

НУЛЕВАЯ ГИПОТЕЗА

выдвинутая гипотеза, обозначаемая H_0, которая используется для её сопоставления с соответствующими характеристиками выборочных данных $x_1, x_2, ..., x_n$. При статистическом исследовании иногда возникает необходимость в формулировке и экспериментальной проверке некоторых предположительных утверждений (гипотез) относительно природы или величины неизвестных параметров анализируемой стохастической системы. Напр., исследователь высказывает предположение о типе закона распределения исследуемой случайной величины, о среднем значении анализируемой ген. совокупности, о том, что выборочные данные извлечены из нормальной ген. совокупности и т.д. Результат подобного сопоставления (проверки) может быть либо отрицательным (данные наблюдения противоречат высказанной гипотезе), либо неотрицательным (данные наблюдения не противоречат высказанной гипотезе). В итоге проверки гипотезы могут быть допущены ошибки двух родов. *Ошибка первого рода* состоит в том, что будет отвергнута правильная нулевая гипотеза. Вероятность ошибки первого рода называют уровнем значимости и обозначают через α. Ошибка второго рода состоит в том, что будет принята неправильная Н.г.; вероятность ошибки второго рода обозначают через β.

О

ОБОБЩЁННАЯ ДИСПЕРСИЯ

многомерной случайной величины – предложена С. Вилксом как скалярная мера рассеяния. О.д. для *p*-мерного случайного вектора *X* вычисляется как |Σ| – определитель *ковариационной матрицы* Σ. Выборочная О.д. вычисляется как |S| – определитель выборочной ковариационной матрицы S. О.д. многомерной случайной величины. Стандартизованная О.д., определяется как корень *p*-ой степени из О.д. и используется для сравнения рассеяния случайных векторов различной размерности.

О.д. зависит от шкалы измерения каждой из переменных и поэтому трудно поддается содержательной интерпретации. О.д. обладает рядом полезных свойств. Для оптимальной оценки многомерного параметра О.д. равна обратной величине определителя информационной матрицы. Ортогональное преобразование случайного вектора оставляет инвариантной О.д. В линейных моделях при предположении о нормальности оптимальный доверительный эллипсоид оцениваемой функции параметров имеет (наименьший) объём, который обратно пропорционален О.д. оптимальной оценки.

О.д. играет важную роль в статическом оценивании многомерных параметров как мера эффективности оценок, используется в *планировании эксперимента, статистическом контроле качества,* некоторых разделах многомерного статистического анализа, в частности,

анализе канонических корреляций, классификации многомерных наблюдений с обучением, дискриминантном анализе и др.

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

статистический метод анализа результатов измерений, предназначенный для проверки зависимости нормально распределённой случайной величины Y, называемой результативным признаком, от одного фактора, который может быть номинальным, порядковым или количественным. По природе фактора однофакторные модели подразделяются на детерминированные (М1) и случайные (М2), в зависимости от того, являются ли уровни факторного признака фиксированными или случайными. Под уровнем фактора здесь понимается некоторая его мера или состояние.

Есть задача О.д.а.: пусть требуется проверить влияние на результативный признак Y одного контролирующего фактора A, имеющего m уровней A_j, $j = 1, 2, ..., m$. Наблюдаемые значения результативного признака Y на каждом из уровней A_j обозначим через y_{ij}, где $i = 1, 2, ..., n_j$ – число наблюдений Y на уровне A_j (в качестве Y можно, напр., рассмотреть объём выполненных работ бригадой на стройке за смену. В нашем случае A_j – номер бригады). Наблюдаемые значения результативного признака обычно представляются в виде матрицы наблюдений:

$$\begin{pmatrix} y_{11} & y_{12} & \dots & y_{1m} \\ y_{21} & y_{22} & \dots & y_{2m} \\ \dots & \dots & \dots & \dots \\ y_{n_1 1} & y_{n_2 2} & \dots & y_{n_m m} \end{pmatrix},$$

$i = \overline{1, n_j}$; $j = \overline{1, m}$.

Однофакторная дисперсионная модель имеет следующий вид:

$$y_{ij} = a + \mu_j + \varepsilon_{ij}, \quad i = \overline{1, n_j}; \; j = \overline{1, m},$$

где y_{ij} – наблюдаемое значение результативного признака, полученного на j-ом уровне фактора с i-ым порядковым номером; a – ген. среднее всех мыслимых результатов наблюдений, или, иначе говоря, *математическое ожидание* результативного признака Y; μ_j – эффект, обусловленный влиянием j-го уровня фактора на Y, или, иначе, отклонение математического ожидания a_j результативного признака при на j-ом уровне фактора от общего математического ожидания a, т.е.

$\mu_j = a_j - a$.

Для модели М1 μ_j – фиксированные величины, удовлетворяющие условию

$$\sum_{j=1}^{m} \mu_j n_j = 0.$$

Для модели М2 μ_j – случайные величины, удовлетворяющие условиям:

$M\mu_j = 0; M\mu_i\mu_j = 0$ для $i \neq j$;

$M\mu_j\varepsilon_{ij} = 0$ для любых i, j;

$M\mu_j^2 = \sigma_\mu^2$ – факторная дисперсия.

ε_{ij} – случайные величины (остатки), отражающие влияние на Y всех неконтролируемых факторов, удовлетворяющие следующим условиям:

$M\varepsilon_{ij} = 0$ для любых i, j;

$M\varepsilon_{kl}\varepsilon_{ij} = 0$ для $k \neq i$, или $l \neq j$;

$M\varepsilon_{ij}^2 = \sigma^2$ – остаточная дисперсия.

Введём обозначения:

$y_j^* = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ – групповые средние (средние уровня A_j);

$y^{**} = \frac{1}{N}\sum_{j=1}^{m} n_j y_j^* = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{n_j} y_{ij}$ – общая средняя всех наблюдаемых значений результативного признака Y,

где $N = \sum_{j=1}^{m} n_j$;

$$Q_A = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_j^* - y^{**})^2 = \sum_{j=1}^{m} n_j (y_j^* - y^{**})^2 -$$

факторная сумма квадратов отклонений (сумма квадратов отклонений групповых средних от общей средней);

$Q_{ост} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - y_j^*)^2$ – остаточная сумма квадратов отклонений (сумма квадратов отклонений наблюдений от групповых средних);

$Q_{общ} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - y^{**})^2$ – общая сумма квадратов отклонений.

Осн. тождество *дисперсионного анализа*:

$Q_{общ} = Q_A + Q_{ост}$, которое показывает, что общая сумма квадратов отклонений результативного признака равна сумме квадратов отклонений групповых средних от общей средней и сумме квадратов отклонений наблюдений от групповых средних.

В дисперсионном анализе анализируются не сами суммы квадратов отклонений, а средние квадраты, которые получаются делением суммы квадратов отклонений на соответствующее число степеней свободы. Число степеней свободы в общем случае равно числу независимых наблюдений, уменьшенному на число параметров, оцениваемых по этим наблюдениям при вычислении статистики (см. табл. 1).

Таблица 1

Параметры однофакторного дисперсионного анализа

Вариация	Сумма квадратов	Число степеней свободы	Средний квадрат
Факторная	Q_A	$m-1$	$S_A^2 = \frac{Q_A}{m-1}$
Остаточная	$Q_{ост}$	$N-m$	$S_{ост}^2 = \frac{Q_{ост}}{N-m}$
Полная	$Q_{общ}$	$N-1$	$S_{общ}^2 = \frac{Q_{общ}}{N-1}$

Осн. гипотеза дисперсионного анализа состоит в утверждении, что уровни фактора A не влияют на изменение результативного признака Y. В случае модели М1 осн. гипотеза формулируется в виде:

$H_0 : \mu_j = 0, \quad j = 1, 2, ..., m,$

а в случае модели М2:

$H_0 : \sigma_\mu^2 = 0.$

Для проверки нулевой гипотезы вычисляется

$F_{набл} = \frac{Q_A/(m-1)}{Q_{ост}/(N-m)}$.

Если $F_{набл} \le F_{крит}(\alpha; m-1; N-m)$,

где $F_{крит}(\alpha; m-1; N-m)$

находится по таблицам F- распределения для заданного уровня значимости α, числа степеней свободы числителя $\nu_1 = m-1$ и числа степеней свободы знаменателя $\nu_2 = N-m$, то гипотеза не отвергается. Из этого следует, что влияние фактора A на результативный признак не доказано.

Если $F_{набл} > F_{крит}(\alpha; m-1; N-m)$,

то гипотеза отвергается с вероятностью ошибки, равной α, т.е. можно сделать вывод о том, что фактор A существенно (значимо) влияет на результативный признак Y. В этом случае для измерения степени влияния фактора A на результативный признак используют выборочный коэффициент детерминации

$R^2 = \frac{Q_A}{Q_{общ}}$,

который показывает какую долю выборочной дисперсии составляет дисперсия групповых средних, или, иначе говоря, какая доля общей выборочной дисперсии объясняется зависимостью от фактора A.

ОПРЕДЕЛЯЮЩАЯ ФУНКЦИЯ

Понятие «О.ф.» связано с понятием средней величины, представляющей собой обобщённую количественную характеристику признака в статистической совокупности в конкретных

условиях места и времени. Показатель в форме средней величины выражает типичные черты и дает обобщённую характеристику однотипных явлений по одному из варьирующих признаков. Он отражает уровень этого признака, отнесённый к единице совокупности. Сущность средней можно раскрыть через понятие её определяющего свойства, сформулированного А.Я. Боярским и О. Кизини: средняя, являясь обобщающей характеристикой всей статистической совокупности, должна ориентироваться на определенную величину, связанную со всеми единицами этой совокупности. Эту величину (среднюю) можно представить в виде функции:

$$f(x_1, x_2, ..., x_n) \quad (1),$$

где $x_1, x_2, ..., x_n$ – значения рассматриваемого признака в статистической совокупности. Так как данная величина (1) в большинстве случаев отражает реальную экономическую категорию, её называют определяющей функцией (показателем). Если в функции (1) все величины $x_1, x_2, ..., x_n$ заменить их средней величиной $\bar{x}$, то значение функции (1) должно остаться прежним:

$$\bar{x} \equiv f(x_1, x_2, ..., x_n) = f(\bar{x}, \bar{x}, ..., \bar{x}). \qquad (2)$$

Исходя из равенства (2) и определяется средняя: *средняя арифметическая; средняя гармоническая; средняя геометрическая; средняя квадратическая*, кубическая и т. д. Перечисленные средние (кроме средней геометрической) объединяются в общей формуле средней степенной (при различной величине *к*):

$$\bar{x} = \sqrt[k]{\frac{\sum_{i=1}^{m}(x_i^k f_i)}{\sum_{i=1}^{m} f_i}},$$

где $\bar{x}$ – средняя величина исследуемого явления; x_i – i-й вариант осредняемого признака (i = 1, 2, …, m); f_i – частота i-го варианта; m – число групп (число различных значений исследуемого признака).

Помимо степенных средних в статистической практике также используются средние структурные, среди которых наиболее распространены *мода* и медиана.

ОПТИМАЛЬНАЯ (БАЙЕСОВСКАЯ) ПРОЦЕДУРА КЛАССИФИКАЦИИ

процедура классификации многомерных наблюдений при известных законах распределения вероятностей классов, минимизирующая вероятность ошибочной классификации либо потери от ошибочной классификации среди всех возможных процедур классификации.

Классифицируемые k – мерные наблюдения интерпретируются как выборка из генеральной совокупности, описываемой смесью классов π_i, заданных одномодальными плотностями распределения $f_i(x)$ и априорными вероятностями p_i. Процедура классификации задает разбиение признакового пространства на непересекающиеся области R_i, внутри которых объект относится к классу i. Задача состоит в оптимальном выборе областей R_i.

Для случая двух классов вероятность ошибочной классификации определяется выражением

$$T(R, f) = p_1 \int_{R_2} f_1(x)dx + \int_{R_1} f_2(x)dx = p_1 + \int_{R_2} (p_2 f_2(x) - p_1 f_1(x))dx$$

Т.е. вероятность ошибочной классификации достигает минимума, если R_1 выбрана т.о., что подынтегральное выражение отрицательно во всех точках множества R_1. Отсюда вытекает правило классификации: отнести наблюдение *x* к классу π_1, если

$$f(x_1) / f(x_2) > p_2 / p_1,$$

иначе – к классу π_2. Вероятности ошибочной классификации равны соответственно

$$P_1 = \int_{R_2} f_1(x)dx, \; P_2 = \int_{R_1} f_2(x)dx.$$

В случае нескольких классов решающее правило, минимизирующее вероятность ошибочной классификации, имеет вид: отнести объект к классу π_i, если

$$p_i f_i(x) = \max_j p_j f_j(x).$$

При практической реализации априорные вероятности p_i и плотности распределения $f_i(x)$ заменяются их оценками, построенными на базе обучающих выборок. Если предполагается, что все классы описываются законом распределения вероятностей одного и того же параметрического семейства,

$$f_i(x) = f(x, \theta_i),$$

то получают случай *параметрического дискриминантного анализа*.

ОПТИМИЗАЦИОННЫЕ ФОРМУЛИРОВКИ СТАТИСТИЧЕСКИХ ЗАДАЧ

Осн. свойства и характеристики выборки, называемые эмпирическими (или выборочными), могут быть проанализированы и вычислены по имеющимся у исследователя выборочным статистическим данным $x_1, x_2,..., x_n$ (случайная выборка, состоящая из n независимых одномерных наблюдений), извлечённым из исследуемой ген. совокупности. Осн. свойства и характеристики ген. совокупности, называемые теоретическими, не известны исследователю. Назначение оптимизационных задач как раз в том и состоит, чтобы с их помощью получить как можно более точное представление о теоретических свойствах и характеристиках ген. совокупности по соответствующим свойствам и характеристикам выборочных данных. Оптимизационные формулировки статистических задач можно подразделить на две категории: оптимизационные задачи, основанные на использовании *метода макс. правдоподобия* (ММП). В соответствии с этим методом оценка $\hat{\theta}_{ii}$ неизвестного параметра θ по наблюдениям $x_1, x_2,..., x_n$ случайной величины ξ (подчиненной закону распределения $f_\xi(x, \theta)$, где f – плотность или вероятность $P\{\xi = x\}$) определяется из условия:

$$L(x_1, x_2,..., x_n; \hat{\theta}_{ii}) = \max L(x_1,..., x_n; \hat{\theta}),$$

где L – функция правдоподобия, определенная соотношением:

$$L(x_1, x_2,..., x_n; \theta) = f(x_1; \theta) f(x_2; \theta) ... f(x_n; \theta);$$

$\hat{\theta}$ – оценка неизвестного параметра θ.

Отсюда, формулировка оптимизационной задачи выглядит: оценка макс. правдоподобия $\hat{\theta}_{ii}$ параметра θ по независимым наблюдениям $x_1, x_2,..., x_n$ может быть представлена в виде:

$$\hat{\theta}_{ii} = \arg \max_{\hat{\theta}} \prod_{i=1}^{n} f(x_i; \hat{\theta}).$$

ММП может быть применён в тех случаях, когда с точностью до неизвестных значений параметров известен общий вид закона распределения вероятностей имеющихся выборочных данных. ММП используется, в частности, также для оценки параметров линейной регрессионной модели и её ошибки – классического случая зависимости двух переменных X и У по их выборкам. В регрессионном анализе в качестве исходной модели рассматривается линейная модель вида:

$$Y_i = b_0 + b_1 X_i + U_i, \; U_i \sim N(0, \sigma^2),$$

$$i = 1,. 2, ..., n,$$

где U_i – случайное слагаемое (ошибка модели);

$$U_i = Y_i - b_0 + b_1 X_i.$$

Функция плотности вероятностей для случайного слагаемого:

$$f(U_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - b_0 + b_1 X_i)^2}{2\sigma^2}}.$$

Для определения трёх неизвестных параметров регрессионной модели b_0, b_1 и σ^2 по данным некоторой случайной выборки может быть использована эффективная функция макс. правдоподобия:

$$L(b_0, b_1, \sigma^2) = \prod_{i=1}^{n} f(U_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - b_0 + b_1 X_i)^2}{2\sigma^2}}.$$

Если прологарифмировать эту функцию по натуральному основанию e, то необходимым условием для оценки параметров линейной регрессионной модели ММП является решение системы нормальных уравнений, получаемых приравниванием нулю частных производных полученной функции по рассматриваемым параметрам.

Оптимизационные задачи, основанные на использовании *метода наименьших квадратов* (МНК) и его модификациях. МНК используется для оценки параметров регрессионных моделей по выборкам независимых переменных $x_{1i}, x_{2i},..., x_{mi}$, i = 1,…, n; m – число независимых переменных, n – число элементов в выборке и выборочным значениям зависимой переменной y: $y_1, y_2,..., y_n$. В соответствии с этим методом минимизируется сумма квадратов отклонений уравнения регрессии в соответствующих точках от эмпирических данных для зависимой переменной (пример линейной регрессии – случай зависимости двух переменных):

$$S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 + b_1 X_i)^2 \to \underset{b_0, b_1}{Min},$$

или $U_i = Y_i - b_0 + b_1 X_i$,

$$S(b_0, b_1) = \sum_{i=1}^{n} U_i^2 \to \underset{b_0, b_1}{Min}.$$

Оценки b_0 и b_1 неизвестных параметров b_0 и b_1 ищутся из условия мин. положительно определённой квадратичной формы. Для этого приравниваются нулю частные производные $S(b_0, b_1)$ по b_0 и b_1. Затем из полученной системы линейных алгебраических уравнений находим оценки b_0 и b_1, а также оценку $\hat{\sigma}^2$ неизвестной дисперсии σ^2.

ОРДИНАЛЬНАЯ (ПОРЯДКОВАЯ) ШКАЛА

предполагает упорядочение объектов относительно какого-либо критерия или свойства. Эта шкала определяется установлением равенства объектов по отношению к какому-либо конкретному значению шкалы и определением отношения «больше – меньше» между объектами. О.ш. предназначена для отнесения объекта к одному из непересекающихся классов, упорядоченных по некоторому критерию так, что у одного из объектов измеряемое свойство выражено сильнее, чем у другого, но при этом нельзя определить, насколько сильнее. В О.ш. допустимыми являются все строго возрастающие преобразования. Примеры О.ш. – образование, социальное положение, тип поселения и т.п. При построении О.ш. классы нумеруются в порядке возрастания или убывания соответствующего признака. Арифметические операции над номерами классов не производятся. Частный случай О.ш. – ранговая шкала, которая применяется, когда некоторый признак не может быть измерен, но объекты могут быть упорядочены по какому-либо критерию, или, когда порядок объектов более важен, чем точный результат измерения, напр., рейтинг вузов по качеству образования. Ранговые шкалы используются также при изучении ценностных ориентаций, мотивов, предпочтений и т.п., где необходимо упорядочить предложенный список объектов по определённому критерию. Другой частный случай О.ш. – оценочная шкала, с помощью которой свойства объекта или отношение респондента к чему-либо оценивается исходя из определённого количества баллов. Оценочные шкалы часто рассматриваются как исключение из шкал порядка, т.к. предполагается, что между баллами на шкале существует примерно одинаковое расстояние. Это свойство позволяет во многих случаях рассматривать оценочные шкалы как квазиинтервальные и использовать их соответствующим образом, напр., определять среднюю успеваемость в классе. С помощью О.ш. можно измерять качественные, не имеющие строгой количественной меры, показатели. Особенно широко эти шкалы используются в гуманитарных науках: социологии, педагогике, психологии. О.ш. используется и во многих иных областях. В эконометрике, напр., это различные методы экспертных оценок.

К О.ш. относятся шкала Мооса (твёрдости минералов), шкала Бофорта оценки силы ветра (отсутствие ветра), сейсмическая шкала Рихтера (шкала магнитуд) и т.д.

ОРТОГОНАЛЬНЫЕ ПОЛИНОМЫ ЧЕБЫШЕВА

используются в прикладной статистике при построении кривых распределения случайной величины ξ по её выборочным наблюдениям. Идеи и методы теории аппроксимации, фундамент которой заложен классическими работами Чебышева, Вейерштрасса, Джексона и Бернштейна о приближении многочленами индивидуальных функций и целых их классов. О.п.Ч., иногда называют многочленами, наименее уклоняющимися от нуля.

Два полинома, заданные на интервале [a,b] являются ортогональными, если выполнено условие

$$\int_a^b p(x)q(x)w(x)dx = 0,$$

где w(x) − неотрицательная весовая функция.

Множество полиномов pn (x), n = 0, 1, 2,..., где n − степень полинома pn (x), образуют систему ортогональных полиномов, если справедливо равенство

$$\int_a^b p_m(x)\,p_n(x)\,w(x)\,dx = c_n\delta_{mn},$$

где cn − заданные константы, а δmn − символ Кронекера.

Различают О.п.Ч. первого и второго родов.

Полиномы Чебышева первого рода

$T_n(x) = \cos(n \arccos(x))$, $n = 0, 1, 2, \ldots$

ортогональны на отрезке [− 1, 1] с весовой функцией

$$h_1(x) = \frac{1}{\sqrt{1-x^2}}:$$

$$\int_{-1}^{1} \frac{T_m(x)T_n(x)}{\sqrt{1-x^2}}dx = \begin{cases} 0, & m \neq n \\ \pi, & m = n = 0 \\ \pi/2, & m = n \neq 0 \end{cases}.$$

Для полиномов Чебышева 1-го рода справедливо рекуррентное соотношение:

$T_{n+1}(x) = 2x\,T_n(x) - T_{n-1}(x)$, $n = 1, 2, \ldots$

при этом

$T_0(x) = 1$; $T_1(x) = x$; $T_2(x) = 2x^2 - 1$;

$T_3(x) = 4x^3 - 3x$; $T_4(x) = 8x^4 - 8x^2 + 1$.

Графики первых пяти полиномов Чебышева1-го рода представлены на рис. 1.

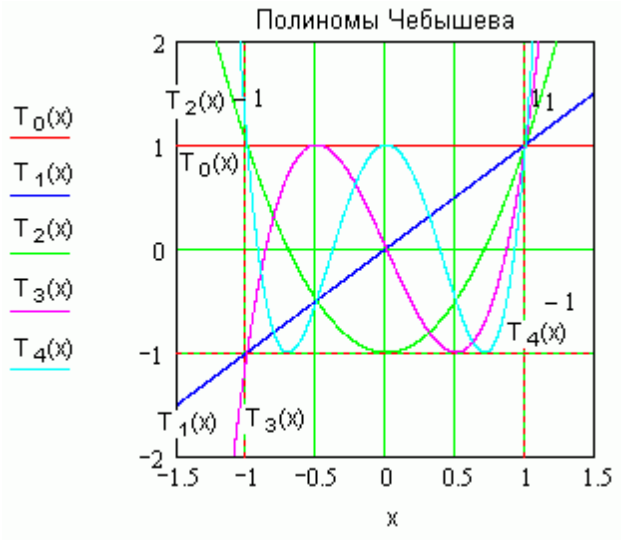


Рис. 1. **Полиномы Чебышева 1-го рода**

Полиномы Чебышева II-го рода

$$U_n(x) = \frac{\sin((n+1)\arccos(x))}{\sin(\arccos(x))},$$

$n = 0, 1, 2, \ldots$ ортогональны на отрезке [-1,1] с весовой функцией

$$h_2(x) = \sqrt{1-x^2}\ :$$

$$\int_{-1}^{1} U_n(x)U_m(x)\sqrt{1-x^2}\,dx = \begin{cases} 0, \text{ если } n \neq m \quad \text{или} \quad n = m = 0, \\ \dfrac{\pi}{2}, \text{ если} \quad n = m \neq 0. \end{cases}$$

Полиномы Чебышева II-го рода

$U_n(x)$

так же удовлетворяют рекуррентному соотношению:

$U_{n+1}(x) = 2x\,U_n(x) - U_{n-1}(x)\ ,\ n = 1, 2, \ldots$

при этом

$U_0(x) = 1;\ U_1(x) = 2x;\ U_2(x) = 4x^2 - 1;$

$U_3(x) = 8x^3 - 4x;\ U_4(x) = 16x^4 - 12x^2 + 1$

Многочлены Чебышёва являются решениями уравнения Пелля:

Tn(x)2 − (x2 − 1)Un − 1(x)2 = 1

в кольце многочленов с вещественными коэффициентами и удовлетворяют тождеству:

$T_n(x) + U_{n-1}(x)\sqrt{x^2-1} = (x + \sqrt{x^2-1})^n.$

Из последнего тождества также следуют явные формулы:

$$T_n(x) = \frac{(x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n}{2} = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n}{2k} (x^2-1)^k x^{n-2k};$$

$$U_n(x) = \frac{(x+\sqrt{x^2-1})^{n+1} - (x-\sqrt{x^2-1})^{n+1}}{2\sqrt{x^2-1}} = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n+1}{2k+1} (x^2-1)^k x^{n-2k}.$$

ОТБОР БЕСПОВТОРНЫЙ

Достоверность рассчитанных по выборочным данным характеристик в значительной степени определяется репрезентативностью (представительностью) выборочной совокупности, которая, в свою очередь, зависит от способа отбора единиц из ген. совокупности. В каждом конкретном случае в зависимости от целого ряда условий, а именно сущности исследуемого явления, объёма совокупности, вариации и распределения наблюдаемых признаков, материальных и трудовых ресурсов, выбирают наиболее предпочтительную систему организации отбора, которая определяется видом, методом и способом отбора. По виду различают индивидуальный, групповой и комбинированный отбор. При индивидуальном отборе в выборочную совокупность отбираются отдельные единицы ген. совокупности, при групповом отборе – группы единиц, а комбинированный отбор предполагает сочетание группового и индивидуального отбора. Метод отбора определяет возможность продолжения участия отобранной единицы в процедуре отбора.

Бесповторным называется такой метод отбора, при котором попавшая в *выборку* единица не возвращается в совокупность, из которой осуществляется дальнейший отбор. А при повторном отборе попавшая в выборку единица после регистрации наблюдаемых признаков возвращается в исходную (ген.) совокупность для участия в дальнейшей процедуре отбора.

Способ отбора определяет конкретный механизм или процедуру выборки единиц из ген. совокупности. В практике выборочных обследований наибольшее распространение получили следующие выборки: собственно-случайная; механическая; типическая; серийная; комбинированная. Для каждого из этих способов отбора при бесповторных выборках средняя ошибка выборки μ, выражающая среднее квадратическое отклонение выборочной средней от ген.

средней, вычисляется по разному. Напр., при собственно-случайной выборке (отбор единиц из ген. совокупности наугад или наудачу, без каких-либо элементов системности) средняя ошибка бесповторной выборки вычисляется по формуле:

$$\mu = \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)},$$

а при типической выборке (применяется в случаях, когда ген.совокупность каким-либо образом можно разбить на несколько типических групп) при пропорциональном объёму групп отборе – по формуле:

$$\mu = \sqrt{\frac{\overline{\sigma}^2}{n}\left(1-\frac{n}{N}\right)},$$

где n – объём выборочной совокупности; N – объём ген. совокупности; σ^2 – ген. дисперсия; $\overline{\sigma}^2$ – средняя из внутригрупповых дисперсий.

ОТБОР ИНФОРМАТИВНЫХ ТИПООБРАЗУЮЩИХ ПРИЗНАКОВ

Подобный отбор продиктован стремлением исследователя снизить размерность исследуемого признакового пространства с целью лаконичного объяснения природы анализируемых многомерных данных. Возможность лаконичного описания анализируемых многомерных данных основана на априорном допущении, в соответствии с которым существует небольшое (в сравнении с числом p исходных анализируемых признаков $x^{(1)}, x^{(2)}, ..., x^{(p)}$ число p' признаков-детерминант (*гл. компонент*, общих факторов, наиболее информативных объясняющих переменных), с помощью которых могут быть достаточно точно описаны как сами наблюдаемые переменные анализируемых объектов, так и определяемые этими переменными свойства (характеристики) анализируемой совокупности. При этом упомянутые признаки-детерминанты могут находиться среди исходных признаков, а могут быть латентными, т.е. непосредственно статистически не наблюдаемыми, но восстанавливаемыми по исходным данным. Гениальный пример практической реализации этой идеи даёт нам периодическая система элементов Менделеева: в этом случае роль идеально информативного единственного признака-детерминанта играет, как известно, заряд атомного ядра.

Отбор наиболее информативных признаков (включая выявление латентных факторов) – отбор из исходного (априорного) множества признаков

$$X = (x^{(1)}, x^{(2)}, ..., x^{(p)})^T$$

или о построении в качестве некоторых комбинаций исходных признаков относительно небольшого числа p' переменных

$$Z(X) = (z^{(1)}(X), z^{(2)}(X), ..., z^{(p')}(X))^T,$$

которые обладали бы свойством наибольшей информативности в смысле, определённом, как правило, некоторым специально подобранным для каждого конкретного типа задач критерием информативности $I_{p'}(Z)$. Так, напр., если критерий $I_{p'}(Z)$ «настроен» на достижение макс. точности регрессионного прогноза некоторого результирующего количественного показателя y по известным значениям предикторных (объясняющих, экзогенных) переменных $x^{(1)}, x^{(2)}, ..., x^{(p)}$, то речь идёт о наилучшем подборе наиболее существенных предикторов в модели регрессии. Если же критерий $I_{p'}(Z)$ устроен т.о., что его оптимизация обеспечивает наивысшую точность решения задачи отнесения объекта к одному из классов по значениям X его описательных признаков, то речь идет о построении (отборе) системы наиболее информативных типообразующих признаков в задачах классификации (расщепление смесей вероятностных распределений; методы кластер-анализа) или о выявлении и интерпретации некоторой сводной (латентной) характеристики изучаемого свойства. Наконец, критерий $I_{p'}(Z)$ может быть нацелен на макс. автоинформативность новой системы показателей X, т.е. на макс. точное воспроизведение всех исходных признаков $x^{(1)}, x^{(2)}, ..., x^{(p)}$ по сравнительно небольшому числу вспомогательных переменных $z^{(1)}, z^{(2)}, ..., z^{(p')}$ $(p' << p)$. В этом случае говорят о наилучшем автопрогнозе и обращаются к моделям и методам *факторного анализа* и его разновидностей (*метод гл. компонент*, *факторный анализ*).

ОТНОСИТЕЛЬНАЯ ЧАСТОТА (ЧАСТОСТЬ)

$p_i^{(n)}$ – величина, которая определяется как отношение числа ν_{x_i} наблюдений в выборке $x_1, x_2, ..., x_n$,

в точности равных x_i, к общему объёму выборки n, т.е.

$$p_i^{(n)} = \frac{\nu_{x_i}}{n}.$$

В случае «группированного», интервального *вариационного ряда*, имеющего к- интервалов, v_{xi} – число наблюдений в i-м интервале и

$$\sum_{i=1}^{k} \nu_{xi} = n.$$

ОТНОШЕНИЕ ПРАВДОПОДОБИЯ

представление о сравнительной правдоподобности имеющихся наблюдений $x_1, x_2,..., x_n$ в отношении проверяемой и *альтернативной гипотез*. Сопоставление соответствующих функций правдоподобия и, в частности, их отношение, называемое в статистике О.п.:

$$\gamma^{(n)} = \frac{L_{H_1}(x_1, x_2,..., x_n; \theta)}{L_{H_0}(x_1, x_2,..., x_n; \theta)} = \frac{L(x_1, x_2,..., x_n; \theta_1)}{L(x_1, x_2,..., x_n; \theta_0)},$$

где L_{H_1} и L_{H_0} – значения функций правдоподобия наблюдений $x_1, x_2,..., x_n$, вычисленные в предположении справедливости соответственно гипотез $H_1 : \theta = \theta_1$ и $H_0 : \theta = \theta_0$. Очевидно, чем правдоподобнее наблюдения в условиях гипотезы H_0, тем больше функция правдоподобия L_{H_0} и тем меньше величина $\gamma^{(n)}$.

Функция правдоподобия в зависимости от постановки задач и целей исследования может рассматриваться либо как функция параметра θ (при заданных фиксированных наблюдениях $x_1^*, x_2^*,..., x_n^*$, либо как функция текущих значений наблюдений $x_1, x_2,..., x_n$ (при заданном фиксированном значении параметра θ), либо как функция обеих переменных X и θ.

ОЦЕНКА ДИСПЕРСИИ

оценка ген. дисперсии σ^2, полученная по результатам выборочных наблюдений $x_1, x_2,..., x_n$. Различают точечные и интервальные О.д. В табл. 1 приведены точечные О.д. ген. совокупности.

Таблица 1

Оцениваемый параметр ген. совокупности	По простой выборке $x_1, x_2,..., x_n$	По сгруппированным данным $x_1, x_2,..., x_L$; m_i – частота встречаемости признака x_i; $n = \sum_{i=1}^{L} m_i$ – объём выборки
Ген. дисперсия σ^2 (математическое ожидание μ известно)	Выборочная дисперсия S^2	
	$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$	$S^2 = \frac{1}{n}\sum_{i=1}^{L}(x_i - \mu)^2 m_i$
Ген.дисперсия σ^2 (математическое ожидание μ неизвестно)	$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ или $S^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2$, где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$	$S^2 = \frac{1}{n}\sum_{i=1}^{L}(x_i - \bar{x})^2 m_i$ или $S^2 = \frac{1}{n}\sum_{i=1}^{L}x_i^2 m_i - (\bar{x})^2$, где $\bar{x} = \frac{1}{n}\sum_{i=1}^{L}x_i m_i$
	Исправленная (несмещённая) выборочная дисперсия $\hat{S}^2$	
	$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	

О.д. выборочной

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

является смещённой. Если *математическое ожидание* неизвестно, то несмещённая О.д. – исправленная выборочная дисперсия

$$\hat{S}^2 = \frac{n}{n-1}S^2$$

(дробь $\frac{n}{n-1}$ называют поправкой выборочной Бесселя).

Оценка интервальная (*доверительный интервал*) ген. дисперсии σ^2 ген. совокупности X, имеющей *нормальный закон распределения*, с заданной надежностью γ при малых объёмах выборки (n<30) осуществляется по формуле:

$$P\left(\frac{nS^2}{\chi_2^2} \le \sigma^2 \le \frac{nS^2}{\chi_1^2}\right) = \gamma.$$

Значения χ_1^2 и χ_2^2 находят по табл. χ^2 – распределения Пирсона для условий:

$$P(\chi_1^2) = P(\chi^2 > \chi_1^2) = 1 - \frac{\alpha}{2} = \frac{1+\gamma}{2},$$

$$P(\chi_2^2) = P(\chi^2 > \chi_2^2) = \frac{\alpha}{2} = \frac{1-\gamma}{2},$$

где χ^2 – случайная величина, имеющая χ^2 – распределение с $\nu = n-1$ степенями свободы; $\alpha = 1-\gamma$ – уровень значимости.

А интервальная оценка (доверительный интервал) ген. дисперсии σ^2 ген. совокупности X, имеющей нормальный закон распределения, с заданной надежностью γ при больших объёмах выборки (n>30) осуществляется по формуле:

$$P\left(\frac{2nS^2}{(\sqrt{2n-3}+t_\gamma)^2} \le \sigma^2 \le \frac{2nS^2}{(\sqrt{2n-3}-t_\gamma)^2}\right) = \Phi(t_\gamma) = \gamma,$$

где t_γ значение нормированной нормальной случайной величины, соответствующее заданной надежности γ: $t_\gamma = \Phi^{-1}(\gamma)$ – определяется по таблице интегральной функции Лапласа $\Phi(t)$ (Φ^{-1} – обратное преобразование).

ОЦЕНКА ДОСТАТОЧНАЯ

оценка θ_n^* параметра θ называется достаточной, если условное распределение $P(x_1,x_2,\ldots,x_n/\theta_n^*)$, где значение статистики θ_n^*, не зависит от неизвестного параметра θ для всех возможных значений θ_n^*. Достаточность связана с объёмом информации, содержащимся в выборке, необходимым для выработки решения относительно параметра θ ген. совокупности.

Достаточность статистики на практике обычно проверяют с помощью критерия факторизации. Согласно критерию оценка будет достаточной тогда и только тогда, когда функция правдоподобия $L(x_1,x_2,\ldots,x_n/\theta)$ может быть представлена в виде произведения двух множителей, первый из которых зависит от параметра θ и статистики θ_n^*, а второй зависит только от результатов наблюдений $x_1,x_2,\ldots,x_n$ и не зависит от θ.

$L(x_1,x_2,\ldots,x_n/\theta)=G(\theta_1\theta_n^*)H_1(x_1,x_2,\ldots,x_n)$ (1)

Свойства достаточной статистики: 1) достаточные статистики инвариантны относительно преобразования параметра. Поэтому если θ_n^* – достаточная статистика параметра θ, то статистика $\varphi(\theta_n^*)$ также будет достаточной статистикой параметра $\varphi(\theta)$, где φ – рациональное преобразование. Для нормального закона $\bar{x}$ и S^2 являются достаточными статистиками, из чего следует, что достаточным будут и оценки $s, \hat{s}^2, \hat{s}$; 2) эффективная оценка является и О.д. Напр., пусть величина X распределена по закону Пуассона, тогда

$$P(x) \quad \frac{e}{x!} \quad .$$

Докажем, что

$$n\bar{x} = \sum_{i=1}^{n} x_i$$

является достаточной статистикой, где

$$x_i = \sum_{j=1}^{N} \xi ij$$

число появления события A в i-й серии испытаний; $\xi ij = I$, если событие A наблюдалось в испытании, и $\xi ij = 0$, если не наблюдалось

(i=1,2,…,n; j=1,2,…,N). Составим функцию правдоподобия:

$$L(x_1, x_2, \ldots, x_n/\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}}{G} \bullet \frac{1}{\prod_{i=1}^{n} (x_i!)} = G(\lambda; \sum_{i=1}^{n} x_i) H(x_1, x_2, \ldots, x_n)$$

Согласно (1) выборочная характеристика

$$n\bar{x} = \sum_{i=1}^{n} x_i$$

является достаточной статистикой параметра λ(λ>0).

Статистики, полученные на основании критерия факторизации, являются миним. достаточными оценками. Для получения таких оценок используется минимально возможная часть информации, содержащейся в выборке. При дальнейшем сокращении используемой информации свойство достаточности не сохраняется.

ОЦЕНКА ИНТЕРВАЛЬНАЯ

оценка с заданной вероятностью неизвестного параметра *ген. совокупности* (т.е. истинного значения) в виде интервала

$[\hat{\theta}_{\min}, \hat{\theta}_{\max}]$,

границами которого являются *оценки точечные*. Границы интервала рассчитываются по определённым правилам, исходя из результатов выборки и являются функциями выборки и, следовательно, случайными величинами. О.и. позволяет определить границы интервала, покрывающего с *доверительной* (желаемой) *вероятностью* (1-α) неизвестное значение параметра θ. Это неизвестное значение называется доверительным, или *доверительной областью*, а его границы – *доверительными границами* (уровнем значимости, или вероятностью ошибки). Уровень доверительной вероятности указывается на то, что при повторении оценивания в среднем в (1-α) × 100% случаев параметр действительно покрывается интервалом. Ширина *доверительного интервала* наглядно показывает точность оценки: чем больше ширина доверительного интервала, тем больше точность оценивания. В зависимости постановки задачи интервал может быть двусторонним или односторонним. При вычислении О.и. обязательно учитывается их распределение и объём *выборки*. При небольших объёмах выборки (n≤ 60) и заданной доверительной вероятности доверительные границы рассчитываются по таблице t-*распределения Стьюдента* с n-1 степенями свободы. О.и. удовлетворяет требованиям: оценка должна быть достаточно точной. Точность О.и. определяется как

$$\Delta_\theta = \frac{\hat{\theta}_{\max} - \hat{\theta}_{\min}}{2},$$

т.о., длина интервала должна быть достаточно малой; оценка должна быть достаточно надёжной. Надёжностью О.и. называют вероятность

$$\gamma = P\left\{\hat{\theta}_{\min} \le \theta \le \hat{\theta}_{\max}\right\}$$

накрытия доверительным интервалом неизвестного параметра θ. Такая доверительная вероятность должна быть близкой к единице, чтобы считать событие

$$\left\{\hat{\theta}_{\min} \le \theta \le \hat{\theta}_{\max}\right\}$$

практически достоверным.

ОЦЕНКА НЕСМЕЩЁННАЯ

оценка θ_n^* параметра θ, если её *математическое ожидание* равно оцениваемому параметру θ, т.е.

$$M\left(\theta_n^*\right) = \theta$$

при всех допустимых значениях θ. В противном случае, говорят, что оценка смещённая, а её смещением является разность

$$B_n = M\left(\theta_n^*\right) - \theta.$$

Напр., $\bar{x}$ является О.н. μ в ген. совокупности, если μ существует. Пусть из ген. совокупности X извлечена выборка $x_1, x_2, ..., x_n$. Следовательно,

$$M(x_i)=\mu, D(x_i)=\sigma^2,\ i=\overline{1,n}.\ \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\ ;$$

$$M(\bar{x})=M\left[\frac{1}{n}\sum_{i=1}^{n}x_i\right]=\frac{1}{n}\sum_{i=1}^{n}\underbrace{M(x_i)}_{\mu}=\frac{1}{n}n\mu=\mu,$$

$$M(\bar{x})=\mu$$

т.е. $\bar{x}$ является О.н. μ. Рассмотрим *выборочную дисперсию*

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Тогда, $MS^2=\frac{1}{n}\sum_{i=1}^{n}M\left[(x_i-\mu)-(\bar{x}-\mu)\right]^2=$

$$=\frac{1}{n}\sum_{i=1}^{n}\underbrace{M(x_i-\mu)^2}_{D(x_i)=\sigma^2}+\frac{1}{n}\sum_{i=1}^{n}M(\bar{x}-\mu)^2-\frac{2}{n}M\sum_{i=1}^{n}(x_i-\mu)(\bar{x}-\mu);$$

$$-\frac{2}{n}M\sum_{i=1}^{n}\{(x_i-\mu)(\bar{x}-\mu)\}=-\frac{2}{n}M(\bar{x}-\mu)\underbrace{\sum_{i=1}^{n}(x_i-\mu)}_{n(\bar{X}-\mu)}=-\frac{2}{n}nD(\bar{x})=-2D(\bar{x});$$

$$D(\bar{X})=D\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n};$$

$$MS^2=\frac{n}{n}\sigma^2+\frac{1}{n}nD(\bar{x})-2D(\bar{x})=\sigma^2-\frac{\sigma^2}{n}\neq\sigma^2$$

Значит, S^2 – смещённая оценка σ^2 ген. совокупности. Величина смещения

$$B_n=MS^2-\sigma^2=-\frac{1}{n}\sigma^2\to 0$$

при $n\to\infty$. Обозначим

$$\hat{S}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

где $(n-1)$ – число степеней свободы равно числу наблюдений за вычетом условий, накладываемых на наблюдения.

$$\hat{S}^2=\frac{n}{n-1}S^2.$$

$$M\hat{S}^2=\frac{n}{n-1}MS^2=\frac{n}{n-1}\frac{\sigma^2(n-1)}{n}=\sigma^2,$$

т.е. $\hat{S}^2$ – О.н. σ^2 в ген. совокупности.

ОЦЕНКА РОБАСТНАЯ

робастность оценки означает её устойчивость к наличию резко выделяющихся значений ("выбросов") или к нарушению предположений, ограничивающих применение соответствующего статистического метода. Наиболее известны – устойчивые оценки Пуанкаре, Винзора и Хубера, которые дают хорошие результаты в случае, если и осн., и засоряющее распределения симметричны и уровень засорения ε известен. Для отбраковки выделяющихся значений усекают выборку, отбрасывая определённую часть миним. и/или макс. наблюдений (отбрасывается αN наименьших и αN наибольших значений выборки), и по оставшейся части оценивают параметры распределения, т.н. α-урезанные асимптотически нормальные оценки Пуанкаре. По другому подходу перед процедурой оценивания винзорируют *выборку*: всем наблюдениям левее и/или правее определённых значений присваивают одинаковые значения, вычисленные оценки параметров распределения – оценки Винзора так же являются асимптотически нормальными. В экономических исследованиях наиболее широко применяется метод Хубера, который позволяет получить оценки с наименьшим средним квадратом смещения при наихудшем засорении. Оценки Хубера являются состоятельными, обладают высокой эффективностью

и достаточной робастностью. При асимметричном распределении оценки Пуанкаре, Винзора и Хубера теряют эффективность и становятся несостоятельными и смещёнными. При несимметричном распределении применяют джекнайф-оценку Тьюки и Квенсула, суть которой состоит в разбиении на группы исходной совокупности и оценке эффекта каждой группы по результату, полученному при исключении данной группы из рассмотрения, что позволяет уменьшить смещение параметра положения при асимметричных распределениях. Более эффективна – взвешенная джекнайф-оценка Хинкли, где весовые коэффициенты выбираются как расстояния, отражающие недостаток симметрии и отражают вклад каждого наблюдения в дисперсию показателя. При оценке параметров вклад наблюдений с наибольшими весами уменьшается.

См. также *Оценивание робастное.*

ОЦЕНКА СОСТОЯТЕЛЬНАЯ

оценка θ_n^* называется О.с. θ, если при $n \to \infty$ она сходится по вероятности к оцениваемому параметру θ, т.е. если $\theta_n^* \xrightarrow{P} \theta$, где n – объём выборки, $n \to \infty$ (или $\lim_{n\to\infty} P\left\{\left|\theta_n^* - \theta\right| < \varepsilon\right\} = 1$). Иногда такие оценки называют слабо состоятельными в отличие от сильно состоятельных оценок, для которых соответствующая сходимость имеет место с вероятностью 1. Так, средняя арифметическая $\overline{X}$ является О.с. *математического ожидания* $M(X) = \mu$ совокупности, так как, согласно *закону больших чисел*, $\overline{X} \xrightarrow[n\to\infty]{P} M(X)$, если существует $M(X)$.

См. также *Состоятельность оценки.*

ОЦЕНКА СТАТИСТИЧЕСКАЯ

функция результатов наблюдений, вычисляемая на основании выборочных данных (выборки) для приближённой замены неизвестного параметра распределения или самого распределения. Напр., если X1,..., Xn – независимые случайные величины, имеющие одно и то же нормальное распределение с неизвестным средним значением μ, то функция – средняя арифметическая результатов наблюдений,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

есть О.с. для неизвестного параметра μ. В качестве О.с. какого-либо параметра θ естественно выбрать функцию θ*(X1,..., Xn) от результатов наблюдений X1,..., Xn, в некотором смысле близкую к истинному значению параметра. Принимая какую-либо *меру "близости"* О.с. к значению оцениваемого параметра, можно сравнивать различные оценки по качеству. Обычно мерой близости оценки к истинному значению параметра служит величина среднего значения квадрата ошибки $E_\theta(\theta^* - \theta)^2 = D_\theta\theta^* + (\theta - E_\theta\theta^*)^2$ (выражающаяся через *математическое ожидание* оценки $E_\theta\theta^*$ и её *дисперсию* $D_\theta\theta^*$).

Из ген. совокупности можно сделать много разных выборок, причём значение О.с. в общем случае будет меняться от выборки к выборке; т.е., выборка является случайной, а значит, случайной величиной является и О.с. Напр., выборочные средние для разных выборок из одной и той же совокупности могут различаться между собой. О.с. обычно обозначают латинскими буквами, а оцениваемые ими параметры – греческими. Различают *точечные оценки* и *оценки интервальные* параметров. При анализе свойств оценок необходимо знание их законов распределения.

ОЦЕНКА СТАТИСТИЧЕСКАЯ

числовая характеристика параметра *ген. совокупности*, полученная на основе *выборки,* позволяющая лаконично описать интересующие свойства исследуемой совокупности путём представления множества обрабатываемых исходных данных в виде небольшого числа сводных характеристик, построенных на основании этих данных. Эти характеристики являются функциями от исходных результатов наблюдения $X_1, X_2, ..., X_n$ и называются статистиками. К ним относятся все выборочные *(эмпирические)* характеристики ген. совокупности: средние значения, дисперсии, коэффициенты эксцесса и асимметрии, ковариации и корреляции,

наконец, эмпирическая функция распределения и эмпирическая плотность. Формально: статистика $\hat{\Theta}$, используемая в качестве приближенного значения неизвестного параметра Θ называется статистической оценкой. Все статистики и О.с. – *случайные величины*: при переходе от одной выборки к другой (даже в рамках одной ген. совокупности) конкретные значения О.с., посчитанные по одной и той же формуле, будут подвержены некоторому неконтролируемому разбросу. Однако значения О.с., подсчитанные по разным выборкам, хотя и подвержены случайному разбросу, должны концентрироваться около истинного значения оцениваемого параметра. Для того чтобы О.с. были надёжными, они должны удовлетворять свойствам: состоятельности (т.е. стремится к Θ с ростом n), несмещённости (т.е. совпадать с Θ в среднем) и эффективности (т.е. обладать наименьшей степенью случайных отклонений от Θ). К принципам, с помощью которых формируется отношение к статистикам, претендующим на роль оценок параметров ген. совокупности, относится принцип размерности. Он состоит в том, что когда Θ не является безразмерной величиной, но обладает физической размерностью, такой, как время или длина, оценка $\hat{\Theta}$ должна иметь такую же физическую размерность, что и Θ. Второй принцип – принцип заменяемости: если оценка $\hat{\Theta}$ базируется на случайной выборке $(x_1, x_2, ..., x_n)$ равноточных наблюдений заданной *СВ X,* то порядок, в котором идут наблюдения, несущественен; оценка должна быть симметрической функцией наблюдений. Примером служат широко известные статистики $\bar{x} = \sum_{i=1}^{n} x_i$ и $s^2 = \sum (x_i - \bar{x})^2 / (n-1)$. Р.Фишер обнаружил, что в некоторых случаях можно собрать в единственной статистике всю информацию, содержащуюся в выборке относительно оцениваемых параметров (пользуясь словом «информация» в бытовом смысле). Такая статистика называется достаточной оценкой данного параметра. Вычисление на основании имеющихся выборочных данных оценки $\hat{\Theta}(x_1, x_2, ..., x_n)$ параметра Θ позволяет получить лишь приближённую *точечную оценку* $\hat{\Theta}$ даже в том случае, когда эта оценка состоятельна, несмещенна и эффективна. При этом остается проблема построения подходящей меры точности этой оценки, которая отвечает на вопрос о том насколько сильно может отличаться приближенное значение (оценка) от истинного. Для этого выбирается величина Δ, которая с определённой, заранее заданной вероятностью, близкой к единице, гарантирует выполнение неравенства $|\hat{\Theta} - \Theta| < \Delta$. Выбираемая исследователем вероятность, близкая к единице, называется *доверительной вероятностью*, а интервал вида $(\hat{\Theta}_1, \hat{\Theta}_2)$, который с заранее заданной вероятностью накрывает истинное значение параметра Θ. – *доверительным интервалом* или *оценкой интервальной* в отличие от точечных оценок $\hat{\Theta}$. Доверительный интервал по своей природе случаен, как по своему расположению, так и по величине. Ширина доверительного интервала существенно зависит от объёма выборки n (уменьшается с ростом n) и от величины доверительной вероятности (увеличивается с приближением доверительной вероятности к единице).

См. также *Оценка достаточная*, *Оценка несмещённая*, *Оценка состоятельная*, *Оценка эффективная*.

ОЦЕНКА ТОЧЕЧНАЯ

см. в ст. Точечная оценка

ОЦЕНКА ЭФФЕКТИВНАЯ

оценка несмещённая θ_n^* для θ называется эффективной, если её дисперсия

$$D(\theta_n^*) = M(\theta_n^* - \theta)^2$$

является наименьшей среди дисперсий всех возможных несмещённых оценок параметра θ, вычисленных по одному и тому же объёму выборки n. Так, оценка $\bar{x}$ является эффективной для *математического ожидания*, причём

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}.$$

Дисперсия любой несмещённой оценки одного параметра θ удовлетворяет неравенству Рао-Крамера

$$D\left(\theta_n^*\right) \geq -\frac{1}{NM\left(\frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2}\right)},$$

где $f(x;\theta)$ – плотность распределения вероятностей *CBX*; N – число произведённых испытаний. В случае дискретной CBX плотность $f(x;\theta)$ заменяется рядом распределения вероятностей $P(X = x;\theta)$. Несмещённая оценка $\theta_n = n(x_1, x_2, ..., x_n)$, для которой в неравенстве Рао-Крамера достигается знак равенства, называется эффективной. Из определения 2 следует определение 1, обратное утверждение, вообще говоря, не имеет места. В математической статистике применяются также асимптотически эффективные оценки, дисперсия которых стремится к нижней границе неравенства Рао-Крамера при неограниченном увеличении объёма выборки, т.е. при $n \to \infty$. Эффективность оценки θ_n^* определяют отношением

$$e = \frac{\sigma^2_{\theta_n^{*(э)}}}{\sigma^2_{\theta_n^*}},$$ где $\sigma^2_{\theta_n^{*(э)}}$ и $\sigma^2_{\theta_n^*}$ – соответственно дисперсии эффективной и данной оценок. Чем ближе e к 1, тем эффективнее оценка. Если $e \to 1$ при $n \to \infty$, то такая оценка называется асимптотически эффективной.

ОЦЕНИВАНИЕ РОБАСТНОЕ

При исследовании статистических совокупностей часто приходится иметь дело с данными, отклоняющимися от основного массива, т.е. с ошибками, или выбросами. Пример: на десяти пр-тиях отрасли легкой пром. произведены контрольные расчёты уровня рентабельности произ-ва по итогам работы в первом полугодии и получены результаты (см. табл. 1):

Таблица 1

Предприятие	1	2	3	4	5	6	7	8	9	10
Уровень рентабельности продукции, %	15.4	13.2	18.3	47.1	12.0	16.3	65.2	17.4	11.0	12.9

В приведённых данных имеются два значения: 47,1 и 65,2, которые значительно больше всех других значений, покрываемых интервалом [11,0; 18,3]. При выявлении подобных «выбросов» возникают вопросы: являются ли отклоняющиеся данные действительно ошибками (напр., регистрации) или это реальные значения и как в этом случае получить адекватные оценки для параметров изучаемой совокупности. Решением подобных вопросов занимается специальный раздел статистики – О.р. (устойчивое), занимающееся разработкой и использованием таких статистических методов, которые позволяют получать достаточно надежные оценки параметров распределения статистической совокупности, нечувствительные к структуре данных, с учётом неявности закона ее распределения и наличия существенных отклонений в значениях выборочных данных. Такие процедуры оценивания параметров называют робастными.

При решении задач О.р. выделяют два типа данных, засоряющих статистическую совокупность. К первому типу относят данные, несущественно отличающиеся от значений, которые наиболее часто встречаются в изучаемой совокупности. Эти данные не вызывают значительных искажений в аналитических результатах и могут обрабатываться обычными методами статистического оценивания. Второй тип данных – резко выделяющиеся на фоне изучаемой совокупности, их называют «засорением» или «грубыми ошибками», они оказывают сильное искажающее воздействие на аналитические результаты. Эти данные должны подвергаться специальной обработке.

При обработке «грубых» ошибок (засорений) можно выделить два осн. подхода. Первый ориентирован на устранение из выборочной совокупности ошибок и оценку параметров по оставшимся «истинным» значениям. Второй

подход предполагает в каждом случае с грубой ошибкой выделение истинных значений признака и собственно ошибки

$x = x_{ист} + \xi$;

при этом осуществляется модификация данных т.о., чтобы искажающий элемент ξ получил нормальное распределение с нулевым *математическим ожиданием*. Тогда для некоторого множества грубых ошибок вариативной величины x сумма

$\sum \xi$

приближается к нулю, а оценки $\hat{x}$ – к истинным значениям параметров выборочной совокупности. Алгоритм обработки «засорений» включает последовательное выполнение шагов: распознавание ошибок в данных; выбор метода и проведение О.р. данных; критериальная или логическая проверка и интерпретация результатов устойчивого оценивания.

Выявление грубых ошибок и оценка степени засорения выборки возможны при визуальном анализе данных или проверке статистической гипотезы на наличие ошибки. Во втором случае предусматривается расчёт специальных статистических критериев.

Простой формальный приём для обнаружения грубых ошибок основывается на расчёте Т-критерия Граббса для выборки

$$x_1, x_2, ..., x_n : T_{наблi} = \frac{x_i - \bar{x}}{s},\ i = 1, 2, ..., n,$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

– выборочная средняя;

$$s = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

– выборочное среднеквадратическое отклонение случайной величины.

Наблюденные значения $T_{наблi}$ Т-критерия сравнивают с пороговым $T_{кр}(\alpha, n)$, заданным соответствующим распределением, определяемым по таблице процентных точек критерия Смирнова – Граббса. Параметр α – задаваемый уровень значимости. Проверяемое признаковое значение x_i относят к классу выбросов, если

$T_{наблi} > T_{кр}(\alpha, n)$,

в противном случае считается, что это значение несущественно отличается от других данных и не будет давать сильного искажающего эффекта.

Более точными по сравнению со статистикой Граббса оценками грубых ошибок признаются *L*- и *E*-критерии, предложенные американскими статистиками Г. Титьеном и Г Муром: 1. L-критерий исчисляется для выявления грубых ошибок в верхней части ранжированного ряда выборочных данных $x_1, x_2, ..., x_n$:

$$L = \sum_{i=1}^{n-k} (x_i - \bar{x}_k)^2 \Big/ \sum_{i=1}^{n-k} (x_i - \bar{x})^2,$$

где *к* – число наблюдений с резко отклоняющимися значениями признака;

$$\bar{x}_k = \sum_{i=1}^{n-k} x_i \Big/ (n-k)$$

– средняя, которую рассчитывают по $n-k$ наблюдениям, остающимися после отбрасывания *к* грубых ошибок«сверху» ранжированного ряда данных; 2. L'- критерий применяется для выявления грубых ошибок в данных, расположенных в нижней части ранжированного ряда данных:

$$L' = \sum_{i=k+1}^{n} (x_i - \bar{x}_k)^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

где $\bar{x}_k = \sum_{i=k+1}^{n-k} x_i \Big/ (n-k)$

– средняя, рассчитанная по $n-k$ наблюдениям, остающимся после отбрасывания *к* грубых ошибок «снизу» ранжированного ряда; 3. E-критерий используется, когда в выборке имеются предположительно грубые ошибки с наибольшими и наименьшими значениями, т.е. расположенные в верхней и нижней частях ранжированного ряда данных:

$$E = \sum_{i=k+1}^{n-k'} (x_i - \bar{x}_{k'})^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

где $\bar{x}_k = \sum_{i=k+1}^{n-k'} x_i \Big/ (n - (k + k'))$

– средняя, рассчитанная по «истинным» данным после отбрасывания из выборки наимень-

ших (k) и наибольших (k') значений, засоряющих совокупность данных.

Все три критерия L, L' и E имеют табулированные критические значения для заданного уровня значимости α при известном объёме выборки *n* и предполагаемом числе ошибок *к*. Если наблюденные значения критериев оказываются меньше пороговых $C_{\alpha,k}$, то ошибки в данных, подвергаемых проверке, признаются грубыми, существенно отклоняющимися от основного массива данных. При L, L',

E> $C_{\alpha,k}$

данные гипотетически предполагаются типичными для изучаемой выборочной совокупности. После обнаружения выбросов в данных решается задача оценивания параметров выборочной совокупности. При этом используются два осн. подхода: экстремальные значения (грубые ошибки) отбрасываются либо модифицируются.

Наиболее простыми представляются оценки по усечённой совокупности данных, остающейся после отбрасывания грубых ошибок. Американский статистик Пуанкаре предложил следующую формулу для расчёта устойчивой (робастной) статистической оценки средней по усечённой совокупности (урезанную среднюю):

$$T(\alpha) = \sum_{i=k+1}^{n-k'} x_i \Big/ (n-2k),$$

где *к* – число грубых ошибок, $k \le \alpha n$ – целая часть от произведения αn, где n – объём выборочной совокупности, а α – некоторая функция величины засорения выборки ξ. Значения α находят по специальным табл. для расчёта устойчивых оценок Пуанкаре и Винзора. Обычно α а колеблется в пределах от нуля до 0,5. Другой подход демонстрирует оценка Винзора, она предполагает замену признаковых значений, засоряющих выборку, на модифицированные (винзорированные) значения с устраненными или уменьшенными ошибками. Средняя по Винзору определяется также с известным заранее уровнем α (0< α < 1/2) по формуле:

$$W(\alpha) = \frac{1}{n}\sum_{i=k+2}^{n-k-1}(x_i + k(x_{k+1} + x_{n-k})).$$

По аналогии с оценками $T(\alpha)$ и $W(\alpha)$, т.е. соответственно по усечённой совокупности, или винзорированным данным, могут быть найдены не только средние величины, но и другие оценки параметров статистической совокупности, напр., вариации, *моды,* медианы и т.п. Приёмы О.р. Пуанкаре и Винзора дают хорошие результаты на выборках с симметричным распределением засорений, когда грубые ошибки группируются примерно на одном расстоянии от центра в нижней и верхней частях статистической совокупности.

Наряду с уже названными методами О.р., широкое распространение имеет ставший классическим подход Хубера. Он напоминает процедуры для последовательного «улучшения» данных по Винзору. При этом используется некоторая исходная величина *к*, определяемая с учётом степени «засорения» статистической совокупности ξ и определяющая шаг модификации резко отличающихся наблюдений (рассчитывается по таблице значений $k = f(\xi)$ для расчета устойчивой оценки Хубера). Оценка средней величины по методу Хубера производится по формуле:

$$\hat{\theta} = \frac{1}{n}\sum_{|x_i-\theta|<k}^{n-k-1}(x_i + k(n_2 - n_1)),$$

где $\hat{\theta}$ – устойчивая оценка, определяется при помощи итеративных процедур; *к* – величина, которая допускается в качестве отклонения от центра совокупности, принимает постоянные значения с учётом удельного веса грубых ошибок в совокупности данных ξ; n_1 – численность группы наблюдений из совокупности, отличающихся наименьшими значениями: $x_i < \theta - k$, или значения в интервале $(-\infty; \theta - k)$; n_2 – численность группы наблюдений из совокупности, отличающихся наибольшими значениями: $x_i > \theta + k$, или значения в интервале $(\theta + k; \infty)$.

При расчётах по приведённой выше формуле в качестве начальной оценки θ может приниматься обычная средняя арифметическая или медиана, оценённая по выборке. Затем на каждой итерации производится разделение выборочной совокупности на три части. В одну часть попадают «истинные» признаковые значения, которые остаются без изменения

$(|x_i - \theta| \le k)$. В две другие части совокупности (для $x_i > \theta + k$ и $x_i < \theta - k$) попадают «ошибки», они не исключаются из рассмотрения, а заменяются соответственно на величины $x_i - k$ и $x_i + k$. По «истинным» и модифицированным данным каждый раз определяется новая оценка средней θ и итерация возобновляется. Итерации повторяются до тех пор, пока все наблюдения не оказываются в интервале «истинных» значений: $|x_i - \theta| \le k$. Оценка $\hat{\theta}$, найденная по методу Хубера, представляется достаточно эффективной, но быстро теряет оптимальные свойства с увеличением засорения выборки (ростом ξ). В многомерном случае «засорением» совокупности данных уже будут не отдельные значения, а вектор значений, представляющий аномальный объект. Чтобы удостовериться, что многомерное наблюдение является действительно выбросом, обычно используют *расстояние Махаланобиса.*

ОЦИФРОВКА

приписывание величинам, измеренным в номинальной или порядковой шкале, количественных значений, называемых метками. Предполагается существование латентных количественных признаков, которые выражены через данные качественные признаки. Имеет место разбиение латентных признаков на интервалы и приписывание им категориальных значений − градаций. Задача О. − реконструкция исходных латентных количественных признаков, а также некоторых их характеристик: для задач измерения связей − корреляционных функций; при автоматической классификации − функций расстояния и т. д. Такая реконструкция требует предположений о природе подлежащих анализу данных, которые определяются конкретной задачей статистической обработки данных. В качестве критерия используется, как правило, оптимизация некоего функционала от значений исходных признаков (напр., в случае решения задачи анализа связей – максимизация корреляционной функции).

ОШИБКА АППРОКСИМАЦИИ

величина отклонения, возникающая при замене одного математического объекта другим, в том или ином смысле близким к исходному. В теории приближения функций О.а. (погрешность приближения) определяется метрикой некоторого функционального пространства. Если функция $f(t)$ приближается функцией $\varphi(t)$ и обе непрерывны на отрезке $[a,b]$, то часто пользуются равномерной метрикой евклидова пространства

$$\mu(f,\varphi) = \max_{a \le t \le b} |f(t) - \varphi(t)|.$$

Если непрерывность приближаемой функции не гарантирована или важна близость между $f(t)$ и $\varphi(t)$ в среднем на отрезке $[a,b]$, то используют интегральные метрики, полагая

$$\mu(f,\varphi) = \int_a^b q(t)\,|f(t) - \varphi(t)|^p\,dt,\ p > 0,$$

где $q(t)$ – некоторая весовая функция. Наиболее употребительным является случай $p = 2$ (среднеквадратическое приближение). О.а. может рассчитываться по отдельным точкам t_k, $k = 1,\ldots,n$ промежутка $[a,b]$, напр.:

$$\mu(f,\varphi) = \max_{1 \le k \le n} |f(t_k) - \varphi(t_k)|$$

или $\mu(f,\varphi) = \sum_{k=1}^{n} q_k\,|f(t_k) - \varphi(t_k)|^p,\ p > 0,$

где q_k – некоторые положительные коэффициенты. См. также *Аппроксимация функций.*

ОШИБКА ВЫБОРКИ

расхождение между характеристиками выборочной и *ген. совокупности*. Другое название – ошибка репрезентативности. Различают два вида О.в.: случайную и систематическую. При определении случайной ошибки предполагается, что ошибка регистрации равна нулю. Случайная ошибка зависит от размера выборки: чем больше размер выборки, тем она ниже. Обычно, когда говорят об О.в., подразумевают именно случайную ошибку. Систематическую ошибку часто называют ошибкой, вызванной смещением. Возникает как инструментальная О.в., следующая из: неадекватности сформированной выбор-

ки задачам исследования; незнания распределения ген. совокупности и применения процедур отбора, которые могут исказить эти распределения; сознательного отбора наиболее удобных и «выигрышных» для решения задач исследования элементов ген. совокупности, которые, однако, не представляют её в целом, и т.д. При повторных измерениях систематические ошибки остаются постоянными. Общая О.в. складывается из случайной ошибки и из смещения (систематической ошибки), если оно существует.

ОШИБКА ИЗМЕРЕНИЯ

отклонение результата измерения от истинного значения измеряемой величины *x* (абсолютная О.и.). Другое название: погрешность измерения.

Относительной О.и. называется отношение абсолютной погрешности измерения к истинному значению. Различают систематические, случайные и грубые О.и. Систематические ошибки обусловлены гл. обр. погрешностями средств измерений и несовершенством методов измерений, случайные − рядом неконтролируемых обстоятельств; грубые ошибки − неисправностью средств измерений, неправильным отсчитыванием показаний, резкими изменениями условий измерений и т.д. При обработке результатов измерения грубые ошибки обычно отбрасывают; влияние систематических погрешностей стремятся уменьшить внесением поправок или умножением показаний на поправочные множители; оценки случайных ошибок осуществляют методами *математической статистики*.

ОШИБКА ВТОРОГО РОДА

состоит в принятии *нулевой гипотезы*, когда в действительности верна *альтернативная гипотеза*. Вероятность О.в.р. обозначается β и записывается:

$$\beta = P_{H_1}(H_0).$$

Ситуация, когда нулевая гипотеза ложна, а исход эксперимента не попал в *критическую область,* приводит к ошибочному принятию нулевой гипотезы и тем самым допускается О.в.р. Решение, принимаемое на основании любого *критерия статистического,* может оказаться ошибочным как в случае отклонения проверяемой нуль-гипотезы H_0 с вероятностью α, так и в случае её принятия с вероятностью β. Из двух критериев, характеризующихся одной и той же вероятностью α отвергнуть в действительности правильную гипотезу H_0, предпочтительным является тот, который сопровождается меньшей О.в.р. β. Снижая уровень значимости α, можно легко сократить вероятность возникновения *ошибки первого рода*, но в этом случае возрастает вероятность О.в.р. В связи с этим вводится понятие *мощности критерия* $1-\beta$, который задаёт вероятность – не допустить О.в.р., т.е. отклонить гипотезу H_0, когда она неверна. Способом уменьшения величины ошибки как первого, так и второго рода является увеличение объёма выборки.

ОШИБКА ПЕРВОГО РОДА

состоит в том, что в процессе проверки нулевая гипотеза H_0 отвергается, в то время как нулевая гипотеза верна. О.п.р. реализуется в случае, когда при истинности гипотезы H_0, наблюдаемое значение статистики критерия, полученное по данным выборки, попадает в критическую область

$\theta_n^* \in W_\alpha$.

Вероятность допустить О.п.р. $P(H_1|H_0)$ является одной из осн. характеристик надежности результатов экспериментального исследования. Чем меньше вероятность О.п.р., тем надежнее полученные экспериментальные результаты, и наобоот: высокая вероятность О.п.р. говорит о недостоверности результатов эксперимента. Вероятность О.п.р. иначе называют уровнем значимости критерия и обычно обозначают греческой буквой α. В процедуре проверки гипотезы уровень значимости определяет границы критической области. О.п.р. является взаимно-симметричной ошибке второго рода (ошибке, состоящей в том, что в процессе проверки *нулевая гипотеза* H_0 не отвергается, в то время как она не верна). В большинстве практических задач нулевая гипотеза H_0 соответ-

ствует естественному, наиболее ожидаемому положению вещей, а альтернативная гипотеза H_1 обозначает противоположную ситуацию, которая обычно трактуется как менее вероятная, неординарная, требующая какой-либо реакции. В связи с этим, О.п.р. иногда называют ложной тревогой.

П

ПЕРВИЧНАЯ СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ

один из этапов статистической обработки данных, в ходе которого, как правило, решаются задачи: а) отображение переменных, описанных текстом, в номинальную или ординальную (порядковую) шкалу б) унификация типов переменных. Используется либо подход, связанный с переходом от индивидуальных к группированным значениям, либо порядковые и номинальные переменные преобразовываются в количественные с помощью методов *оцифровки* или м*ногомерного шкалирования*; в) статистическое описание исходных совокупностей с определением пределов варьирования переменных; г) выявление резко выделяющихся наблюдений. Резкие отклонения в данных могут быть обусловлены как случайными колебаниями, так и искажениями условий сбора данных или ошибками регистрации. В последних двух случаях наблюдения исключают из рассмотрения: обработка пропущенных наблюдений, сводка и группировка данных. Сводка – научная обработка первичных данных с целью получения обобщенных характеристик изучаемого социально-экономического явления по ряду существенных для него признаков с целью выявления типичных черт и закономерностей, присущих изучаемому явлению в целом. Проведение сводки включает следующие этапы: выбор группировочного признака; определение порядка формирования групп; разработка системы статистических показателей для характеристики групп и объекта в целом; разработка макетов статистических таблиц для представления результатов сводки. В ходе сводки выполняется контроль собранных данных, их систематизация, а также построение таблиц и графиков, расчет итогов и производных показателей в виде средних и относительных величин. Группировка – разбиение общей совокупности единиц объекта наблюдения по одному или нескольким существенным признакам на однородные группы, различающиеся между собой в количественном и качественном отношении и позволяющие выделить социально-экономические типы, изучить структуру совокупности и проанализировать связи между отдельными признаками. Результатом группировки является преобразование данных в упорядоченную статистическую информацию. Простейшая группировка, в которой каждая выделенная группа характеризуется только частотой, является рядом распределения. Также первичная статистическая обработка данных включает учёт размерности и алгоритмической сложности задачи и возможностей вычислительной техники; формулировку задачи на входном языке прикладного программного обеспечения и т.п.

См. также *Вариационный ряд*, *Восстановление пропущенных наблюдений*, *Шкала измерений*.

ПОЛИГОН

наиболее распространённый вид графического изображения *вариационного ряда*. Его обычно используют для изображения дискретного вариационного ряда, но иногда и для изображения непрерывных (интервальных) рядов. Для изображения дискретного ряда в прямоугольной системе координат наносят точки с координатами (x_i, m_i) или (x_i, w_i), где x_i – значение i-го варианта, а m_i (или w_i) – соответствующие частоты (или частота). Затем отмеченные точки соединяют отрезками прямой линии. Полученная ломанная называется П. частот (или частостей, т.е. относительных частот). Если ряд интервальный, то П. строится следующим способом: на оси абсцисс откладываются интервалы значений величины, в серединах интервалов стоятся ординаты, пропорциональные частотам или частостям, и концы ординат соединяются.

ПОСЛЕДОВАТЕЛЬНАЯ СХЕМА НАБЛЮДЕНИЙ

методика, согласно которой число наблюдений, на основании которых статистик принимает решение, не фиксируется заранее, а ставится в зависимость от результатов зарегистрированных на каждой данной стадии эксперимента наблюдений в отличие от классической схемы наблюдений, которая строится на базе *выборки* заранее заданного объёма. Поскольку результаты наблюдений на каждой фиксированной стадии эксперимента представляют собой случайную выборку из *ген. совокупности* и, следовательно, случайны по своей природе, то и момент прекращения наблюдений (определение которого зависит от этих результатов) также является величиной случайной. Впервые идея об использовании П.с.н. возникла в ходе конструирования экономных планов выборочного *статистического контроля качества* продукции. При такой методике достигается заметный выигрыш (в среднем) в числе наблюдений, необходимом для различения интересующих нас гипотез с заданными характеристиками точности. Поэтому к последовательной схеме наблюдений целесообразно обращаться в ситуациях, когда каждое наблюдение является дорогостоящим или труднодоступным и по условиям эксперимента исследователь имеет практическую возможность реализовать эту схему.

ПРИНЦИП ОТНОШЕНИЯ ПРАВДОПОДОБИЯ

принцип построения наилучших критериев путём максимизации *мощности критерия* по отношению к *альтернативной гипотезе* при заданном уровне значимости критерия. Представление о сравнительной правдоподобности имеющихся наблюдений $x_1, x_2, \ldots, x_n$ в отношении *нулевой* и *альтернативной гипотез* даёт сопоставление соответствующих *функций правдоподобия*. Математически принцип реализуется на основе леммы Неймана-Пирсона. Пусть с помощью критерия проверяется нулевая гипотеза

$H_0 : \theta = \theta_0$

против альтернативной

$H_1 : \theta = \theta_1$ на основе выборки $x_1, x_2, \ldots, x_n$ из ген. совокупности с функцией распределения, зависящей от параметра θ, и функция правдоподобия $L(\theta)$ рассчитана по выборке. Для заданного уровня значимости α, критерий, максимизирующий вероятность $(1-\beta)$, где β – вероятность ошибки второго рода, определяется из условия

$$\frac{L(\theta_0)}{L(\theta_1)} < k,$$

где k – число, выбираемое так, чтобы выполнялось условие заданного уровня значимости α.

См. также *Критерий статистический*, *Метод макс. правдоподобия*.

ПРОБИТ (PROBIT) - МОДЕЛЬ БИНАРНОГО ВЫБОРА

модель эконометрическая, основанная на законе нормального распределения N (0,1), в которой зависимая дискретная переменная y является *бинарной переменной*, т.е. принимающей лишь два значения 0 или 1. П.-м.б.в. имеет вид:

$$P(y_i = 1 \mid x_i^{\mathrm{T}}) = \int_{-\infty}^{x_i^{\mathrm{T}}\beta} \phi(t)dt = \Phi(x_i^{\mathrm{T}}\beta)$$

и $P(y_i = 0 \mid x_i^{\mathrm{T}}) = 1 - \Phi(x_i^{\mathrm{T}}\beta)$,

где y_i – дискретная зависимая переменная, x_i^{T} – вектор независимых переменных, Ф(.) – функция плотности вероятностей стандартного нормального закона распределения; Ф(.) – функция распределения нормального закона. П.-м.б.в. называют линейными вероятностными моделями**.**

Графики функции распределения нормального и логистического распределения достаточно близки. На интервале $z \in [-1{,}2; 1{,}2]$ они практически одинаковы. Однако логистическая функция имеет более "тяжелые хвосты", т.е. медленнее стремится к нулю при x $\to$-∞ или единице при x $\to\infty$. Поэтому *логит* и пробит модели дают похожий результат, если только изучаемая вероятность не слишком близка к нулю или единице.

Оценки параметров П.-м.б.в. находят *методом макс. правдоподобия*:

$$\frac{\partial \log L(y_i)}{\partial \beta} = \sum_{i=1}^{n}\left[\frac{y_i - \Phi(x_i^{\mathrm{T}}\beta)}{\Phi(x_i^{\mathrm{T}}\beta)(1-\Phi(x_i^{\mathrm{T}}\beta))}\phi(x_i^{\mathrm{T}}\beta)\right]x_i = 0$$

Решение этого уравнения находят численными методами. Полученная оценка $\hat{\beta}$ называется оценкой макс. правдоподобия.

Для П.-м.б.в.: $\hat{p}_i = \Phi(x_i^{\mathrm{T}}\hat{\beta})$.

Интерпретация коэффициентов отличается от обычной интерпретации коэффициентов линейной модели. В П.-м.б.в. коэффициенты соответствуют предельному (маржинальному) эффекту k-независимой переменной. Этот эффект является функцией всех объясняющих переменных:

$$\frac{\partial \Phi(x_i^{\mathrm{T}}\beta)}{\partial x_{ik}} = \phi(x_i^{\mathrm{T}}\beta)\beta_k \cdot$$

Знак предельного эффекта j-й переменной соответствует знаку коэффициента βj и легко интерпретируется.

Для проверки значимости уравнения, т.е. проверки гипотезы

H_0: $\beta_1=0, \beta_2=0, \ldots, \beta_k=0$

используют *критерий отношений правдоподобия*:

$LR = 2(\ln L - \ln L_0)$,

где lnL – найденное значение логарифма функции правдоподобия; $\ln L_0$ – логарифм правдоподобия при нулевой гипотезе, то есть для тривиальной модели

$$P(y_i = 1) = F(\beta_0) = P_0 \ .$$

При выполнении нулевой гипотезы величина LR имеет хи-квадрат распределение с k степенями свободы. Если вычисленное значение хи-квадрат попадает в критическую область, то есть $LR > \chi^2_{\text{крит}}(\alpha;\ \nu = 1)$, то гипотеза H_0 отвергается, т.е. уравнение в целом значимо. В противном случае H_0 не отвергается.

Для моделей бинарного выбора трудно предложить естественную меру качества аппроксимации такую как коэффициент детерминации R^2 для линейной модели. Чаще всего такие меры строятся путём прямого или косвенного сравнения текущей модели и тривиальной модели. По аналогии с коэффициентов детерминации построен коэффициент:

$$R^2_{pseudo} = 1 - \frac{1}{1 + 2(\ln L - \ln L_0)/n}.$$

Альтернативная мера, называемая индексом отношения правдоподобия, предложена Макфадденом:

$$R^2_{McFadden} = 1 - \frac{\ln L}{\ln L_0}.$$

Если коэффициенты логит или пробит модели незначимы, т.е. все

$\beta_1=0, \beta_2=0, \ldots, \beta_k=0$, то $\ln L = \ln L_0$

и псевдо R^2 и R^2 Макфаддена равны нулю. Если модель совершенно точна, т.е. прогнозные вероятности совпадают с наблюдаемыми значениями $\hat{p}_i = y_i$, тогда все сомножители функции правдоподобия будут равны 1, а логарифм правдоподобия равен нулю. Поэтому верхняя граница, равная 1, может достигаться для индекса отношения правдоподобия Макфаддена.

Альтернативный способ построения мер качества состоит в вычислении прогноза и сравнения его с фактическими значениями. Будем считать, что если предсказанная по модели вероятность больше ½, то прогнозное значение равно 1, если меньше ½, то 0.Так как плотность распределения нормального закона симметрично относительно нуля, это соответствует правилу:

$\hat{y}_i = 1$, *если* $x_i^{\mathrm{T}}\beta > 0$

и $\hat{y}_i = 0$, *если* $x_i^{\mathrm{T}}\beta < 0$.

Доля неправильных прогнозов задаётся формулой:

$$wr_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Логично сравни-вать ошибку прогноза текущей и тривиальной модели. Для тривиальной модели все предсказанные вероятности будут равны между собой и равны доле успехов в выборке $p=n_1/n$. Прогноз для всех наблюдений будет одинаков (все нули или единицы, если p меньше (больше) ½). Число ошибок прогноза будет равно:

$wr_0 = 1 - \hat{p}$, если $\hat{p} > 0{,}5$ и $wr_0 = \hat{p}$, если $\hat{p} \le 0{,}5$. Качество подгонки может быть оценено как

$$R_p^2 = 1 - \frac{wr_1}{wr_0}.$$

Возможно, что данный коэффициент примет отрицательное значение, если число ошибок текущей модели окажется больше, чем для простейшего предсказания.

ПРОСТОЙ СЛУЧАЙНЫЙ (СОБСТВЕННО-СЛУЧАЙНЫЙ) ОТБОР

способ извлечения в *выборку* заданного объёма n единиц *ген. совокупности* общим объёмом N единиц, при котором каждая из возможных выборок имеет равную вероятность быть отобранной. Отбор производится из всей ген. совокупности без разделения на какие-либо группы (серии), единица отбора совпадает с единицей наблюдения. Простую случайную выборку получают последовательно, единица за единицей, осуществляя отбор посредством жеребьевки или с помощью табл. случайных чисел из конечной ген. совокупности. При жеребьевке на каждую единицу ген. совокупности заводится жребий (карточка, жетон, шар), содержащий присвоенный ей номер от 1 до N или иной отличительный признак (название, фамилия, адрес и т.п.). Все жребии тщательно перемешиваются, и из них последовательно в случайном порядке отбирается n жребиев. Отобранные т.о. единицы совокупности образуют выборку. Отбор по табл. случайных чисел предполагает нумерацию единиц ген. совокупности от 1 до N. При этом используется такое количество разрядов чисел табл. случайных чисел, чтобы объём ген. совокупности N не превышал образуемое ими число. В выборку отбираются единицы ген. совокупности, имеющие порядковые номера, соответствующие числам табл. П.с.о. можно осуществить с помощью компьютерных генераторов псевдослучайных чисел. Полученная с помощью генераторов последовательность псевдослучайных чисел носит детерминированный характер, однако вполне обеспечивает выполнение условий случайности отбора и может использоваться так же, как и обычная табл. случайных чисел. П.с.о. может быть повторным и бесповторным. При повторном отборе (схема возвращенного шара) отобранная единица (жребий) после извлечения регистрируется и возвращается в ген. совокупность, откуда может быть извлечена вновь. При *отборе бесповторном* (схема невозвращенного шара) – отобранная единица обратно в ген. совокупность не возвращается и не может быть обследована повторно. При повторном отборе вероятность быть отобранной для каждой единицы ген. совокупности остаётся неизменной на каждом шаге отбора и равна

$$\frac{1}{N};$$

количество выборок заданного объёма n определяется:

$$\left(C_N^n\right)_{повт} = \frac{(N+n-1)!}{n!(N-1)!}.$$

При бесповторном – вероятность быть отобранной изменяется на каждом шаге отбора от

$\frac{1}{N}$ – для первой отбираемой единицы до

$\frac{1}{N-n+1)}$ – для последней,

но для всех единиц, оставшихся в ген. совокупности, вероятность попадания в выборку на каждом очередном шаге отбора одинакова; количество выборок заданного объёма n определяется как

$$C_N^n = \frac{N!}{n!(N-n)!}.$$

На практике, как правило, используется бесповторный отбор.

См. также *Случайные числа* (таблица).

Р

РАВНОМЕРНО НАИБОЛЕЕ МОЩНЫЙ КРИТЕРИЙ

см. в ст. Критерий статистический наиболее мощный

РАВНОТОЧНЫЕ ИЗМЕРЕНИЯ

ряд измерений, выполненных одинаковыми по точности средствами измерений в одних и тех же условиях. С точки зрения статистической обработки результаты измерений $x_1, x_2, \ldots, x_n$ представляют собой *случайные величины*, которые отличаются от истинного значения a из-за случайной $\Delta_i = x_i - a, i = 1, \ldots, n$ погрешности, причём $M(\Delta_i) = 0$ (систематическая погрешность отсутствует) и $M(\Delta_i{}^2) = \sigma^2$. Р.и. в широком смысле истолковываются как одинаковая распределённость погрешностей $\Delta_i, i = 1, \ldots, n$, в узком смысле – как одинаковая мера точности всех результатов измерений, т.е. значение *дисперсии* σ одинаково для всех наблюдений. Наличие грубых ошибок означает нарушение равноточности для некоторых измерений. Если при многократных измерениях $\Delta_i, i = 1, \ldots, n$ независимы и нормально распределены, то величина

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

принимается в качестве оценки истинного значения *a*, причём такая *оценка* является *несмещённой* и *эффективной*.

РАЗМАХ

характеристика вариации *распределения случайной величины* (признака в *ген. совокупности*). Р. определяется как разность между макс. и миним. значениями *случайной величины* с ненулевыми значениями *плотности распределения вероятностей* (для *случайной величины непрерывной*) или *функции распределения вероятностей* (для *случайной величины дискретной*.

См. также *Размах выборки*.

РАЗМАХ ВЫБОРКИ

одна из статистических характеристик распределения признака в выборочной совокупности (R). Р.в. характеризует вариацию изучаемого признака в *выборке*. Определяется как разность между макс. ($x_{\max}$) и миним. ($x_{\min}$) индивидуальными выборочными значениями изучаемого признака:

$R = x_{\max} - x_{\min}$.

В теории статистики обычно используется термин «размах вариации», под которым понимают разность между макс. и миним. индивидуальными значениями изучаемого признака в *ген.* или выборочной *совокупности*. Достоинство Р.в. – простота исчисления. Он необходим тогда, когда важно знать пределы колеблемости изучаемого признака. В практической деятельности используется при контроле качества продукции для выявления влияния систематических факторов на производственный процесс; при анализе инвестиционных проектов для сравнения уровней риска и др. Недостаток Р.в. – чувствительность к случайным выбросам, т.е. к ситуациям, когда макс. и (или) миним. индивидуальные значения признака в выборке обусловлены случайными факторами. В этом случае Р.в. будет смещённой оценкой *размаха* в ген. совокупности. Кроме того, Р.в. не учитывает колеблемость всех индивидуальных значений признака. Более устойчив к значениям крайних вариантов интерквартильный размах (IR) – разность между третьим (Q_3) и первым (Q_1) квартилями распределения изучаемого признака:

$IR = Q_3 - Q_1$. Р.в.

и интерквартильный размах относятся к абсолютным характеристикам вариации.

РАНГОВАЯ КОРРЕЛЯЦИЯ

Ранговые методы базируются на переходе от исходных наблюдений *случайных величин* $X_1, X_2, ..., X_n$ к их рангам $R_1, R_2, ..., R_n$. Рангом R_i наблюдения X_i среди *СВ* $X_1, X_2, ..., X_n$ называется порядковый номер (условная числовая метка), который получит значение X_i при расстановке чисел $X_1, X_2, ..., X_n$ в порядке возрастания. Равным значениям (т.н. связям – от англ .– ties) соответствует среднее ранговое число, которое может быть дробным. Поскольку значения $X_1, X_2, ..., X_n$ являются случайными величинами, то и их ранги – случайные величины. В

статистической практике обычно рассматривается двумерная модель, где данные ранжируются отдельно по каждому компоненту. Выборочная модель: есть выборка

$(X_1,Y_1),...,(X_n,Y_n)$ из $F(x,y)$, где $F(.,.)$

– абсолютно непрерывная функция распределения с абсолютно непрерывными частными функциями распределения

$F_x(.)$ и $F_y(.)$;

наблюдения независимы. Представляем данные двумя строками:

$X_1, X_2, ..., X_n$

$Y_1,\ Y_2,\ ...,\ Y_n$

и предполагаем, что $X_1 < X_2 < ... < X_n$, общность при этом не теряется. Пусть $R_1, R_2, ..., R_n$ – ранги, соответствующие $Y_1, Y_2, ..., Y_n$, тогда получается табл. рангов:

$1 \quad 2 \ ... \ n$

$R_1 \ R_2 \ ... R_n$.

Под Р.к. понимается статистическая связь между порядковыми переменными.

См. также *Коэффициент вариации.*

РАСПРЕДЕЛЕНИЕ ЭМПИРИЧЕСКОЕ

(распределение выборки, выборочное распределение) – статистический аналог *распределения вероятностей*. Р.э. строится по выборке значений $x_1, x_2, ..., x_n$ изучаемого признака (случайной величины) X и может быть представлено в виде *вариационного ряда*, *гистограммы*, *полигона*, *кумуляты* и т.д. Р.э. задается в виде последовательности отдельных значений признака или в виде последовательности интервалов с помощью частот n_i и относительных частот

$$w_i = \frac{n_i}{n}$$

значений признака (i=1,…,k, где k – число групп, на которые разбивается вся совокупность наблюдений). Для эмпирического распределения можно определить те же характеристики, что и для распределения вероятностей, напр., *эмпирическую функцию распределения*, выборочную среднюю, выборочные квантили и т.д. Эмпирическое распределение может быть использовано для приближенного представления теоретического распределения (распределения *ген. совокупности*). При этом характеристики Р.э. (выборочные характеристики) служат *статистическими оценками* соответствующих характеристик ген. совокупности. Напр., относительная частота w_i появления в выборке интересующего события (значения признака) служит оценкой соответствующей вероятности (что следует из *закона больших чисел*). Для решения задачи установления теоретического закона распределения случайной величины, характеризующей изучаемый признак, по эмпирическому распределению и для оценки степени расхождения между теоретическим и эмпирическим распределениями используют *критерий согласия*. Они основаны на использовании различных мер расстояний между анализируемой эмпирической функцией распределения, определяемой по выборке, и функцией распределения генеральной совокупности. Наиболее часто на практике используются критерии согласия χ^2 - *критерий Пирсона* и λ - к*ритерий согласия Колмогорова*.

См. также *Выборочные характеристики*.

РЕАЛЬНЫЙ КОМПЛЕКС УСЛОВИЙ

совокупность условий (факторов) случайного эксперимента, при осуществлении которых происходит некоторое *случайное событие*.

Предполагается неизменность Р.к.у. и возможность его воспроизведения сколь угодно большое число раз, т.е. эксперимент проводится неоднократно при неизменном комплексе условий. В силу неизбежности влияния большого числа случайных (не поддающихся строгому учёту и контролю) факторов, точный результат таких действий невозможно предсказать с полной уверенностью и сделать выводы о том, произойдет или не произойдет интересующее нас событие. Симметрия и стационарность (т.е. неизменность во времени) Р.к.у., при котором происходит случайный эксперимент, приводят к принятию гипотезы о «равновозможности» его исходов и лежат в основе возникающих при

этом закономерностей и классического определения вероятности.

РЕПРЕЗЕНТАТИВНОСТЬ ВЫБОРКИ (ПРЕДСТАВИТЕЛЬНОСТЬ ВЫБОРКИ)

способность выборочной совокупности полно и адекватно представлять свойства анализируемой ген. совокупности. Р.в. определяется исходя из того, насколько интересующие нас функции и параметры выборочной совокупности соответствуют функциям и параметрам изучаемой ген. совокупности, что приводит к необходимости оценки надёжности результатов выборки и возможности их распространения на ген. совокупность. Выборка тем более репрезентативна, чем больше её объём. Р.в. обычно достигается сочетанием случайного и направленного отбора. Случайный отбор предполагает равную вероятность каждой из единиц, а также любой комбинации единиц ген. совокупности быть отобранными в выборку. Случайность отбора обеспечивается различными способами организации выборки. Случайная выборка позволяет выносить вероятностные суждения об оценках параметров ген. совокупности. Невыполнение требования случайности отбора приводит к смещению оценок. Направленный отбор используется ввиду сложностей организации обследований, не позволяющих провести полностью случайный отбор, а также для обеспечения представительства в выборке относительно малочисленных слоев (типических групп) ген. совокупности.

См. также *Выборка простая случайная.*

РЯД ВАРИАЦИОННЫЙ

см. в ст. Вариационный ряд

С

СЕРИЙНАЯ (ГНЕЗДОВАЯ) ВЫБОРКА

способ организации выборочной совокупности, при котором *выборка* объёмом s серий (гнезд) извлекается из общего числа S серий, образующих ген. совокупность. При этом все единицы ген. совокупности объёмом N единиц распределены между непересекающимися сериями, так что объём серий $N_1, N_2, ..., N_s$ полностью исчерпывает ген. совокупность, т. е.

$$N_1 + N_2 + ... + N_s = N .$$

Выделенные серии должны быть максимально идентичными по осн. характеристикам, тогда как единицы совокупности внутри серий – отличаться друг от друга как можно сильнее (серии должны быть максимально однородными «снаружи» и максимально разнородными – «внутри»). Отбор серий осуществляется либо по жребию, либо по таблице случайных чисел, либо механически через одинаковый интервал. Внутри серий обследуются все единицы совокупности. Отбор серий может быть повторным и бесповторным. На практике, как правило, используется *отбор бесповторный*. Серии могут содержать как одинаковое, так и разное количество единиц совокупности. С.в. используется в тех случаях, когда организационно легче отбирать серии единиц совокупности, а не отдельные единицы. В пром-сти к С.в. прибегают при обследованиях серийного произва, когда единицы продукции образуют серии (партии) изделий обычно одинакового объёма. Напр., С.в. имеет широкое распространение при выборочном *статистическом контроле качества* продукции, особенно в случаях применения т.н. мерной тары. При проведении обследований нас. в сел. хоз-ве, в случае отсутствия надёжной основы выборки, подлежащие обследованию адм.-терр. единицы делятся на р-оны, участки, дома, кварталы, посёлки, фермы и т.п., выступающие в данном случае в качестве гнезд более мелких единиц обычно неравного объёма.

СКОШЕННОСТЬ ВАРИАЦИОННОГО РЯДА

см. в ст. Коэффициент асимметрии

СОВОКУПНОСТЬ ГЕНЕРАЛЬНАЯ

см. в ст. Генеральная совокупность

СОСТОЯТЕЛЬНОСТЬ ОЦЕНКИ

Состоятельность – один из принципов, с помощью которого формируется отношение к *оценкам точечным* параметров данного распределения вероятностей**.** Первым идею формализации принципа состоятельности высказал Р.Фишер. Достаточное условие С.о.:

1. смещение оценки $B_n \to 0$ (при $n \to \infty$);

2. $D\theta_n^* \to 0$ при $n \to \infty$. Согласно *неравенству Чебышева* из *закона больших чисел* можно записать, что

$$P\left\{\left|\theta_n^* - M\left(\theta^*\right)\right| < \varepsilon\right\} \geq 1 - \frac{D\left(\theta_n^*\right)}{\varepsilon^2}.$$

Смещением оценки называется разность $B_n = M\theta_n^* - \theta$. Отсюда $M\left(\theta_n^*\right) = \theta + B_n$.

Тогда $P\left\{\left|\theta_n^* - \theta - B_n\right| < \varepsilon\right\} \geq 1 - \frac{D\theta_n^*}{\varepsilon^2}$

и $\lim\limits_{n\to\infty} P\left\{\left|\theta_n^* - \theta - B_n\right| < \varepsilon\right\} \geq \lim\limits_{n\to\infty}\left(1 - \frac{D\theta_n^*}{\varepsilon^2}\right)$.

В случае, когда $B_n \to 0$ и $D\theta_n^* \to 0$

при $n \to \infty$, получаем:

$$\lim_{n\to\infty} P\left\{\left|\theta_n^* - \theta\right| < \varepsilon\right\} \geq 1 \Rightarrow \lim_{n\to\infty} P\left\{\left|\theta_n^* - \theta\right| < \varepsilon\right\} = 1.$$

Если *оценка состоятельная*, то увеличение объёма наблюдений n ведёт к улучшению оценки и есть смысл увеличить n.

СПОСОБЫ ОРГАНИЗАЦИИ ВЫБОРКИ

способы извлечения единиц (серий) *ген. совокупности* в *выборку*; предназначены для обеспечения случайности отбора единиц (серий) ген. совокупности в выборку, что является условием *репрезентативности выборки*. В каждом конкретном случае выбор С.о.в. зависит от ряда факторов: целей и задач обследования, объёма и структуры ген. совокупности, выбора единицы наблюдения, вариации изучаемых признаков, размеров финансирования, наличия квалифицированного персонала и др. Осн. С.о.в.: простой случайный (собственно-случайный) отбор, механический (систематический) отбор, серийный (гнездовой) отбор, расслоенный (стратифицированный, типический, районированный) отбор. Перечисленные способы отбора предполагают отбор единиц ген. совокупности в выборку непосредственно для наблюдения. В этом случае отбор называют одноступенчатым. Организация масштабных обследований, как правило, приводит к необходимости проведения отбора единиц совокупности в несколько этапов. Такая выборка называется многоступенчатой. Если на разных этапах (ступенях) используются разные способы отбора, то выборка называется комбинированной. При этом сначала отбираются крупные единицы совокупности (серии), а затем из них – более мелкие до тех пор, пока не будут отобраны единицы (серии), подвергающиеся обследованию. Если на каждой последующей ступени (фазе) отбора осуществляется подвыборка из уже отобранных единиц (серий), обследуемых на каждом этапе по расширяющейся от фазы к фазе программе, выборку называют многофазной. При многофазной выборке единица отбора на каждом этапе остается неизменной.

См. также *Выборка простая случайная*, *Выборка расслоенная*, *Серийная выборка*.

СРАВНЕНИЕ ДВУХ ГЕНЕРАЛЬНЫХ ДИСПЕРСИЙ

проверка гипотезы

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

Пусть $\xi_1 \sim N\left(\theta_1; \sigma_1^2\right)$

$\xi_2 \sim N\left(\theta_2; \sigma_2^2\right)$

– две независимые нормально распределённые СВ, параметры которых неизвестны. Из первой извлечена выборка объёма n:

$X = \left(x_1, ..., x_n\right)$,

из второй – объёма m: $Y = \left(y_1, ..., y_m\right)$.

Для этих выборок существуют статистики

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \; ; \; \hat{S}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 ;$$

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i \; ; \hat{S}_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(y_i - \bar{y}\right)^2 .$$

Пусть наибольшей дисперсией является $\hat{S}_1^2$, тогда статистика критерия имеет вид

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}.F$$

образуют так, чтобы дробь была >1, т.е. в числителе стояла наибольшая выборочная дисперсия. Рассмотрим случай, когда

$$\left.\begin{array}{l} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 > \sigma_2^2 \end{array}\right\}$$

Для проверки H_0 при заданном уровне значимости α. Критическая область имеет вид

$$\left(f_{1-\alpha}(n-1,m-1),+\infty\right),$$

где $f_{1-\alpha}$ – квантиль F распределения с числом степеней свободы $n-1$ и $m-1$. Если F – статистика попадает в критическую область, то H_0 – отклоняется, в противном случае – нет оснований отклонить H_0, при заданном уровне значимости α. Рассмотрим случай, когда

$$\begin{array}{l} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{array}.$$

В этом случае также существует равномерно наиболее мощный несмещённый критерий размера

$$\alpha = 2P\left\{F_{n-1,m-1} > f_{1-\frac{\alpha}{2}}(n-1,m-1)\right\}.$$

Если $F_{набл} > f_{1-\frac{\alpha}{2}}(n-1,m-1)$,

то гипотеза H_0 отклоняется. В противном случае нет оснований отклонить гипотезу H_0. Замечания: если принимается решение о равенстве $\sigma_1 = \sigma_2$ (т.е. принимается H_0), то в этом случае можно образовать объединенную оценку общей дисперсии

$$\hat{S}^2 = \frac{(n-1)\hat{S}_1^2 + (m-1)\hat{S}_2^2}{m+m-2}.$$

СРАВНЕНИЕ НЕСКОЛЬКИХ ВЕРОЯТНОСТЕЙ

статистическая проверка гипотезы о значениях вероятностей в условиях полиномиальной модели

$H_0 : p_1 = \pi_1, p_2 = \pi_2, ..., p_l = \pi_l$, г

де p_i – вероятность i-го исхода в n независимых испытаниях ($i = 1,...,l$). Гипотеза проверяется с помощью статистики критерия

$$\chi^2 = \frac{\sum_{i=1}^{l}(n_i - n\pi_i)^2}{n\pi_i},$$

имеющей χ^2-распределение с $k = l-1$ степенями свободы, где n_i – частота i-го исхода,

$$\sum_{i=1}^{l} n_i = n.$$

Критерий основан на аппроксимации полиномиального распределения распределением χ^2, поэтому применяется при больших значения n (напр., $n > 25$) и не слишком малых математических ожиданиях исходов ($n\pi_i \geq 5$, однако по результатам исследований У. Кохрена в некоторых случаях возможно $n\pi_i \geq 1$).

При сравнении вероятностей биномиальных распределений проверяется статистическая гипотеза о равенстве ген. долей вероятностей некоторого признака в *l* ген. совокупностях

$$H_0 : p_1 = p_2 = ... = p_l = p.$$

Проверка гипотезы основана на сравнении выборочных долей признака

$$w_i = \frac{m_i}{n_i},$$

рассчитанных по *l* независимым выборкам достаточно большого объёма n_i, где m_i – число наблюдений *i*-й выборки, обладающих данным признаком, ($i = 1,...,l$). Наилучшее приближение для неизвестной вероятности *p* даёт выборочная доля признака объединенной выборочной совокупности объёма

$$n = n_1 + ... + n_l \quad \hat{p} = \frac{m_1 + ... + m_l}{n_1 + ... + n_l}.$$

Критерием проверки *нулевой гипотезы* является величина

$$\chi^2 = \frac{\sum_{i=1}^{l} n_i(w_i - \hat{p})^2}{\hat{p}(1-\hat{p})},$$

которая при условии справедливости нулевой гипотезы и при $n \to \infty$ имеет χ^2-распределение с $k = l-1$ степенями свободы.

При сравнении долей признака в двух совокупностях при достаточно большом числе испытаний n_1 и n_2 в качестве критерия проверки нулевой гипотезы используется величина

$$z = \frac{w_1 - w_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

которая в условиях справедливости H_0 имеет стандартное нормально распределение.

См. также *Полиномиальное распределение.*

СРАВНЕНИЕ НЕСКОЛЬКИХ ГЕНЕРАЛЬНЫХ СРЕДНИХ

статистическая проверка гипотезы о равенстве *математических ожиданий l ген. совокупностей*

$X_1, X_2, ..., X_l$ $H_0 : a_1 = a_2 = ... = a_l$.

Альтернативная гипотеза состоит в том, что не все средние равны. Если ген. совокупности

$X_1, X_2, ..., X_l$

распределены нормально с постоянной (хотя и неизвестной) дисперсией σ_0^2, то для сравнения нескольких средних используется *дисперсионный анализ* и проверка *нулевой гипотезы* H_0 сводится к сравнению *несмещённых оценок* неизвестной ген. дисперсии: факторной дисперсии (является несмещённой в случае справедливости H_0)

$$S^2_{факт} = \frac{\sum_j (\bar{x}_j - \bar{x})^2}{l-1}$$

и остаточной дисперсии $S^2_{ост} = \frac{\sum_{j=1}^{l} n_j s_j^2}{n_1 + ... + n_l - l}$,

где $\bar{x}$ – общая выборочная средняя, $\bar{x}_j, s_j^2$ – выборочные средние дисперсии, полученные по *j*-й выборке, n_j – объём *j*-й выборки (*j*=1,..,*l*). Критерием проверки нулевой гипотезы является *F*-статистика

$$F = \frac{S^2_{факт}}{S^2_{ост}},$$

имеющая в условиях справедливости H_0 распределение Фишера-Снедекора с числом степеней свободы числителя $l-1$ и знаменателя $n_1 + ... + n_l - l$. В случае двух ген. совокупностей *F*-критерий эквивалентен *t-критерию Стьюдента.* В качестве критической статистики используется величина

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

имеющая в условиях справедливости H_0 распределение Стьюдента с $n_1 + n_2 - 2$ степенями свободы. Непараметрическим критерием для сравнения нескольких ген. средних является *критерий ранговый* Краскела-Уоллиса, основанный на статистике

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{l} \frac{R_j^2}{n_j} - 3(n+1),$$

где $R_j = \sum_{i=1}^{n_j} R_{ij}$

– сумма рангов *j*-й выборки, $n = n_1 + ... + n_l$. При небольших объёмах данных ($n_j \le 5$ или $l \le 3$) эмпирические значения сравниваются с критическими значениями критерия Краскела-Уоллиса, приведённых в специальных таблицах. При больших объёмах выборок распределение статистики H асимптотически сходится к распределению

χ^2 с $k = l - 1$

степенями свободы. Непараметрическим аналогом критерия Стьюдента для двух выборок является критерий Вилкоксона

$$W = \sum_{i=1}^{n_1} R_i^{(1)},$$

где $R_i^{(1)}$ – ранг *i*-го элемента первой выборки в общем вариационном ряду, построенном по двум выборкам, и *U*-критерий Манна-Уитни

$U = n_1 n_2 + n_1(n_1+1)/2 - W$.

Поскольку статистики W и U линейно связаны, то часто говорят об одном критерии Вилкоксона-Манна-Уитни. В условиях справедливости нулевой гипотезы распределение статистики W асимптотически сходится к нормальному. Для проверки гипотезы используется стандартизированное значение статистики

$$W_{\text{ñò}} = \frac{W - \frac{1}{2}n_1(n_1+n_2+1)}{\sqrt{\frac{1}{12}n_1n_2(n_1+n_2)}}.$$

Для небольших выборок ($n_1+n_2 \le 8$) используются специальные таблицы критических значений для критерия Вилкоксона.

СРЕДНЕЕ ЗНАЧЕНИЕ ВЫБОРОЧНОЕ

среднее значение р*аспределения эмпирического*. Среднее значение – осн. характеристика распределения выборочной совокупности и центра группирования значений исследуемой случайной величины. Для выборки $x_1, x_2, \ldots, x_n$ объёма n С.з.в. рассчитывается как средняя арифметическая простая

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Для сгруппированных данных среднее значение вычисляется по формуле средней арифметической взвешенной

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i f}{\sum_{i=1}^{n} f_i}_{\,i},$$

где f_i – частота появления значения x_i. С.з.в. широко используется для обобщённых характеристик массовых процессов. С его помощью устраняются индивидуальные различия, выявляются общие условия и закономерности, осуществляются расчёты по *прогнозированию* и планированию, анализу явлений.

СРЕДНЕЕ ЗНАЧЕНИЕ ГЕОМЕТРИЧЕСКОЕ

число, равное корню степени n из произведений отдельных положительных значений величины x. Для значений $x_1, x_2, \ldots, x_n$ С.з.г. рассчитывается как

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}.$$

Для сгруппированных данных С.з.г. вычисляется по формуле средней геометрической взвешенной

$$G = \sqrt[\sum_{i=1}^{n} f_i]{x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots \cdot x_n^{f_n}},$$

где f_i – частота появления значения x_i. Геометрическое среднее всегда меньше арифметического среднего, кроме случая, когда все значения $x_1, x_2, \ldots, x_n$ равны между собой. Геометрическое среднее двух чисел называется средним пропорциональным. С.з.г. находит применение при расчётах средних темпов (коэффициентов роста), прироста и, в частности, в тех случаях, когда имеют дело с величиной, изменения которой происходят приблизительно в прямо пропорциональной зависимости с достигнутым к этому моменту уровнем самой величины (напр., численность нас.), или же когда имеют дело со средней из отношений, напр., при расчётах индексов цен. Напр., если инфляция за первый год составила 12%, за второй год – 14%, то среднее значение инфляции

$G = \sqrt{1{,}12 \cdot 1{,}14} = 1{,}129$ или 12,9%.

СРЕДНЕЕ ЗНАЧЕНИЕ ГАРМОНИЧЕСКОЕ

число, обратная величина которого является средним арифметическим обратных величин данных чисел $x_1, x_2, \ldots, x_n$, т.е.

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}.$$

Для сгруппированных данных С.з.г. вычисляется по формуле средней гармонической взвешенной

$$H = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n}\frac{1}{x_i} f_i},$$

где f_i – частота появления значения x_i. С.з.г. ряда чисел всегда меньше *среднего значения геометрического* тех же чисел, которое в свою очередь меньше их среднего арифметического. Область его применения весьма ограничена. В экономике, в частности, пользуются гармоническим средним при анализе средних норм времени, а также в некоторых видах индексных расчётов. Напр., один из работников на обработку одной детали затрачивает 5 мин., а другой − 15 мин., и необходимо вычислить сред-

ние затраты времени на обработку одной детали. При условии, что общая продолжительность рабочего времени у работников одинаковая, получим по формуле средней гармонической

$$H = \frac{1}{\frac{1}{2}\left(\frac{1}{5} + \frac{1}{15}\right)} = 7,5 \text{ минут.}$$

СТАТИСТИЧЕСКАЯ ГИПОТЕЗА

см. в ст. Гипотеза статистическая

СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

система приёмов *математической статистики*, предназначенных для проверки соответствия опытных данных некоторому предположительному утверждению (*гипотезе статистической)*. Процедуры С.п.г. позволяют принимать или отвергать статистические гипотезы, возникающие при обработке или интерпретации результатов измерений во многих практически важных разделах науки и произ-ва, связанных с экспериментом.

Первый этап процедуры С.п.г. состоит в формулировке гипотез исходя из требований прикладной задачи. Чаще всего рассматриваются две гипотезы – *нулевая* H_0 и альтернативная H_1. В некоторых случаях число альтернатив может быть больше. Если *альтернативная гипотеза* не задана явно, то полагают, что она формулируется как отрицание нулевой гипотезы.

На втором этапе по данным выборки $x^{(n)} = (x_1, x_2, \ldots, x_n)$ объёмом n строится специально составленная выборочная характеристика (статистика) $\theta_n^* = f(x_1, x_2, \ldots, x_n)$, $\theta_n^*: X^n \to R$. Для большинства гипотез можно предложить целый набор статистик, и выбор между ними не всегда однозначен. Осн. требование, предъявляемое к статистике θ_n^*, состоит в том, что точный или приближённый закон её распределения при истинности гипотезы H_0 известен или может быть выведен. Вывод функции распределения статистики при заданных H_0 и $x^{(n)}$ является строгой математической задачей, которая решается методами теории вероятностей.

Третий этап процедуры – построение критической области W_α – множества наименее вероятных значений статистики θ_n^*. В зависимости от вида конкурирующей гипотезы выбирают правостороннюю, левостороннюю или двустороннюю критические области. Границы критической области определяются величиной уровня значимости α – допустимой для данной задачи вероятности того, что гипотеза на самом деле верна, но будет отвергнута процедурой проверки. Это должно быть достаточно малое число

$0 \le \alpha \le 1$,

как правило, полагают α = 0,05. При заданном уровне значимости α границы критической области находят из соотношений: для правосторонней критической области

$P(\theta_n^* > \theta_{кр}) = \alpha$;

для левосторонней критической области

$P(\theta_n^* < \theta_{кр}) = \alpha$;

для двусторонней симметрической области

$P(\theta_n^* > \theta_{кр\_пр}) = \alpha / 2$ $(\theta_n^* < \theta_{кр\_лев}) =$

$= \alpha / 2$ $\theta_{кр\_лев} < \theta_{кр\_пр}$.

Вычисление границ критической области является строгой математической задачей, которая решается на основе функции распределения статистики θ_n^*, полученной на предыдущем этапе. В практических задачах границы критической области часто определяются из соответствующей табл. закона распределения, которому подчиняется статистика.

Последний этап проверки гипотезы – применение *критерия статистического* – однозначно определённого правила, устанавливающего условия, при которых проверяемую гипотезу (H_0) следует либо отвергнуть, либо не отвергнуть. Каждый критерий разбивает все множество возможных значений статистики

$\theta_n^* = f(x_1, x_2, \ldots, x_n)$

на два непересекающихся подмножества (области): критическая область W_α и область принятия гипотезы. Осн. принцип проверки гипотезы состоит в следующем: если наблюденные зна-

чения статистики критерия попадают в критическую область

$\theta_n^* \in W_\alpha$,

то гипотеза отвергается на уровне значимости α. В противном случае, если

$\theta_n^* \notin W_\alpha$,

гипотезу не отвергают и полагают, что данные не противоречат нулевой гипотезе на уровне значимости α. При реализации процедуры проверки гипотезы возможны четыре случая: 1. гипотеза H_0 верна и её принимают согласно критерию; 2. гипотеза H_0 не верна и её отвергают согласно критерию; 3. гипотеза верна H_0, но её отвергают согласно критерию (*ошибка первого рода* или «ложная тревога»); 4. гипотеза H_0 не верна, но её принимают согласно критерию (ошибка второго рода или «пропуск цели»).

Т.о. в двух случаях принимается правильное решение, а два других сопряжены с возможностью совершения ошибок – первого рода

$P(H_1|H_0) = \alpha$

или второго рода

$P(H_0|H_1) = \beta$.

При проверке гипотез желательно добиваться минимизации значений ошибок обоих родов. Но в большинстве задач невозможно одновременно минимизировать обе ошибки. Стремление минимизировать одну из них приводит к возрастанию другой.

Выбор границы критической области неоднозначен, поэтому налагают еще одно условие, состоящее в том, что вероятность отвержения гипотезы H_0 при истинности гипотезы H_1, т.е. вероятность не совершить ошибку второго рода, была максимальна. Т.е. требования к критической области аналитически можно записываются в виде:

$$\begin{cases} P(\theta_n^* \in W_\alpha | H_0) = \alpha \\ P(\theta_n^* \in W_\alpha | H_1) = 1 - \beta \Rightarrow \max \end{cases}$$

Второе условие выражает требование максимума мощности критерия (см. рис. 1).

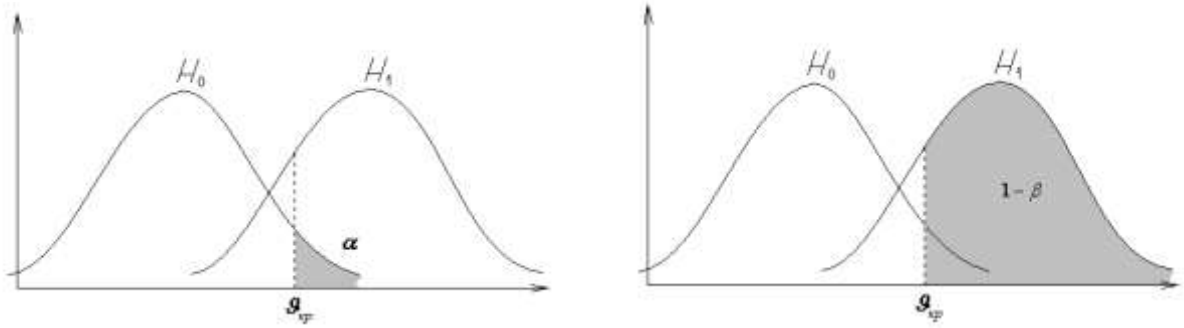


Рис.1 **Уровень значимости и мощность критерия**

Процедура проверки гипотезы может быть осуществлена и в другом порядке. Этот метод основан на использовании т.н. достигаемого уровня значимости $p(\theta_n^*)$ – наименьшей величины уровня значимости, при котором нулевая гипотеза H_0 отвергается для данного значения статистики критерия

θ_n^*: $p(\theta_n^*) = \min\{\alpha : \theta_n^* \in W_\alpha|\}$.

Достигаемый уровень значимости сравнивается с некоторым фиксированным уровнем, и в случае, если первый превышает последний, гипотеза отвергается.

Если эмпирические данные согласуются с предполагаемой гипотезой, это не исключает возможности согласования тех же данных с другой гипотезой. Принятие гипотезы не означает, что она является единственно верной (или даже самой правильной). Поэтому принятие гипотезы – не более чем достаточно правдоподобное, не противоречащее опыту, предположение. Строго доказать справедливость проверяемой гипотезы – невозможно. Допустимо лишь утверждать, что гипотеза не противоречит опытным данным.

Кроме этого, полученные в результате проверки гипотезы выводы о её непротиворечивости данным наблюдений не всегда можно рассматривать как подтверждение истинности гипотезы. Во-первых, многое зависит от объёма выборки n: при малых объемах выборки, данных может быть недостаточно для обнаружения их несоответствия гипотезе. Во-вторых, результаты проверки гипотезы определяются выбором статистики θ_n^*, и если она отражает не всю информацию о гипотезе H_0, увеличивается вероятность принятия гипотезы, когда она не верна.

СТАТИСТИЧЕСКИЙ КОНТРОЛЬ КАЧЕСТВА

раздел *математической статистики*, методы которого используются при оценке достигнутого уровня качества и тенденций его изменения. Статистический контроль играет важную роль в системе мероприятий по управлению качеством продукции. Это обусловлено в первую очередь тем, что изменчивость числовых характеристик осн. показателей качества изделий носит случайный характер. Стремление сделать контроль более объективным приводит к необходимости использования методов случайной *выборки*, что также обусловливает необходимость использования вероятностных и статистических методов.

С.к.к. используется как для регулирования хода технологического процесса, так и для оценки качества партий продукции. В первом случае преследуется цель предупреждения брака путём периодического наблюдения за ходом процесса и своевременного вмешательства в него. Во втором случае решается чисто контрольная задача – проверка соответствия партии сырья (входной контроль) или готовой продукции (приемочный контроль) техническим условиям. Оба эти метода базируются на выборочных процедурах математической статистики. При выборочном контроле необходимо по результатам проверки части изделий сделать вывод о качестве всей партии.

В зависимости от вида контрольной операции различают контроль по альтернативному, качественному и количественному признакам. При контроле по альтернативному признаку изделия по результатам измерений разбивают на пригодные и дефектные. При контроле по качественному признаку изделия классифицируются на несколько групп. Напр., по результатам контроля изделие может быть отнесено к 1, 2, 3-й группе качества или к браку. При контроле по количественному признаку измеряется числовое значение параметра.

Каждый из видов контроля имеет свои преимущества и недостатки. Напр., при контроле по качественному признаку каждое наблюдение несёт меньшую информацию, чем при контроле по количественному признаку, в связи с чем для получения обоснованных решений требуется большее число наблюдений. Но при качественном контроле процесс наблюдений проще и его легче автоматизировать. Кроме того, методика качественного контроля не связана с видом распределения контролируемого признака и поэтому является более универсальной. Если С.к.к. определяет степень годности продукции, то он называется статистическим приёмочным контролем, если цель – оценка состояния технологического процесса для решения вопроса о необходимости его наладки, то С.к.к. называется статистическим регулированием технологических процессов.

С.к.к. производится путём отбора экземпляров продукции (из каждой партии, через равные интервалы времени на конвейере и т. п.) и измерения свойств, определяющих качество продукции. Результаты измерений сравниваются с предельными значениями свойств, при превышении которых продукция бракуется (приёмочный контроль) или принимается решение о наладке технологического процесса (статистическое регулирование технологических процессов). Осн. числовая характеристика планов приёмочного контроля по альтернативному признаку – оперативная характеристика P(q), равная вероятности принять партию продукции с долей дефектных изделий q по результатам выборочного контроля. Если контроль проводят с использованием одноступенчатых планов, то решение о качестве партии принимают по результатам одной выборки. При двухступенчатом плане решение о качестве партии принимают либо по результатам контроля одной вы-

борки, либо двух. Как правило, при этом контролируется лишь вошедшая в выборку часть партии изделий, поэтому возможны ошибочные решения. Разрабатываются статические методы оценки эффективности планов контроля на основе информации, накапливаемой в ходе контроля. С.к.к. по количественному признаку наиболее часто основан на предположении, что наблюдения в выборке являются взаимно независимыми одинаково распределёнными случайными величинами с заданным, обычно нормальным, законом распределения.

Т

ТАБЛИЦЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

необходимые для статистических выводов специальные математические табл., содержащие информацию об осн. типах функций распределений случайных величин, процентных точках и квантилях (табл. нормального распределения, χ^2-распределения, *t*-распределения Стьюдента, *F*-распределения Фишера-Снедекора и др.). Используются на заключительном этапе статистических исследований для определения границ интервальных оценок параметров ген. совокупности, критических значений используемых статистик при проверке статистических гипотез. Законы распределения статистик критериев зависят от параметров задачи: объёма выборки, степени сложности, уровня значимости, на котором проверяется критерий и т.п. в Т.м.с. даются распределения статистик различных критериев, квантили, процентные точки и критические значения при различных значениях параметров. В пояснениях к табл. обычно указывают способы их интерполяции и экстраполяции на внетабличные значения, даются асимптотические формулы для вычисления значений затабулированных функций во всей естественной области их определения. Т.м.с. создаются с помощью компьютерных технологий и содержат также различные вспомогательные табл., напр., случайных чисел, значения факториалов, элементарных математических функций. В наиболее полных Т.м.с. содержатся обоснования и вывод функций распределения, интервальных оценок и критических статистик критериев, пояснения с примерами решения конкретных, наиболее часто встречающихся математических задач. К наиболее полным и известным Т.м.с. относятся табл. Большева Л.Н. и Смирнова Н.В.

Несмотря на возросшие возможности компьютерных и статистических вычислений, сборники таких табл. по-прежнему остаются важными и актуальными, т.к. служат полезным справочным руководством по применениям вероятностных и статистических методов, анализу и интерпретации полученных результатов.

ТЕОРИЯ АСИМПТОТИЧЕСКОГО ОЦЕНИВАНИЯ

часть теории статистического оценивания, исследующая построение и сравнение асимптотических оценок, сравнение по их свойствам при достаточно большом объёме *выборки* (при $n \to \infty$). Т.а.о. получила развитие в задачах, в которых строятся оценки, приближающиеся к оптимальным, при условии, что те или иные параметры стремятся к предельным значениям. Пусть оценивается тот или иной параметр ген. распределения по выборке $x_1, x_2, \ldots, x_n$, объём которой безгранично растёт. Оценка рассматриваемого параметра θ есть функция

$$\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$$

наблюдений. В силу поставленной задачи Т.а.о. – теория предельного поведения оценки

$$\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$$

при $n \to \infty$. При исследовании предельного поведения оценки устанавливается её состоятельность. Оценка состоятельная – оценка, которая с ростом объёма выборки стремится по вероятности к оцениваемому параметру. Для определения *состоятельности оценок* в Т.а.о. используют *метод макс. правдоподобия*, *метод моментов*, *метод наименьших квадратов*, *байесовскую оценку*.

Состоятельность является одним из асимптотических требований, которые следует предъявлять к оценке. *Асимптотические свойства оценок* помимо состоятельности включают асимптотическую эффективность, асимптотическую

нормальность. Асимптотически эффективные такие оценки, выборочные распределения которых при $n \to \infty$ меньше рассеиваются около рассматриваемого значения параметра, т.е. оценки с возможно меньшей стандартной ошибкой. Последовательность оценок называется асимптотически нормальной, если функция распределения сходится к *нормальному закону*. При достаточно большом объёме выборки п асимптотическая нормальность позволяет построить приближенные *доверительные интервалы* и асимптотические критерии.

ТЕОРИЯ ОШИБОК

раздел *математической статистики*, посвящённый построению уточнённых выводов о численных значениях приближённо измеренных величин, а также об ошибках (погрешностях) измерений. Повторные измерения одной и той же постоянной величины дают, как правило, различные результаты, т.к. каждое измерение содержит некоторую ошибку. Различают три осн. вида ошибок: систематические, грубые и случайные. Систематические ошибки всё время либо преувеличивают, либо преуменьшают результаты измерений и происходят от определённых причин (неправильной установки измерительных приборов, влияния окружающей среды и т. д.), систематически влияющих на измерения и изменяющих их в одном направлении. Оценка систематических ошибок производится с помощью методов, выходящих за пределы математической статистики. Грубые ошибки возникают в результате просчёта, неправильного чтения показаний измерительного прибора и т. п. Результаты измерений, содержащие грубые ошибки, сильно отличаются от других результатов измерений и поэтому часто бывают хорошо заметны. Случайные ошибки происходят от различных случайных причин, действующих при каждом из отдельных измерений непредвиденным образом то в сторону уменьшения, то в сторону увеличения результатов.

Т.о. занимается изучением лишь грубых и случайных ошибок. Осн. задачи Т.о.: определение законов распределения случайных ошибок, оценок неизвестных измеряемых величин по результатам измерений, установление погрешностей таких оценок и устранение грубых ошибок.

Пусть в результате n независимых равноточных измерений некоторой неизвестной величины a получены значения $x_1, x_2, \ldots, x_n$. Разности $d_1 = x_1 - a, \ldots, d_n = x_n - a$ называются истинными ошибками. В терминах вероятностной Т.о. все d_i трактуются как случайные величины, независимость измерений понимается как взаимная независимость случайных величин $d_1, \ldots, d_n$. Равноточность измерений в широком смысле истолковывается как одинаковая распределённость: истинные ошибки равноточных измерений суть одинаково распределённые случайные величины. При этом математическое ожидание случайных ошибок $b = Md_1 = \ldots = Md_n$ называется систематической ошибкой, а разности $d_1 - b, \ldots, d_n - b$ – случайными ошибками. Т.о., отсутствие систематической ошибки означает, что $b = 0$, и в этой ситуации $d_1, \ldots, d_n$ суть случайные ошибки. Величину

$$1 / \sigma\sqrt{2},$$

где σ – квадратичное отклонение, называют мерой точности (при наличии систематической ошибки мера точности выражается отношением

$$1 / \sqrt{2(b^2 + \sigma^2)}).$$

Равноточность измерений в узком смысле понимается как одинаковость меры точности всех результатов измерений. Наличие грубых ошибок означает нарушение равноточности (как в широком, так и в узком смысле) для некоторых отдельных измерений. В качестве оценки неизвестной величины *a* обычно берут арифметическое среднее из результатов измерений

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

а разности $D_1 = x_1 - \bar{x}, \ldots, D_n = x_n - \bar{x}$ называются кажущимися ошибками. Выбор $\bar{x}$ в качестве оценки для *a* основан на том, что при достаточно большом числе *n* равноточных измерений, лишённых систематической ошибки, оценка $\bar{x}$

с вероятностью, сколь угодно близкой к единице, сколь угодно мало отличается от неизвестной величины *a*; оценка $\bar{x}$ лишена систематической ошибки (оценки с таким свойством называются несмещенными); дисперсия оценки:

$$D\bar{x} = M(\bar{x} - a)^2 = \frac{\sigma^2}{n},$$

где $\sigma^2 = Dx_i$ – дисперсия отдельного измерения x_i. Опыт показывает, что практически очень часто случайные ошибки d_i подчиняются распределениям, близким к нормальному (причины этого вскрыты так называемыми предельными теоремами *теории вероятностей*). В этом случае величина $\bar{x}$ имеет распределение, мало отличающееся от нормального, с математическим ожиданием *a* и дисперсией $\frac{\sigma^2}{n}$. Если распределения d_i в точности нормальны, то дисперсия всякой другой *оценки несмещённой* для a, напр., *медианы*, не меньше $D\bar{x}$. Если же распределение d_i отлично от нормального, то последнее свойство может не иметь места.

Если дисперсия σ^2 отдельных измерений заранее известна, то для её оценки пользуются величиной

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} D_i^2 \quad (Ms^2 = \sigma^2,$$

т. е. s^2 – несмещённая оценка для σ^2), если случайные ошибки d_i имеют нормальное распределение, то отношение

$$t = \frac{(\bar{x} - a)\sqrt{n}}{s}$$

подчиняется *распределению Стьюдента* с *n* - 1 степенями свободы. Этим можно воспользоваться для оценки погрешности приближённого равенства $a \cong \bar{x}$.

Величина $\frac{1}{(n-1)}\sum_{i=1}^{n} d_i^2$

при тех же предположениях имеет *распределение-* χ^2 с *n* - 1 степенями свободы. Это позволяет оценить погрешность приближённого равенства $\sigma \cong s$. Можно показать, что относительная погрешность $|\sigma - s|/\sigma$ не будет превышать числа *q* с вероятностью

w = *F* (z_2, *n* - *1*) - *F* (z_1, *n* - *1*),

где F (z, n - 1) – функция распределения c^2,

$$z_1 = \frac{\sqrt{n-1}}{1+q}, \quad z_2 = \frac{\sqrt{n-1}}{1-q}.$$

ТЕОРИЯ СТАТИСТИЧЕСКИХ РЕШЕНИЙ

часть *математической статистики и* теории игр, позволяющая единым образом охватить такие разнообразные задачи, как *статистическая проверка гипотез*, построение статистических оценок параметров и *доверительных границ* для них, *планирование эксперимента* и др. В основе Т.с.р. лежит предположение, что распределение вероятностей F наблюдаемой случайной величины X_F принадлежит некоторому априори данному множеству G. Осн. задача Т.с.р. состоит в отыскании наилучшего статистического решения или решающего правила (функции) d = d (x), позволяющего по результатам наблюдений x над X судить об истинном (но неизвестном) распределении F. Для сравнения достоинств различных решающих правил вводят в рассмотрение функцию потерь W [F, d (x)], представляющую убыток от принятия решения d (x) (из заданного множества D), когда истинное распределение есть F. Естественно, было бы считать решающее правило d* = d*(x) наилучшим, если средний риск r (F, d*) = M_FW [F, d (X)] (M_F – усреднение по распределению F) не превышает r (F, d) для любого $F \in G$ и любого решающего правила d = d (x). Однако такое "равномерно наилучшее" решающее правило в большинстве задач отсутствует, в связи с чем наибольший интерес в Т.с.р. представляет отыскание т.н. минимаксных и байесовских решений. Решение $\tilde{d} = \tilde{d}(x)$ называется минимаксным, если $\sup r(F, \tilde{d}) = \inf_d \sup r(F, d)$.

Решение $\bar{d} \equiv \bar{d}(x)$ называется бейесовским (относительно заданного априорного распределения *n* на множестве σ), если для всех решающих правил *d* $R(\theta, \bar{d}) \le R(\theta, d)$, где, между минимаксными и байесовскими решениями существует тесная связь, заключающаяся в том, что в весьма широких предположениях о данных задачи минимаксное решение является байесовским относительно "наименее благоприятного" априорного распределения p.

Совр. концепция статистического решения выдвинута А.Вальдом и считает поведение оптимальным, если оно минимизирует риск в последовательных экспериментах, т.е. математическое ожидание убытков статистического эксперимента. В такой постановке любая задача статистических решений может рассматриваться как игра двух лиц, в которой одним из игроков является "природа".

Выбор наилучших решений в условиях неполной информации – одно из осн. занятий людей. Если процесс определяется повторяющимися ситуациями, то его усреднённые характеристики испытывают тенденцию к стабилизации и появляется возможность либо замены случайного процесса детерминированным, либо использования каких-то методов исследования стационарных случайных процессов, в частности, методов *теории массового обслуживания*. Однако большинство процессов характеризуется "дурной неопределённостью" и невозможно найти законы распределения и другие вероятностные характеристики. В таких ситуациях приходится прибегнуть к экспертным оценкам.

Возникает и проблема выбора критерия оптимальности, поскольку решение, оптимальное для каких-то условий, бывает неприемлемым в других, поэтому приходится искать некоторый компромисс.

Пусть задан некоторый вектор

$S = (S_1,S_2,..,S_n)$,

описывающий n состояний внешней среды, и вектор

$X = (X_1,X_2,..,X_m)$,

описывающий m допустимых решений. Требуется найти вектор

$X^* =(0,0,..,0, X_i ,0,..,0)$,

который обеспечивает оптимум некоторой функции полезности W(X,S) по некоторому критерию K. Информация об указанной функции представляют матрицей размерности m x n с элементами

$W_{ij} = F(X_i, S_j)$,

где F – решающее правило.

Предположим, что в нашем распоряжении имеются статистические данные, позволяющие оценить вероятность того или иного события, и этот опыт может быть использован для оценки будущего. При известных вероятностях P_j для спроса S_j можно найти математическое ожидание W(X,S,P) и определить вектор X^*, дающий

$$W = \max_{i=1,m} \sum_{j=1}^{n} W_{ij} P_j \ .$$

Критерий Лапласа. В основе этого критерия лежит "принцип недостаточного основания".

Если нет достаточных оснований считать, что вероятности того или иного спроса имеют неравномерное распределение, то они принимаются одинаковыми, и задача сводится к поиску варианта, дающего

$$W = \max_{i=1,m} \frac{1}{n} \sum_{j=1}^{n} W_{ij} \ .$$

Критерий Вальда обеспечивает выбор осторожной, пессимистической стратегии в той или иной деятельности. Его суждения близки к тем суждениям, которые мы использовали в теории игр для поиска седловой точки в пространстве чистых стратегий: для каждого решения X_i выбирается самая худшая ситуация (наименьшее из W_{ij}), и среди них отыскивается гарантированный макс. эффект:

$$W = \max_{i=1,m} \min_{j=1,n} W_{ij} \ .$$

Можно принять и критерий выбора оптимистической стратегии:

$$W = \min_{i=1,m} \max_{j=1,n} W_{ij} \ .$$

Критерий Гурвица. Ориентация на самый худший исход является своеобразной перестраховкой. Однако опрометчиво выбирать политику, которая излишне оптимистична. Критерий Гурвица предлагает некоторый компромисс:

$$W = \max_{i=1,m} \left[\alpha \max_{j=1,n} W_{ij} + (1-\alpha) \min_{j=1,n} W_{ij} \right],$$

где параметр α принимает значение от 0 до 1 и выступает как коэффициент оптимизма.

Критерий Сэвиджа. Суть этого критерия заключается в нахождении миним. риска. При выборе решения по этому критерию сначала

матрице функции полезности (эффективности) сопоставляется матрица сожалений:

$$D_{ij} = W_{ij} - \max_i W_{ij},$$

элементы которой отражают убытки от ошибочного действия, т.е. выгоду, упущенную в результате принятия i-го решения в j-м состоянии. Затем по матрице D выбирается решение по пессимистическому критерию Вальда, дающее наименьшее значение макс. сожаления.

ТОЛЕРАНТНЫЕ ГРАНИЦЫ

границы, для которых с заданной *вероятностью* β доля распределения, содержащаяся в интервале между ними, не меньше чем γ. Это границы, удовлетворяющие следующему условию:

$$P\left\{\int_{l_1}^{l_2} f(x)dx \geq \gamma\right\} = \beta,$$

где l_1 и l_2 – нижняя и верхняя Т.г., $f(x)$ – неизвестная непрерывная функция плотности. Левая часть данного выражения имеет значение, независящее от $f(x)$, тогда (Уилкс, 1942) и только тогда (Роббинс, 1944), когда l_1 и l_2 являются порядковыми статистиками. Понятие Т.г. (или толерантные пределы) ввёл Шьюарт в 1931. Интервал, находящийся между Т.г., называют толерантным. Толерантный интервал может быть двусторонним, когда имеются обе Т.г. и выполняется соотношение:

$$P\{P(l_1 \leq X \leq l_2) \geq \gamma\} = \beta;$$

либо односторонним, когда имеется верхняя или нижняя граница:

$$P\{P(-\infty < X \leq l_2) \geq \gamma\} = \beta$$

$$(P\{P(l_1 \leq X < \infty) \geq \gamma\} = \beta).$$

Т.г. можно строить без предположений (кроме непрерывности) о виде исходного распределения. Для нормального распределения вывод Т.г. рассмотрели в своих работах Вальд и Волфовиц. Фрэзэр и Гаттмэн определили Т.г. для интервала, покрывающиго в среднем заданную часть исходного нормального распределения. При определении Т.г. используются параметрический и непараметрический методы для нормального распределения. Непараметрический метод определения толерантных интервалов не требует знания вида функции распределения совокупности, но применим лишь в случаях, когда известно, что функция распределения непрерывна. В случае нормального распределения Т.г. определяются по выборочным характеристикам (выборочной средней $\bar{x}$ и среднему квадратическому отклонению S) с помощью множителей, которые являются функциями от *квантилей* нормального распределения или χ^2 – распределения, по следующим формулам:

$$l_1 =, \bar{x} - kS \quad \text{и} \quad l_2 = \bar{x} + kS,$$

где коэффициент k зависит от β, γ и объёма выборки n.

ТОЧЕЧНАЯ ОЦЕНКА

оценка статистическая $\hat{\theta}$, значение которой принимают в качестве приближенного значения оцениваемого параметра θ или функции выбранной параметрической модели. Т.о. является функцией

$$\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$$

независимых случайных величин $X_1, X_2, ..., X_n$

с распределением, зависящим от распределения исходной *случайной величины* X, в т.ч. от самого оцениваемого параметра θ и от объёма выборки n. Если будет произведена иная выборка, то Т.о. примет другое значение. *Оценка* $\hat{\theta}$ является лишь приближённым значением оцениваемого параметра, если даже она *состоятельная, несмещённая и эффективная*. Оценка $\hat{\theta}$ называется состоятельной, если с ростом объёма *выборки* она стремится по вероятности к оцениваемому параметру

$$\theta: \hat{\theta} \xrightarrow{P} \theta \quad при \quad n \to \infty.$$

Оценка $\hat{\theta}$ называется несмещённой, если её *математическое ожидание* равно оцениваемому параметру при любом фиксированном объеме выборки, т.е.

$$M(\hat{\theta}) = \theta.$$

Несмещённость означает отсутствие систематической ошибки. Если математическое ожидание оценки сходится к параметру

$M(\hat{\theta}) \to \theta$ при $n \to \infty$,

то оценка $\hat{\theta}$ называется асимптотически несмещённой. Несмещённая оценка $\hat{\theta}$ называется эффективной, если она имеет миним. *дисперсию* среди всех возможных Т.о. при заданном объёме выборки. На практике не всегда удается удовлетворить всем трём требованиям, но наиболее важные – состоятельность и несмещённость. Для построения Т.о. параметра выбранной параметрической модели используют методы: *метод моментов*, предложенный К.Пирсоном, и метод наибольшего правдоподобия, предложенный Р.Фишером. Т.о. оцениваемых параметров ген. совокупности могут быть приняты в качестве первоначальных результатов обработки выборочных данных. Их недостаток заключается в том, что неизвестно, с какой точностью оценивается параметр θ. Если для выборок большого объёма точность обычно бывает достаточной (при условии несмещённости, эффективности и состоятельности оценок), то для выборок небольшого объёма вопрос точности оценок становится очень важным.

См. также *Метод макс. правдоподобия.*

ТОЧНОСТЬ ИНТЕРВАЛЬНОЙ ОЦЕНКИ

величина, показывающая наибольшее значение абсолютной разности

$$\left|\theta - \hat{\theta}\right| < \delta$$

оцениваемого параметра θ и *точечной оценки* $\hat{\theta}$, при заданной *доверительной вероятности* γ. Характеристика δ – некоторое положительное число, полученное по выборке из п независимых наблюдений

$X_1, X_2, ..., X_n$.

Чем меньше значение характеристики δ, тем оценка $\hat{\theta}$ точнее описывает оцениваемый параметр θ. *Оценка интервальная* оцениваемого параметра – это *доверительный интервал*. Доверительный интервал – это числовой интервал, который с заранее выбранной вероятностью γ покрывает оцениваемый параметр θ. Границы доверительного интервала являются *случайными величинами*. Доверительный интервал для оценки *математического ожидания* нормального распределения при известном среднеквадратическом отклонении σ имеет вид:

$$\left|\bar{x} - a\right| < t\sigma/\sqrt{n},$$

где a – неизвестное математическое ожидание, которое оценивается по выборочной средней $\bar{x}$. Т.и.о. равна

$$\delta = t\sigma/\sqrt{n},$$

аргумент t находится при заданном уровне вероятности по табл. функции Лапласа ($\Phi(t) = \gamma/2$). При возрастании объёма *выборки* n значение δ уменьшается, а следовательно Т.и.о. уменьшается. Если происходит увеличение надёжности, то это приведет к увеличению точности δ. Для оценки математического ожидания при неизвестном среднеквадратическом отклонении используют точность равную

$$\delta = \frac{t\hat{s}}{\sqrt{n-1}},$$

где $\hat{s}$ – исправленное среднее квадратическое отклонение случайной величины X; вычисляется по выборке:

$$\hat{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}.$$

Аргумент t имеет *распределение Стьюдента* с п-1 числом степеней свободы.

См. также *Доверительная вероятность.*

У

УРЕЗАНИЕ РАСПРЕДЕЛЕНИЯ

Понятие связано с ситуациями, когда исследуемый признак не может быть наблюдаем в какой-либо части области его возможных значений. Урезание – свойство распределения, в отличие от цензурирования не всегда приводит к потере эффективности оценивания. В отличие от *цензурированных выборок* в выборках урезанных распределений мы не имеем возможно-

сти оценить даже доли наблюдений, располагающихся за пределами порога урезания. Напр., если исследуется распределение семьи по доходу, но по условиям выборочного обследования лишены возможности наблюдать семьи со среднедушевым доходом меньшим некоторого заданного уровня a (тыс. руб.), то в подобных случаях говорят, что распределение урезано слева в точке a.

УРОВЕНЬ ЗНАЧИМОСТИ КРИТЕРИЯ

положительное число, которое в задаче проверки статистических гипотез ограничивает сверху вероятность *ошибки первого рода*. Обычно У.з.к. принимают равным одному из стандартных значений 0,1, 0,05, 0,01 и т.д. При данном объёме наблюдений уменьшение У.з.к., т.е. вероятности ошибки первого рода, понижает «чувствительность» критерия за счёт увеличения вероятности *ошибки второго рода*. Поэтому выбор У.з.к. в реальных задачах основывается на содержательных соображениях, учитывающих практические последствия ошибок первого и второго рода.

УСТОЙЧИВОСТЬ СТАТИСТИЧЕСКАЯ

свойство *выборочных характеристик* (выборочные начальные и центральные моменты, выборочная относительная частота, выборочные функции распределения и плотности), полученных при достаточно большом числе наблюдений мало изменять свои номинальные значения. Иными словами, выборочные характеристики, рассчитанные для различных *выборок* (достаточно большого объёма n) из одной и тоже *ген. совокупности*, будут приближённо равны между собой. У.с. позволяет использовать выборочные характеристики распределения в качестве приблизительного описания свойств ген. совокупности в целом. Теоретическим обоснованием подобного замещения служит *закон больших чисел*, согласно которому при увеличении объёма выборки n выборочная характеристика стремится по вероятности к ген. характеристике.

УСТОЙЧИВЫЕ СТАТИСТИЧЕСКИЕ ВЫВОДЫ

статистические выводы, мало меняющиеся при переходе от одного критерия к другому. Для статистического вывода необходима формализованная система статистических методов, сводящихся к попытке описать свойства большого объема данных путем обследования малой его части. Статистические методы позволяют на основе обрабатываемых данных уточнить вероятностную модель с помощью методов оценки параметров и проверки гипотез. Для одних и тех же данных можно определить множество методов статистической обработки. Каждому методу обработки данных будет соответствовать свой статистический вывод. Но чтобы из множества возможных методов анализа данных выбрать наиболее подходящий, необходимо ввести критерий качества, который позволит определить методы, дающие устойчивые выводы.

Ф

ФУНКЦИЯ МОЩНОСТИ КРИТЕРИЯ

функция, которая характеризует качество *критерия статистического*. Пусть по реализации x случайного вектора X, принимающего значения в выборочном пространстве (χ, β, P_θ), $\theta \in \Theta$, надлежит проверить гипотезу H_0, согласно которому распределение вероятностей P_θ случайного вектора X принадлежит подмножеству

$H_0 = \{P_\theta, \theta \in \Theta_0\}$

против альтернативы H_1, согласно которому

$P_\theta \in H_1 = \{P_\theta, \theta \in \Theta_1 / \Theta_0\}$

и пусть $\varphi(\bullet)$ – критическая функция статистического критерия, предназначенного для проверки H_0 против H_1, где Θ_1 – область, в которой гипотеза H_0 не согласуется с результатами наблюдений. Тогда

$$\beta(\theta) = \int \varphi(x) dP_\theta(x)$$

называется *мощностью критерия* статистического, имеющего критическую функцию φ. Из формулы следует, что Ф.м.к. $\beta(\theta)$ показывает, с какими вероятностями статистический критерий, предназначенный для проверки H_0 против

H_1 отклоняет проверяемую гипотезу H_0, если подчиняется закону $P_\theta, \theta \in \Theta$. В теории проверки *гипотез статистических*, основанной Нейманом и Пирсоном, задача проверки сложной гипотезы H_0 против сложной альтернативы H_1 формулируется в терминах Ф.м.к. и заключается в построении статистического критерия, максимизирующего Ф.м.к. $\beta(\theta)$, когда $\theta \in \Theta_1$ при условии, что $\beta(\theta) \le \alpha$ для всех $\theta \in \Theta_0$, где число $\alpha (0 < \alpha < 1)$, называется *уровнем значимости критерия*. Т.е. Ф.м.к. показывает допустимую вероятность ошибочного отклонения гипотезы H_0, в то время, как проверяемая гипотеза верна. Критическая функция – измеримая функция φ ($0 \le \varphi \le 1$), определённая на выборочном пространстве. Данная функция задаёт статистический критерий, согласно которому гипотеза H_0 отвергается (принимается) с вероятностью $\varphi(x)(1-\varphi(x))$, которая зависит от результата наблюдений.

ФУНКЦИЯ ПРАВДОПОДОБИЯ

функция вида L(X*,Θ), выражающая совместную вероятность (или плотность вероятности) появления набора значений X*= $\left(x_1^*, x_2^*, ..., x_n^*\right)$ при извлечении из *ген. совокупности* выборки объёма n. Значение Ф.п. определяется соотношением:

$$L(X^*,\Theta) = \begin{cases} P\left(\prod_{i=1}^{n} I\left(X_i = x_i^*\right)\right) = \prod_{i=1}^{n} P\left(X_i = x_i^*\right) = \prod_{i=1}^{n} P\left(X_i = x_i^*, \Theta\right), & \text{если } X \text{ дискретна} \\ f_{X_1, X_2, ..., X_n}\left(x_1^*, x_2^*, ..., x_n^*\right) = \prod_{i=1}^{n} f_{X_i}\left(x_i^*\right) = \prod_{i=1}^{n} f_{X_i}\left(x_i^*, \Theta\right), & \text{если } X \text{ непрерывна}. \end{cases}$$

Из соотношения следует, что чем вероятнее (правдоподобнее) набор $x_1^*, x_2^*, ..., x_n^*$ при заданном Θ, тем больше значение L(X*,Θ). Т.о. функция L при фиксированном Θ служит мерой правдоподобия набора X*. Часто для целей упрощения расчётов вместо L используют функцию log L, в этом случае ф.п. называется *логарифмической функцией правдоподобия*. Ф.п. лежит в основе *метода макс. правдоподобия* Понятие «Ф.п.» было введено Р.Фишером.

Х

ХАРАКТЕРИСТИКИ ВЫБОРОЧНЫЕ

см. в ст. Выборочные характеристики

Ц

ЦЕНЗУРИРОВАНИЕ ВЫБОРКИ

приписывание ряду «хвостовых» членов *вариационного ряда* нулевых весов, а остальным – одинаковых положительных значений. Если приписывание производится по признаку выхода текущих значений наблюдений за пределы заданного диапазона $[a, b]$, т.е.

$$\omega(x_i) = \begin{cases} \omega_0 > 0, & \text{если } a \le x_i \le b; \\ 0, & \text{если } x_i < a \text{ или } x_i > b, \end{cases}$$

то говорят о цензурировании 1-го типа. Очевидно, в этом случае число ν оставшихся в рассмотрении наблюдений – *величина случайная* (ν<n). Если же нулевые веса приписываются фиксированной доле α крайних малых значений и фиксированной доле β крайних больших значений, то говорят, что производится цензурирование 2-го типа уровня (α, β). В этом случае число ν оставшихся в рассмотрении наблюдений является величиной, заранее заданной и равной, в частности, $n(1-\alpha-\beta)$. Исследователь может прибегнуть к цензурированию вынужденно или добровольно. Вынужденное цензурирование обусловлено соответствующими условиями эксперимента. Напр., на разрушающее испытание ставится n изделие, но эксперимент может производиться в течение ограниченного времени T. Необходимо вынужденно произвести цензурирование 1-го типа, при котором из дальнейшего рассмотрения исключаются точные значения долговечности (времени до разрушения) всех тех изделий, которые не разрушились за время T. С другой стороны, в классе

оценок, построенных по Ц.в., часто можно найти оценки, хоть и не являющиеся наилучшими в жестких рамках ген. совокупности определённого типа, но обладающие выгодными условиями устойчивости своих хороших качеств по отношению к тем или иным отклонениям от априорных допущений.

Для Ц.в. необходимо применять свои методы оценки показателей, *статистических проверок гипотез.* Теория обработки Ц.в. сложнее традиционных методов математической статистики и далека от своего завершения.

ЦЕНТР РАССЕИВАНИЯ

точка, относительно которой наблюдается разброс элементов выборочной совокупности. Это точка c на прямой, относительно которой находится мера разброса выборки $X_1, \ldots, X_n$, равная

$$Q_n = \sum_{i=1}^{n} (x_i - c)^2 .$$

В одномерном случае Ц.р. могут выступать выборочные средняя, *медиана* или *мода*, построенные по исходной выборке $X=(X_1, \ldots, X_n)$, подчиняющаяся нормальному закону распределения. Самый распространенный вид Ц.р. – выборочная средняя, или *математическое ожидание* эмпирического распределения вероятностей, построенного по выборке $X_1, \ldots, X_n$. Рассеивание выборки минимально, если оно вычислено относительно выборочного среднего.

ЦЕНТРАЛЬНЫЕ МОМЕНТЫ ВЫБОРОЧНЫЕ

эмпирический аналог центральных моментов. Ц.м.в. k - го порядка определяют как среднее значение k –й степени отклонений наблюдаемых значений *случайной величины* от выборочной средней. Ц.м.в. находят по формуле

$$\hat{m}_k^o = \frac{1}{n} \sum_{i-1}^{n} (x_i - \bar{x})^k .$$

Для группированных выборочных данных, когда n наблюдений разбиты на P групп с n_i наблюдениями в i-й группе *(1,2,...,p)* Ц.м.в. определяются по формуле:

$$\hat{m}_k^o = \frac{1}{n} \sum_{i-1}^{n} n_i (x_i - \bar{x})^k).$$

Для случайной величины *непрерывной* Ц.м.в. второго порядка равен *выборочной дисперсии*. При большом объёме *выборки* (при $n \to \infty$) каждый Ц.м.в. распределён асимптотически нормально. Ц.м.в. используют для вычисления *выборочных характеристик*. С помощью $\hat{m}_2^0$, $\hat{m}_3^0$ и $\hat{m}_4^0$ вычисляют асимметрию и эксцесс.

Ч

ЧАСТОСТЬ

см. в ст. Относительная частота (частость)

ЧАСТОТА ОТНОСИТЕЛЬНАЯ

см. в ст. Относительная частота (частость)

ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ

число «свободных» единиц совокупности, на которые не наложены ограничения при вычислении выборочной характеристики, т.е. объём *выборки* минус число оцененных параметров. Ч.с.с. указывает на то, какой объём единиц совокупности остался свободным после их частичного использования в конкретном виде анализа. При определении *дисперсии* Ч.с.с. должно показать, сколько независимых отклонений из n возможных

$$[(y_1 - \bar{y})^2, (y_2 - \bar{y})^2, \ldots, (y_n - \bar{y})^2]$$

требуется для образования данной суммы квадратов. Так, для общей суммы квадратов

$$\sum (y_i - \bar{y})^2$$

необходимо n–1 независимых отклонений, т.к. после расчёта среднего уровня по совокупности из n единиц свободно варьируют лишь $(n-1)$ число отклонений. Ч.с.с. также используется при построении разного рода *критериев статистических* и *оценок интервальных* параметров, таких как *распределение* «*хи-квадрат*» (χ^2); *распределение Стьюдента(t)*; F - распределение (распределение дисперсионного отношения); β- распределение (*бета-распределение*) и γ - распределение (*гамма - распределение*).

Ш

ШКАЛА ИЗМЕРЕНИЙ

определённое упорядоченное отображение качественного или количественного свойства измеряемых объектов на упорядоченное множество чисел или другую систему логически связанных знаков. Измерение – одна из древнейших операций, применявшихся человеком в практической деятельности. Практика измерения восходит в своих началах к истокам науки, однако логические основания измерения не изучались вплоть до кон. 19 – нач. 20 вв., когда Гельмгольц изложил осн. идеи репрезентационной теории измерения, а Гельдер развил аксиоматику измерения экстенсивных величин. В 19 в. вопрос об измерении возникал гл. обр. в отношении физических величин. Он сводился к счёту прерывных и измерению непрерывных величин в натуральных единицах. С развитием психологии и социологии, начиная с 30-х гг. 20 в., появилась острая необходимость в сопоставлении величин, не удовлетворяющих условиям аддитивности. Это привело к созданию новой теории измерений. Идея её была выдвинута американским психологом Стивенсом С., который в 1950 ввёл понятие «измерительная шкала». По его мнению, измерение – это приписывание числовых форм объектам или событиям в соответствии с определёнными правилами, что создает шкалу. Далее идеи Стивенса были восприняты математиками (П. Суппес, И. Пфанцагль, Д. Кранц и др.).

Различные меры повторяемости, воспроизводимости фактов являются измерениями или шкалами. Ш.и. – средство адекватного сопоставления и определения численных значений отдельных свойств и качеств различных объектов. При измерении одного свойства объекта, приходиться пренебрегать всеми другими свойствами, что приводит к тому, что несходные объекты могут стать эквивалентными, напр., все объекты одинаковой длины считаются эквивалентными независимо от веса, формы и т.д. Измерение включает элементы: объект, его свойство или качество, единицу измерения, технические средства, метод измерения, наблюдателя. Измеряемый объект находится в определённых условиях и под воздействием ряда факторов. Реакция объекта на воздействие выражается в виде сигналов, которые содержат информацию о свойствах объекта. То, что измеряется называется переменной, то чем измеряют – инструментом измерения, от которого непосредственно зависит тип Ш.и. Тип шкалы задает группу допустимых преобразований, т.е. преобразованные объекты, которые находятся в пределах рассматриваемого класса. Полученные результаты называются данными, либо результатами измерения и могут относиться к одной из Ш.и. Каждая шкала предполагает использование определённых математических операций, и, соответственно, ограничивает применение методов *математической статистики*. При изучении реального явления или процесса следует прежде всего установить типы шкал, в которых измерены те или иные переменные. Типы шкал характеризуются свойствами эквивалентности и упорядочения. В исследовании различные данные измеряются в различных шкалах. Формализация данных осуществляется в наиболее удобной форме для их передачи и обработки. Данные могут быть представлены в аналоговом или цифровом (кодовом) виде. Шкалы разделяют на *количественные*, когда может быть установлена единица измерения, и, в противном случае, на неколичественные. С. Стивенс предложил следующую классификацию Ш.и.: 1) номинативная или номинальная, или шкала наименований; 2) порядковая или ранговая шкала; 3) интервальная или шкала равных интервалов; 4) шкала равных отношений. Соответственно имеются четыре типа переменных: номинальные, используемые только для качественной классификации и позволяющие разбивать исследуемую совокупность объектов по анализируемому свойству на неподдающиеся упорядочиванию однородные группы; порядковые (ординальные), позволяющие ранжировать объекты и указывающие какие из них в большей или меньшей степени обладают тем или иным качеством; интервальные и относительные, скалярно измеряющие в определённой шкале степень проявления изучаемого свойства объекта и позволяющие не только упорядочивать объекты

измерения, но и численно выразить и сравнить различия между ними.

Статистические данные могут измеряться на различных уровнях с применением измерительных шкал различной степени точности. Низший уровень представлен номинальной шкалой, указывающей лишь на то, в какой группе относится рассматриваемая величина. Следующим уровнем измерения является порядковая шкала, на которой категории располагаются иерархически, но она не позволяет измерить различия между категориями. Шкалы интервального уровня также являются упорядоченными, но имеют дополнительное свойство, заключающееся в том, что расстояние между делениями на шкале одинаково. Величины на интервальной шкале можно сравнивать с точки зрения их упорядоченности или расстояния между ними. Шкала отношений (пропорциональный уровень) сочетает свойства интервальных шкал и связано с точкой абсолютного нуля. Измерение на интервальном и пропорциональном уровнях является метрическим, тогда как номинальный и порядковый уровни – неметрические.

Уровни измерения влияют на виды *математико-статистических методов*, которые используются при анализе данных. На номинальном уровне проведение исследований основано на частотах или расчёте таких показателей, как *мода*, χ^2 (хи-квадрат), *табл. сопряжённости*. Измерение на порядковом уровне соответствует всему кругу непараметрической статистики. Интервальный же и пропорциональный уровни позволяют использовать все статистические методы. В зависимости от природы изучаемых данных используются различные методы статистического анализа при исследовании зависимостей. Так, если независимые и зависимые переменные оценены в номинальной шкале, применяют анализ табл. сопряжённости, если в количественной – *корреляционный и регрессионный анализы*. Если же независимая переменная задается в номинальной шкале, а зависимая – в количественной, то можно использовать *дисперсионный анализ*, если имеет место противоположная ситуация: независимая переменная – в количественной шкале, а зависимая – номинальной, то целесообразнее использовать методы *дискриминантного или кластерного анализа*. Выбранная Ш.и. определяет характер информации, которой будет располагать исследователь при исследовании какого-либо объекта. При этом выбор шкалы должен соответствовать как типу изучаемого показателя и характеру отношений между объектами, так целям и задачам исследования. Любому измерению предшествует качественный анализ, учитывающий цели исследования. Качественный анализ необходим и после того, как измерение произведено, для того чтобы оценить адекватность результатов измерения объектов поставленным целям. В естественных науках проблема точности измерения связана, прежде всего, с самим процессом измерения, а именно, с разработкой правил и методов измерений, формированием системы принципов, постулатов и других теоретических положений, формирующих базис точности, основанием выбора типа шкал при конструировании измерителя и т.д.

Э

ЭМПИРИЧЕСКАЯ (ВЫБОРОЧНАЯ) ДИСПЕРСИЯ

статистическая оценка теоретической *дисперсии ген. совокупности*, определяемая как средняя арифметическая квадрата отклонения значений выборочных данных $x_1, x_2, ..., x_n$ от выборочной средней $\bar{x}$:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 .$$

Она обладает свойствами асимптотической несмещённости в силу

$$Ms^2 = \left(1 - \frac{1}{n}\right)m_2^0$$

и состоятельности в силу

$$Ds^2 = \frac{m_4^0 - m_2^{0\,2}}{n} - \frac{2\left(m_4^0 - 2m_2^{0\,2}\right)}{n^2} + \frac{m_4^0 - 3m_2^{0\,2}}{n^3},$$

где m_2^0 и m_4^0 – второй и четвертый центральные моменты ген. совокупности. Оценка s^2 является смещённой оценкой ген. дисперсии. В качестве *оценки несмещённой* ген. дисперсии

принято использовать «исправленную» дисперсию

$$\hat{s}^2 = \frac{n}{n-1} s^2.$$

ЭМПИРИЧЕСКАЯ (ВЫБОРОЧНАЯ) ФУНКЦИЯ ПЛОТНОСТИ

статистическая оценка теоретической функции плотности случайной величины непрерывной ξ, характеризующая плотность частости попадания выборочных значений $x_1, x_2, ..., x_n$ в интервал $[x, x + \Delta x)$.

Э.ф.п. $\hat{f}^{(n)}(x)$ является первой производной *эмпирической функции распределения*

$$\hat{f}^{(n)}(x) = \hat{F}^{(n)\prime}(x) \approx \frac{\hat{F}^{(n)}(x + \Delta\tilde{o}) - \hat{F}^{(n)}(x)}{\Delta x}.$$

Условие её существования –непрерывность и дифференцируемость теоретической функции распределения. В этом случае на основе имеющейся выборки $x_1, x_2, ..., x_n$ можно построить Э.ф.п.: *вариационный ряд* $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, содержащий выборку $x_1, x_2, ..., x_n$, разбивают на k промежутков и подсчитывают *частоты* $\nu_{k(x)}$ попадания выборочных значений x_i в промежутки $\Delta_{k(x)}$. Тогда Э.ф.п. определяется по формуле: $\hat{f}^{(n)}(x) = \frac{\nu_{k(x)}}{n\Delta_{k(x)}}$, где $\frac{\nu_{k(x)}}{n}$ – частость попадания наблюдаемых значений непрерывной случайной величины ξ в *интервал группирования*. Э.ф.п. $\hat{f}^{(n)}(x) \ge 0$ – неотрицательная функция, принимающая нулевое значение при $x \le x_{(1)}$ и $x \ge x_{(n)}$. При увеличении числа наблюдении n и уменьшении длины интервала Δ Э.ф.п. стремится к теоретической функции плотности, а её графическое представление – *гистограмма* – приближается к кривой распределения.

ЭМПИРИЧЕСКАЯ (ВЫБОРОЧНАЯ) ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

статистическая оценка теоретической функции распределения *F(x)*, выражающая зависимость между значениями количественного признака и накопленной частотой. Значение Э.ф.р. $\hat{F}^{(n)}(x)$ для каждого наблюдения в *выборке* $x_1, x_2, ..., x_n$ соответствует накопленной относительной частоте события $x_i \le x$, т.е. равно относительному числу наблюдений в выборке, не превосходящих *x*. Э.ф.р. задается соотношением

$$\hat{F}^{(n)}(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{r}{n}, & x_{(r)} \le x < x_{(r+1)}, 1 \le k \le n-1, \\ 1, & x_{(n)} \le x \end{cases}$$

где $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ – *вариационный ряд*.

Э.ф.р. является *оценкой несмещённой*, асимптотически нормальной теоретической функции распределения и обладает всеми её свойствами. График Э.ф.р. – ломаная линия, в промежутках между соседними членами вариационного ряда $\hat{F}^{(n)}(x)$ сохраняет постоянное значение (см. рисунок). При переходе через точки оси x, равные членам выборки, $\hat{F}^{(n)}(x)$ претерпевает разрыв, возрастая на величину 1/n, а при совпадении l наблюдений – на l/n.

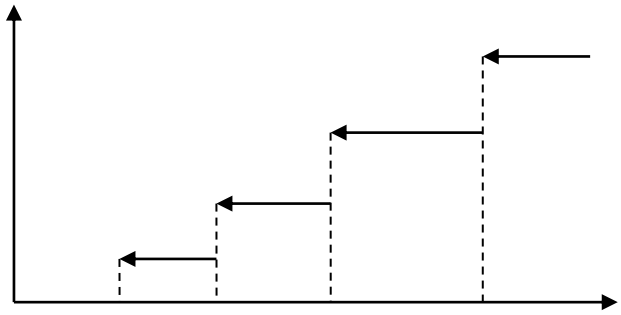


Рис. **Эмпирическая функция распределения**

Э.ф.р. при неограниченном увеличении объёма выборки $n \to \infty$ равномерно сходится к теоретической функции распределения, т.е.

$$P\left\{\lim_{n\to\infty}\sup_{x}\left|\hat{F}^{(x)}(x)-F(x)\right|\to 0\right\}=1.$$

Этот факт известен как теорема Гливенко. Для проверки согласия Э.ф.р. с полностью известной теоретической функцией распределения используются непараметрические *критерии согласия Колмогорова,* Колмогорова-Смирнова, *омега-квадрат*.

ЭМПИРИЧЕСКИЕ АНАЛОГИ НАЧАЛЬНЫХ МОМЕНТОВ

см. в ст. Начальные моменты выборочные

ЭМПИРИЧЕСКИЕ АНАЛОГИ ЦЕНТРАЛЬНЫХ МОМЕНТОВ

см. в ст. Центральные моменты выборочные

ЭМПИРИЧЕСКИЕ АНАЛОГИ ЦЕНТРА ГРУППИРОВАНИЯ

числовые характеристики центра группирования *ген. совокупности*, рассчитанные по выборочным данным $x_1, x_2, ..., x_n$. Э.а.ц.г. – среднее значение выборочное, выборочная *мода,* выборочная *медиана*. Выборочное среднее – наиболее употребительная характеристика центра группирования и обладает свойствами несмещённости и состоятельности. В зависимости от характера анализируемых данных выборочное среднее значение может рассчитываться по формуле средней арифметической, геометрической или гармонической. Мода служит хорошим показателем центра группирования, если переменная имеет категориальный характер. Применение моды оправданно и для неоднородных ген. совокупностей, кривая распределения которых имеет несколько вершин. Выборочная медиана рассчитывается легче, чем выборочное среднее или мода. Однако для приближённого нормального распределения медиана является менее точной оценкой центра группирования. Отметим, что в случае симметричной плотности вероятности и, в частности, для нормального распределения выборочное среднее, мода и медиана совпадают.

ЭФФЕКТИВНОСТЬ КРИТЕРИЯ АСИМПТОТИЧЕСКАЯ

понятие, позволяющее осуществлять в случае больших *выборок* количественное сравнение двух различных *критериев статистических* при проверке одной и той же *гипотезы статистической*. Существует несколько различных подходов к определению асимптотической эффективности. При распределении наблюдений P_θ, определяемое действительным параметром θ, требуется проверить гипотезу

$H_0 : \theta = \theta_0$ против $H_1 : \theta \neq \theta_0$.

Пусть некоторому критерию с уровнем значимости α требуется N_1 наблюдений для достижения мощности β против заданной альтернативы θ, а другому критерию того же уровня значимости требуется N_2 наблюдений. Тогда можно определить относительную эффективность первого критерия по отношению ко второму:

$e_{12}=N_2/N_1$.

Понятие относительной эффективности дает исчерпывающую информацию при сравнении критериев, но оказывается неудобным для применения, поскольку величина e_{12} является функцией трех аргументов: α, β и θ и обычно не поддается вычислению в явном виде. Поэтому используют предельные переходы. Т.о. Э.к.а., по Питмену – величина

$$\lim_{\theta\to\theta_0} e_{12}(\alpha,\beta,\theta)$$

при фиксированных α, β (если этот предел существует). Аналогично определяются Э.к.а. по Бахадуру, когда при фиксированных β и θ устремляют к нулю α, и по Ходжесу-Леману, когда при фиксированных α и θ вычисляют предел при $\beta \to 1$. Каждое из этих определений имеет свои достоинства и недостатки. Так, напр., Э.к.а. по Питмену вычисляется обычно легче, чем Э.к.а. по Бахадуру (вычисление средней связано с нетривиальной задачей изучения асимптотики вероятностей больших уклонений тестовых статистик), однако в ряде

случаев оказывается менее чувствительным средством для сравнения дух критериев.

ЭФФЕКТИВНОСТЬ ОЦЕНКИ

свойство оптимальности *оценок несмещённых*, которая характеризует разброс случайных значений оценки около истинного значения оцениваемого параметра. Среди всех несмещённых оценок более предпочтительной является та, значения которой теснее сконцентрированы около значения параметра. Пусть $\hat{\Theta}(x_1, x_2, ..., x_n)$ – произвольная несмещённая оценка параметра Θ, величина I(Θ) – количество информации о параметре Θ, содержащееся в одном наблюдении, nI(Θ) – количество информации о Θ, содержащееся в n независимых наблюдениях $x_1, x_2, ..., x_n$. При некоторых условиях регулярности имеет место неравенство Крамера – Рао, которое дает нижнюю границу для *дисперсии* несмещённой оценки. Тогда Э.о. определяется отношением нижней границы для дисперсии оценки к фактической дисперсии оценки

$$eff(\Theta^*) = \frac{1}{nI(\Theta) D\hat{\Theta}^*}$$

и удовлетворяет неравенствам

$$0 \le eff(\Theta^*) \le 1.$$

Если Э.о. стремится к 1 при неограниченном увеличении числа наблюдений *n*, т.е.

$$eff(\Theta^*) = \lim_{n\to\infty} \frac{1}{nI(\Theta) D\hat{\Theta}^*},$$

то имеет место асимптотическая эффективность оценок; для *оценки эффективной*

$$eff = 1.$$

Э.о. – решающее правило качества оценок неизвестного параметра. В случае несовместности требований несмещённости и эффективности предпочтительным является соблюдение условия Э.о. На практике не всегда удается получить в явном виде эффективные оценки, поэтому приходится использовать оценки, обладающие эффективностью менее 1. Понятие «Э.о.» введено Р.Фишером.

ЭФФЕКТИВНЫЙ КРИТЕРИЙ

критерий, имеющий наибольшую *вероятность* попадания в *критическую область* среди всех *критериев статистических* с заданной вероятностью *ошибки первого рода* α, предназначенных для проверки простой гипотезы H_0 против простой конкурирующей гипотезы H_1. Э.к. так же определяется, как статистический критерий, имеющий наименьшую вероятность *ошибки второго рода* β при проверке простой гипотезы H_0 против простой альтернативы H_1 с заданной

вероятностью ошибки первого рода. Иными словами Э.к. с наибольшей вероятностью отвергает проверяемую гипотезу Н, если она ошибочна, и с наименьшей вероятностью отвергает проверяемую гипотезу Н, если она верна. Наиболее Э.к. – *критерий Пирсона*, который используется при малых объёмах *выборки*. Вопросами выбора Э.к. занимались Дж. Нейман, Е. Пирсон.

См. также *Мощность критерия.*

Подраздел 2.2. Многомерный статистический анализ и эконометрика

Рубрика 2.2.1. Многомерные статистические методы

А

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ

разделение рассматриваемой совокупности объектов или явлений на однородные, в определённом смысле, группы, либо отнесение каждого из заданного множества объектов к одному из заранее известных классов (при этом классифицируемое заданное множество может состоять из одного объекта).

См. также *Классификация многомерных наблюдений без обучения.*

АНАЛИЗ ТИПОЛОГИЧЕСКИЙ

метод изучения сложных социально-экономических объектов, состоящий в выделении значимых, качественно отличных друг от друга, внутренне однородных групп объектов, характеризующихся совокупностью признаков произвольной природы. А.т. на эмпирическом уровне реализуется с помощью математических методов. Осн. понятия А.т. – «объект типологии», «основание типологии», «тип объекта», «априорная типология».

Объектом типологии называется совокупность тех свойств изучаемых объектов, которые позволяют рассматривать их как носителей определенных типов явлений. На эмпирическом уровне эта совокупность условно называется совокупностью типообразующих признаков. При осуществлении А.т. всю совокупность признаков, характеризующих объекты, можно разделить на три подсовокупности, функционально играющие различную роль: признаки, определяющие априорную типологию и тем самым описывающие объективные условия существования искомых типов; типообразующие признаки, описывающие само явление; признаки, участвующие в процессе интерпретации и объясняющих изучаемое явление. Основание типологии – совокупность содержательных предпосылок для определения того, какие объекты считать близкими, «похожими», однотипными, а какие далекими, «непохожими», разнотипными. Реализация этого понятия в каждом конкретном случае А.т. может быть неоднозначной. Одно и то же основание типологии часто можно формализовать различными способами. Формализация происходит при подготовке данных, при выборе математических методов анализа, а также при интерпретации полученных результатов.

Процесс А.т. структурно распадается на несколько этапов, обусловленных логикой исследователя. Этапы А.т.: 1. Построение априорной типологии. 2. Определение объекта типологии и тем самым типообразующих признаков; 3. Операционализация понятия «основание типологии». На следующих двух этапах происходит его формализация, в процессе которой само понятие может уточняться; 4. Построение признакового пространства на базе анализа свойств типообразующих признаков с учетом типа используемых шкал измерения. На этом этапе решаются проблемы нормирования, взвешивания признаков и выбора количественных характеристик типообразующих (классификационных) признаков. 5. Выбор формального аппарата классификации, т. е. способа разбиения объектов на однородные группы. Такой выбор осуществляется на основе предыдущих этапов, а также с учетом требований, которые обусловлены формализацией закономерностей, определяющих искомые классы. При этом предполагается комплексное использование различных процедур классификации (см. ст. Классификация многомерных наблюдений), каждая из которых выявляет различные, структурные особенности исходных данных. 6. Интерпретация результатов: выбор методов описания классов, анализ возможных причин несовпадения типологии и классификации, обеспечение условий перехода от формальной классификации к содержательной типологии объектов. В результате реализации этого этапа, в частности, может видоизмениться априорная типология, формальная классификация (посредством объединения или разделения отдельных классов) и т. д. Все этапы А.т. взаимосвязаны

между собой так, что в процессе реализации одного из них может происходить переосмысление других и, соответственно, возможны коррекции даже в дефинициях основных понятий А.т.

АНАЛИЗ ДАННЫХ

область математики и информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных (в широком смысле) данных.

Первоначально А.д. практически сводился к прикладным разделам математической статистики, поскольку объектом анализа являлись в основном случайные выборки из ген. совокупности. Прикладная математическая статистика – наиболее обширный раздел А.д. Вместе с тем необходимость исследования больших массивов данных, не являющихся случайными выборками, напр., содержимого *баз данных* и данных в сети Интернет привели к созданию других подходов, среди которых *DM-методы*, математическая морфология, связывающая методы количественного анализа и визуального представления данных, теория интерпретации эксперимента, сближающая традиционный А.д. и математическое моделирование.

Последовательность действий, которую необходимо выполнить для А.д. (получения знания) получила название поиска знаний в базах данных (от англ. – knowledge discovery in databases (KDD)). Данная методика не зависит от предметной области, это набор атомарных операций, комбинируя которые, можно получить нужное решение (см. рис.1). Алгоритм KDD состоит из пяти этапов.

1. Выборка данных. Первый шаг в анализе – получение исходной выборки для построения модели. На этом шаге необходимо активное участие эксперта для выдвижения гипотез и отбора факторов, влияющих на анализируемый процесс. В качестве источника чаще всего рекомендуется использовать специализированное *хранилище данных*, агрегирующее всю необходимую для анализа информацию.

2. Очистка данных. Необходимость предварительной обработки при анализе данных возникает независимо от того, какие технологии и алгоритмы используются. Более того, эта задача может представлять самостоятельную ценность в областях, не имеющих непосредственного отношения к анализу данных. К задачам очистки данных относятся: заполнение пропусков, редактирование аномалий, сглаживание, обнаружение дубликатов и противоречий и прочие.

3. Трансформация данных. Различные алгоритмы анализа требуют специальным образом подготовленные данные. К задачам трансформации данных относятся: скользящее среднее, приведение типов, выделение временных интервалов, преобразование непрерывных значений в дискретные и наоборот, сортировка, группировка и прочее.

Рис. 1 **Методика Knowledge Discovery in Databases (KDD)**

4. DM-методы – процесс обнаружения в «сырых» данных ранее неизвестных нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других. Задачи, решаемые DM-методами: классификация; регрессия; кластеризация; ассоциация (нахождение зависимости, что из события X следует событие Y); последовательные шаблоны (установление закономерностей между связанными во времени событиями) и др. Наиболее популярные алгоритмы: *деревья решений*; искусственные *нейронные сети* (многослойный перцептрон, обучение при помощи алгоритма обратного распространения ошибки); линейная регрессия (классическая линейная модель); *самоорганизующиеся карты Кохонена*; ассоциативные правила (алгоритм APriori).

Т.к. DM-метод развивался и развивается на стыке таких дисциплин, как статистика, теория информации, машинное обучение, теория баз данных, вполне закономерно, что большинство алгоритмов DM-методов разработаны на основе различных методов этих дисциплин. В общем случае, не принципиально, каким именно алгоритмом будет решаться одна из задач DM-методов – главное иметь метод решения для каждого класса задач.

5. Интерпретация. В случае, когда извлеченные знания непрозрачны для пользователя, должны существовать методы обработки, позволяющие привести их к интерпретируемому виду. Для оценки качества полученной модели используют как формальные методы, так и знания эксперта. Полученные модели являются, по сути, формализованными знаниями эксперта, а следовательно их можно тиражировать. Найденные знания должны быть применимы и на новых данных с некоторой степенью достоверности.

Использование методов построения моделей позволяет получать новые знания, которые невозможно извлечь другим способом. Кроме того, полученные результаты –формализованное описание некоего процесса, а, следовательно, поддаются автоматической обработке. Недостатком же является то, что такие методы более требовательны к качеству данных, знаниям эксперта и формализации самого изучаемого процесса. К тому же почти всегда имеются случаи, не укладывающиеся ни в какие модели.

На практике подходы комбинируются, напр., *визуализация данных* наводит эксперта на некоторые идеи, которые он пробует проверить при помощи различных способов построения моделей, а результаты моделирования подаются на

вход механизмам визуализации. Полнофункциональная система анализа не должна замыкаться на применении только одного подхода или одной методики анализа. Механизмы визуализации и построения моделей должны дополнять друг друга. Макс. отдачу можно получить, комбинируя методы и подходы к анализу данных.

Б

БЛОЧНАЯ МАТРИЦА

матрица, разбитая на матрицы (блоки) меньших размеров, которые рассматриваются как её элементы.

Пусть $A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ Б.м. порядка $m\times n$, где A_{11}, A_{12}, A_{21}, A_{22} – подматрицы размера соответственно $m_1\times n_1$, $m_1\times n_2$, $m_2\times n_1$, $m_2\times n_2$, где $m_1+m_2=m$, $n_1+n_2=n$.

Если внедиагональные блоки A_{12}, A_{21} равны нулю, а подматрицы, расположенные на главной диагонали ненулевые, то такая матрица называется блочно-диагональной:

$$A=\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}.$$

Операции над Б.м.: при сложении Б.м. необходимо, чтобы подматрицы были одного размера: если матрица B разбита аналогично матрице A, то сумма матриц определяется:

$$A+B=\begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ A_{21}+B_{21} & A_{22}+B_{22} \end{pmatrix};$$

при умножении Б.м. на число λ каждая подматрица умножается на это число λ; при транспонировании Б.м. подматрицы на главной диагонали остаются на месте и транспонируются, остальные – транспонируются и меняются местами –

$$A^T=\begin{pmatrix} A_{11}^T & A_{21}^T \\ A_{12}^T & A_{22}^T \end{pmatrix};$$

пусть матрица C размером $m\times p$, разбита на подматрицы т.о., что подматрицы C_{11} и C_{12} имеют n_1 строк, а C_{21} и C_{22} – n_2 строк. Тогда матрицу A можно умножить справа на матрицу C:

$$AC=\begin{pmatrix} A_{11}C_{11}+A_{12}C_{21} & A_{11}C_{12}+A_{12}C_{22} \\ A_{21}C_{11}+A_{22}C_{21} & A_{21}C_{12}+A_{22}C_{22} \end{pmatrix}.$$

Следовательно, при перемножении Б.м. необходимо согласование размеров подматриц.

В

ВАРИМАКС-МЕТОД

наиболее распространенный способ ортогонального поворота системы координат, относящийся к проблеме вращения в факторном анализе. Заключается в выборе углов поворота m-мерной системы координат и служит для проведения содержательной интерпретации. Основан на том, что изменение матрицы факторных нагрузок не приводит к изменению редуцированной по этой матрице выборочной *корреляционной матрицы* исходных признаков. Ортогональное вращение – наиболее простое. Оно производится умножением матрицы факторных нагрузок на некоторую ортогональную матрицу T, задающую угол поворота, размерностью m x m по числу гл. компонент. Поворот может задаваться по часовой стрелке или против нее. Напр., для матрицы факторных нагрузок A с числом общих факторов $m=2$ можно записать:

$$T=\begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$$

Вращение по часовой стрелке

$$T=\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$

Вращение против часовой стрелки

Если матрица факторных нагрузок содержит данные более чем по двум латентным факторам, строится несколько матриц преобразования T для всех парных комбинаций факторов. Так, для трёхмерной матрицы A будут исполь-

зоваться три матрицы преобразований T (вращение против часовой стрелки):

$$T_{12} = \begin{pmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}; \quad T_{13} = \begin{pmatrix} \cos\varphi & 0 & -\sin\varphi \\ 0 & 1 & 0 \\ \sin\varphi & 0 & \cos\varphi \end{pmatrix}; \quad T_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{pmatrix}.$$

Полная матрица преобразования для трехмерного случая: $T = T_{12} \times T_{13} \times T_{23}$. Этот подход обобщается на случай, когда число общих факторов $r > 3$, для четырёхмерной матрицы A полная матрица преобразования T будет произведением уже шести матриц вращения, для всех пар общих факторов: $T = T_{12} \times T_{13} \times T_{14} \times T_{23} \times T_{24} \times T_{34}$ и т.д. При условии ортогонального вращения всегда $T^T T = E$, где E – единичная матрица. Предложены различные методы вращения факторов. Цель вращения – повернуть факторы так, чтобы выбрать простейшую для интерпретации факторную структуру. Термин простая структура придуман Трестоуном в 1947 для описания условий, когда факторы отмечены высокими нагрузками на некоторые переменные и низкими для других. Наиболее стандартные вычислительные методы вращения для получения простой структуры – В.-м. вращения, предложенный Кайзером в 1958. Вращение имеет целью максимизацию дисперсии квадратов исходных факторных нагрузок по переменным для каждого фактора, что эквивалентно максимизации дисперсий в столбцах матрицы квадратов исходных факторных нагрузок. Другие методы, предложенные Харманом в 1967 – методы квартимакс, биквартимакс и эквимакс.

ВЕРОЯТНОСТЬ СЛУЧАЙНОГО СОБЫТИЯ A

объективная мера возможности наступления события A, удовлетворяющая условию $0 \le P(A) \le 1$. Классическое определение: вероятность любого события A можно определить по формуле $P(A) = M/N$, где N – число всех простейших возможных исходов, а $M = M_A$ – число возможных случаев, благоприятствующих наступлению случайного события A, причём $M_A \le N$. Здесь вероятности элементарных событий считаются равными 1/N. Геометрическое определение вероятности – обобщение классического определения. Геометрически вероятность P(A) определяется как отношение длин, пл., объёмов и т.д. При этом используется обобщение равновозможности элементарных исходов на вероятности исходных областей.

Статистическое определение вероятности заключается в том, что за вероятность события A принимается относительная частота (частость) или число, близкое к ней: $P(A) \cong \frac{m}{n}$, где m – общее число появлений события A в n независимых практически одинаковых испытаниях, в каждом из которых может произойти или не произойти событие A; кроме того n должно быть достаточно большим, а частость по физической природе устойчивой.

Классическое и статистическое определения вероятности в совокупности до некоторой степени компенсируют друг друга и лишены недостатков, присущих им в отдельности. Точным, строгим с математической точки зрения является аксиоматическое определение вероятности. Такое построение *теории вероятностей* опирается на теорию меры и интегрирования, и исходит из некоторого списка не определяемых формально осн. понятий и аксиом, на основе которого все дальнейшие понятия отчётливо определяются, дальнейшие предложения доказываются. В теории вероятностей принята система аксиом, сформулированная акад. А.Н.Колмогоровым.

Заметим, что операции определяемые в алгебре событий: сумма, произведение событий и взятие противоположного события соответствуют в алгебре множеств операциям объединения, пересечения и дополнительного множества; кроме того достоверное событие соответствует универсальному множеству, а невозможное событие соответствует пустому множеству.

ВЗВЕШЕННОЕ ЕВКЛИДОВО РАССТОЯНИЕ

способ (мера, метрика) нахождения расстояния между объектами в задачах *кластерного анализа*. Эти расстояния могут определяться в одномерном или многомерном пространстве. Выбор метрики (расстояния) является узловым моментом в задачах автоматической классификации, от которого зависит окончательный вариант разбиения объектов на *классы* при заданном алгоритме разбиения. При этом решение данного вопроса зависит в основном от целей исследования, физической и статистической природы вектора наблюдений X, полноты априорных сведений о характере вероятностного распределения X. В.е.р. применяется в тех случаях, когда каждой компоненте x_l вектора наблюдений X удается приписать некоторый «вес» w_l пропорциональный степени важности признака в задаче классификации:

$$d_{BE}(X_i, X_j) = \sqrt{\sum_{k=1}^{m} w_k \left(x_{ik} - x_{jk}\right)^2},$$

где $d_{BE}(X_i, X_j)$ – расстояние между i-м и j-м объектами; x_{ik}, x_{jk} – значение k-й переменной у i-го и j-го объекта ($k = 1,2,\ldots, m$; $i, j = 1,2,\ldots, n$). Обычно принимают $0 \le w_k \le 1$, где $k=1,2,\ldots,m$. Выбор меры расстояния и весов для классифицирующих переменных – очень важный этап кластерного анализа, так как от этой процедуры зависят состав и количество формируемых кластеров, а также степень сходства объектов внутри кластеров. Вопрос о придании переменным соответствующих весов должен решаться после проведения дополнительных исследований, напр., опроса экспертов и обработкой их мнений. Веса задаются пропорционально степени важности переменных. Определение весов w_l только по данным *выборки* может привести к ложным выводам.

ВЗВЕШИВАНИЕ ВЫБОРОЧНЫХ ДАННЫХ

в общем случае наблюдению x_i выборочных данных $x_1, x_2, \ldots, x_n$ приписывается вес $w_i = w(x_i) \ge 0$, который определяется как некоторая функция от его текущего значения. Обычно веса подчиняются условию нормировки

$$\sum_{i=1}^{n} w(x_i) \equiv 1.$$

В частности, можно рассматривать w-взвешенные моменты случайной величины ξ с плотностью $f_\xi(x)$, как выборочные $\hat{m}_k(n,W)$, так и теоретические $m_k(W)$

$$\hat{m}_k(n,W) = \sum_{i=1}^{n} x_i^k w(x_i);$$

$$m_k(W) = \int v^k w(x) f_\xi(x) dx.$$

Под W понимается вектор весов $w(x_1), w(x_2), \ldots, w(x_n)$ в выражении для выборочных моментов и функция со значением $w(x)$ в выражении для теоретических моментов.

Если имеют дело с результатами наблюдения одномерной случайной величины $x_1, x_2, \ldots, x_n$, то часто вес наблюдения x_i определяются в зависимости от его порядкового номера в упорядоченном (по возрастанию) ряду наблюдений и каждому члену *вариационного ряда* x_i ставят в соответствие некоторый вес w_i.

В.в.д. позволяет повысить репрезентативность выборочной совокупности. Способ подбора весов зависит от наличия дополнительной информации. В случае, когда известно ген. распределение одного или нескольких признаков, можно определить веса как отношение значений ген. и выборочной функций плотности (законов распределения) признака. Можно «настроиться» на ген. характеристики совокупности, такие как среднее значение, мода, медиана и прочие.

Предварительно проводится группировка ген. и выборочной совокупностей, причём, необходимо подобрать интервалы т.о., чтобы избежать наличия пустых и перенасыщенных интервалов. Далее выборочным наблюдениям присваиваются веса, такие, что при умножении выборочной частости на соответствующий вес, получается ген. частость. Другими словами, искомые веса определяются как отношения ген. и выборочной частостей для каждого интервала.

ВРАЩЕНИЯ ФАКТОРОВ ПРОБЛЕМА

вращение общих факторов в *факторном анализе* с целью улучшения их интерпретируемости. Конечная цель факторного анализа – получение содержательно интерпретируемых факторов, которые воспроизводили бы выборочную корреляционную матрицу между переменными. Это достигается путём вращения общих факторов, которое может быть ортогональным, когда взаимодействие факторов исключается, и косоугольным, порождающим корреляционные связи латентных факторов. При выборе каждого из двух типов вращения требуется последовательное решение таких вопросов: какие вычислительные процедуры следует выполнить для вращения факторного пространства; сколько раз должна повторяться операция вращения; какой угол лучше установить для поворота. Вопрос достаточности числа поворотов пространства факторов решается путём построением графиков распределений наблюдаемых объектов в пространстве повернутых факторов или с использованием специальных критериев для оценки структуры общих факторов. Все критерии конструктивно базируются на представлении величины дисперсии факторных нагрузок как меры сложности структуры факторов. В ортогональном преобразовании факторного пространства наиболее популярны критерии: квартимакс – вычисляется по формуле:

$$q = \sum_{j=1}^{m} \frac{\sum_{k}^{r} (a_{jk})^4 - \left(\sum_{k=1}^{r} a_{jk}^2 \right)^2}{r^2},$$

где a_{jk} – элементы матрицы факторного отображения, величины факторных нагрузок; r – число общих факторов.

Его использование предусматривает вращение факторных осей т.о., чтобы величины факторных нагрузок максимизировали q, одновременно учитывается качество структуры всех участвующих в анализе общих факторов. *Варимакс-метод* – рассчитывается V_r критерий качества структуры каждого фактора:

$$V_r = \frac{m\sum_{j=1}^{m} a_{jr}^4 - \left(\sum_{j=1}^{m} a_{jr}^2 \right)^2}{m^2}.$$

В данном случае достигают макс. упрощения в описании столбцов матрицы факторного отображения. Возможно раздельное улучшение структуры факторов. Наилучшим также будет макс. значение критерия, как и в предыдущем случае. Если в анализе используется косоугольное вращение пространства факторов, то наиболее часто пользуются критериями облимакс, квартимин, облимин. Вращение общих факторов даёт возможность максимально сократить разброс субъективных суждений относительно их названий, неизменно присутствующих при их содержательной интерпретации.

Г

ГЛАВНЫЕ КОМПОНЕНТЫ

обобщённые показатели, построенные на основе исходных признаков. Обычно исходные признаки весьма существенно коррелируют между собой. Это затрудняет проведение исследований, т.к. большинство многомерных статистических методов предполагает, по крайней мере, неявно, некоррелированность признаков. Предпочтительнее разрабатывать методы, учитывающие коррелированность признаков, или преобразовывать исходное косоугольное пространство в ортогональное. Вторую идею реализует *метод гл. компонент* (МГК).

Сначала на основе матрицы исходных признаков X строят соответствующую матрицу стандартизованных признаков Z. Затем по Z строят *корреляционную матрицу*: R= (Z'*Z)/n, которая и служит основой МГК. Для однозначности полученного решения налагается дополнительное условие: упорядочение по убыванию дисперсий Г.к. Метод множителей Лагранжа преобразует задачу поиска условного экстремума в задачу поиска безусловного экстремума. А эта, в свою очередь, сводится к задаче ортогонализации пространства переходом к системе собственных векторов матрицы R. В результате решения проблемы собственных чисел и собственных векторов строятся две матрицы: диагональная матрица собственных чисел (Λ) и ортогональная матрица собственных векторов (U). Далее определяется матрица нагрузок: $A=U*\Lambda^{0.5}$,

элементы которой [A={a $_{j\nu}$ }; j,ν =1, … ,k] являются *коэффициентами парной корреляции* между исходными признаками (расположенными по строкам) и построенными главными компонентами (расположенными по столбцам). a $_{j\nu}$ =r $_{X_jF_\nu}$. Это позволяет содержательно интерпретировать первые наиболее весомые Г.к. Кроме того, можно объяснить связь между исходными признаками, как следствие их связи с Г.к.

Далее строится матрица индивидуальных значений Г.к. на объектах: F=Z*U. Обобщённые показатели Г.к. располагаются по столбцам этой матрицы. Они являются ортогональными (некоррелированными) центрированными величинами, с дисперсиями, равными соответствующим собственным числам. Это позволяет успешно использовать Г.к. при классификации объектов или при построении уравнения регрессии (с дальнейшим пересчетом в исходные признаки). На практике используются несколько первых наиболее весомых Г.к.

Большинство реальных статистических исследований матрицы данных: «объект-признак» выполняется с использованием МГК, чему способствует наличие программ во всех статистических пакетах прикладных программ. Надо учитывать, что при составлении программ разработчики могли внести модификации, напр., опираться не на корреляционную, а на *ковариационную матрицу*, или включить атрибут факторного анализа: возможность вращения матрицы нагрузок для улучшения интерпретации, или при построении матрицы F использовать формулу: F=Z*A, вместо указанной, и т.д.

ГЛАВНЫХ КОМПОНЕНТ АНАЛИЗ

один из наиболее распространённых в статистической практике методов снижения размерности исследуемого признакого пространства, ориентированный на выявление сравнительно небольшого числа таких обобщённых вспомогательных показателей (каждый из которых строится в виде нормированной линейной комбинации исходных переменных), которые обнаруживают наибольшую изменчивость (наибольший разброс) при переходе от одного «носителя» анализируемых свойств к другому.

В частности, первой *гл. компонентой* $z^{(1)}(X)$ исследуемой системы показателей $X=(x^{(1)},\ldots,x^{(p)})$ называется такая нормированно-центрированная линейная комбинация этих показателей, которая среди всех прочих нормированно-центрированных линейных комбинаций переменных $x^{(1)},\ldots,x^{(p)}$ обладает наибольшей дисперсией. И далее: k-й гл. компонентой $(k=2,\ldots,p)$ исследуемой системы показателей X называется такая нормированно-центрированная линейная комбинация этих показателей, которая не коррелирована с $k-1$ предыдущими гл. компонентами и среди всех прочих нормированно-центрированных и не коррелированных с предыдущими $k-1$ гл. компонентами линейных комбинаций переменных $x^{(1)},\ldots,x^{(p)}$ обладает наибольшей дисперсией.

В оптимизационной постановке задачи снижения размерности решение, получаемое с помощью *метода гл. компонент*, максимизирует критерий информативности, определяемый суммарной дисперсией заданного (небольшого) числа искомых вспомогательных переменных (при соответствующих условиях их нормировки). Для вычисления *k*-й гл. компоненты $z^{(k)}(X)\,(k=1,\ldots,p)$ следует найти собственный вектор $l_k=(l_{k1},\ldots,l_{kp})$ ковариационной матрицы $\sum$ исходного набора показателей $X=(x^{(1)},\ldots,x^{(p)})$, т.е. решить систему уравнений:

$$\left(\Sigma-\lambda_k\, I\right) l_k' = 0,$$

где λ_k – *k*-й по величине корень (при их расположении в порядке убывания) характеристического уравнения:

$$\left|\Sigma-\lambda I\right|=0.$$

Компоненты $l_{kj}\,(j=1,p)$ собственного вектора l_k – искомые весовые коэффициенты, с помощью которых осуществляется переход от исходных показателей $x^{(1)},\ldots,x^{(p)}$ к гл. компоненте $z^{(k)}(X)$, т.е. $z^{(k)}(X)=l_k(X-a)'$, где $a=(a^{(1)},a^{(2)},\ldots,a^{(p)})$ – вектор средних значений компонент вектора *X* (т.е. $a^{(j)}=Ex^{(j)}$).

Осн. числовые характеристики вектора $Z=(z^{(1)},\ldots,z^{(p)})'$ гл. компонент выражаются

через осн. числовые характеристики исходных показателей и собственные числа их ковариационной матрицы $\sum$. В частности:

$$EZ = 0$$

$$\Sigma_z = E(Z \cdot Z') = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \ddots \\ 0 & & \lambda_p \end{pmatrix}$$

$$\sum_{j=1}^{p} Dz^{(j)} = \sum_{j=1}^{p} Dx^{(j)} = \sum_{j=1}^{p} \lambda_j; \ |\Sigma_z| = |\Sigma|$$

,

где Σ_z – ковариационная матрица вектора Z.

Вектор $p'\,(p' < p)$ первых гл. компонент $Z^{(p')}(X) = (z^{(1)}(X), \ldots, z^{(p')}(X))'$ обладает рядом экстремальных свойств: свойство наименьшей ошибки автопрогноза или наилучшей самовоспроизводимости: с помощью p' первых гл. компонент $z^{(1)}, \ldots, z^{(p')}$ исходных показателей $x^{(1)}, \ldots, x^{(p)}$ $(p' < p)$ достигается наилучший (в определённом смысле) прогноз этих показателей среди всех прогнозов, которые можно построить с помощью p' линейных комбинаций набора из p произвольных признаков; свойство наименьшего искажения некоторых геометрических характеристик совокупности исходных многомерных наблюдений $X_1, \ldots, X_n$ при их проецировании в пространство меньшей размерности, натянутое на p' первых гл. компонент.

Гл. компоненты, построенные не по истинной ковариационной матрице Σ вектора исходных показателей $X = (x^{(1)}, \ldots, x^{(p)})$, а по её выборочному аналогу (оценке) $\hat{\Sigma}$ называются выборочными гл. компонентами и в определённых (достаточно широких) условиях обладают (вместе с собственными числами и векторами матрицы $\hat{\Sigma}$) всеми традиционными свойствами «хороших» оценок: состоятельностью, асимптотической эффективностью, асимптотической нормальностью. Однако в условиях растущей размерности, т.е. в «асимптотике А.Н. Колмогорова», анализируемые выборочные характеристики могут вести себя некоторым специальным образом.

Геометрически определение первой гл. компоненты равносильно построению новой координатной оси $Oz^{(1)}$ т.о., чтобы она шла в направлении наибольшего разброса исходных данных, т.е. – в направлении вытянутости анализируемого «облака» многомерных наблюдений. Затем среди направлений, перпендикулярных к $Oz^{(1)}$, отыскивается направление «наибольшей вытянутости» $Oz^{(2)}$ и т.д. Очевидно, если характер вытянутости анализируемого «облака» данных в исходном признаковом пространстве существенно отличен от линейного, то линейная модель гл. компонент может оказаться неэффективной. В подобных ситуациях обращаются к нелинейным версиям метода гл. компонент.

Гл. компоненты используются при решении осн. типов задач анализа данных: 1. упрощение, сокращение размерностей анализируемых моделей статистического исследования зависимостей или классификации с целью облегчения счёта и интерпретации получаемых статистических выводов; 2. наглядное представление (визуализация) исходных многомерных данных, получаемое с помощью их проецирования в пространство, натянутое на первую, первые две или первые три гл. компоненты; 3. предварительная ортогонализация объясняющих переменных в задачах построения регрессионных зависимостей как средство «борьбы» с *мультиколлинеарностью*; 4. сжатие объёмов хранимой статистической информации; 5. построение агрегированных показателей, являющихся измерителями различных синтетических латентных категорий, таких как уровень коррупции, уровень инновационного развития, качество жизни нас. и т.п.

ГРЕБНЕВАЯ РЕГРЕССИЯ

(от англ. – ridge – хребет, гребень) – один из подходов к оцениванию параметров множественной линейной регрессии в условиях *мультиколлинеарности*, позволяющий получить оценки с меньшей дисперсией и большей устойчивостью по сравнению с оценками *метода наименьших квадратов*; основана на использовании гребневых оценок или ридж-оценок.

Г.р. связана с отказом от свойства несмещённости, характерного для оценок *метода наименьших квадратов*, т.е. повышение устой-

чивости оценок достигается за счёт перехода к оценкам с небольшим смещением.

В общем виде ридж-оценка вектора параметров регрессии имеет вид:

$$b(\Delta) = (X^T * X + \Delta)^{-1} * X^T * Y ,$$

где Y – вектор значений зависимой переменной; X – матрица значений независимых переменных; Δ – симметричная неотрицательно определённая матрица.

Осн. часть применения Г.р. заключается в правильном выборе матрицы Δ. Часто матрицу Δ выбираю диагональной с элементами, пропорциональными диагональным элементам матрицы плана $X^T * X$ или в виде $\Delta = \delta * E$, где E – единичная матрица, а δ – параметр, который обычно выбирают равным от 0,1 до 0,4. В этом случае добавление к диагональным элементам матрицы $X^T * X$ небольшого положительного числа – «гребня» δ преобразует эту матрицу в условиях мультиколлинеарности из слабо обусловленной в хорошо обусловленную, а получаемые при этом оценки – в смещённые.

ГРЕКО-ЛАТИНСКИЙ КВАДРАТ

план эксперимента, предназначенный для исследования влияния на результативный показатель четырёх факторов, каждый из которых имеет n уровней. План этого типа позволяет в n^2 раз снизить объём наблюдений по сравнению с четырехфакторным дисперсионным анализом. При этом предполагается отсутствие влияния взаимодействий факторов на результативный показатель.

Г.-л.к. получается путём наложения на *латинский квадрат* другого латинского квадрата такой же размерности n×n и «ортогонального» первому. В данном случае ортогональность означает, что каждая буква как одного, так и другого латинского квадрата только один раз встречалась на каждой строке и в каждом столбце. Обычно во втором латинском квадрате употребляются греческие буквы, отсюда и название Г.-л.к. (см. табл.).

Таблица

Уровни фактора I	Уровни фактора II			
	1	2	3	4
1	Aδ	Dβ	Bα	Cγ
2	Bγ	Cα	Aβ	Dδ
3	Cβ	Bδ	Dγ	Aα
4	Dα	Aγ	Cδ	Bβ

В табл. уровни третьего фактора обозначаются латинскими буквами, записанными в каждой клетке, а четвёртого – греческими буквами.

В данном случае игнорируется влияние на результативный показатель X значительно большего числа взаимодействий факторов, чем в плане типа латинский квадрат. Негативных последствий этого влияния, приводящих к увеличению остаточного отклонения среднего квадратического, следует ожидать гораздо чаще при использовании планов типа Г.-л.к.

Проверка влияния факторов осуществляется с помощью *F*-критерия.

Д

ДЕНДРОГРАММА

(от греческого dendron – «дерево») – древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров. Д. описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения кластеров. С построением Д. связаны иерархические алгоритмы, которые являются результатом иерархического *кластерного анализа*. Д. также называют древовидной схемой, деревом объединения кластеров, деревом

иерархической структуры. Д. – вложенная группировка объектов, которая меняется на различных уровнях иерархии. Существует много способов построения Д.: объекты могут располагаться вертикально или горизонтально. В горизонтальной Д. объекты располагаются вертикально слева, результаты кластеризации – справа. Значения расстояний или сходства, отвечающие строению новых кластеров, изображаются по горизонтальной прямой поверх Д. (см. рис.1).

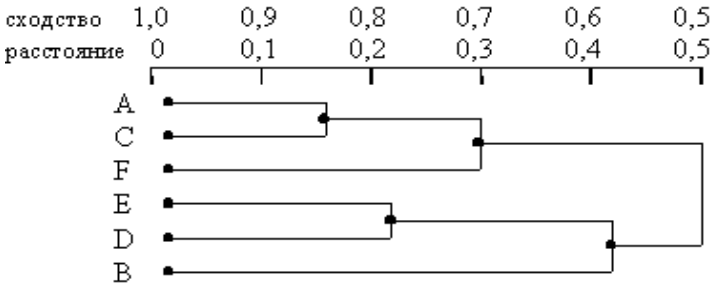


Рис. 1**. Горизонтальная дендрограмма**

На рис. 1 видно, что схема соответствует случаю шести объектов (n=6) и k характеристик (признаков). Объекты A и C наиболее близки и поэтому объединяются в один кластер на расстоянии, равном 0,1. Объекты D и E объединяются на расстоянии 0,2. Имеются 4 кластера: (A, C), (F), (D, E), (B). Далее образуются кластеры (A, C, F) и (E, D, B), соответствующие расстояниям объединения, равным 0,3 и 0,4. Окончательно все объекты группируются в один кластер на расстоянии 0,5. Вид Д. зависит от выбора метрики расстояния между объектом и кластером, а также метода кластеризации. Наиболее важным моментом является выбор меры сходства или меры расстояния между объектами и кластерами.

ДЕТЕРМИНАНТ (ОПРЕДЕЛИТЕЛЬ) МАТРИЦЫ

одна из числовых характеристик квадратной матрицы A. Обозначается как det A или |A|. Определитель n × n-матрицы A вычисляется по формуле:

$$\det A = \sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\dots\sum_{j_n=1}^{n}(-1)^{\nu(j_1,j_2,\dots j_n)}a_{1j_1}a_{2j_2}\dots a_{nj_n},$$

где суммирование ведётся по всем возможным комбинациям различных столбцов (т.е. по всем возможным перестановкам вторых индексов), а $\nu(j_1, j_2,..., j_n)$ – миним. число инверсий (т.е. парных обменов местами), которое надо совершить с элементами исходной перестановки (1, 2, …, *n*), чтобы получить перестановку $(j_1, j_2,..., j_n)$. Очевидно, общее число слагаемых в правой части составит при таком определении n-факториал (n!). Для матриц малых размерностей это определение приводит к результатам:

а) $n = 2$: $\det A = \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}$;

б) $n = 3$: $\det A = \det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} =$

$a_{11}a_{22}a_{33} + a_{13}a_{21}a_{32} + a_{12}a_{23}a_{31} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}$.

Важные свойства определителя матрицы:

1. $\det(\mathbf{AB}) = \det(\mathbf{BA}) = \det\mathbf{A} \cdot \det\mathbf{B}$;

2. $\det(\lambda\mathbf{A}) = \lambda^{n} \det\mathbf{A}$ (λ - число, n - размерность матрицы $\mathbf{A}$);

3. $\det[\mathrm{diag}(a_{11}, a_{22}, ..., a_{nn})] = a_{11}a_{22} \cdot ... \cdot a_{nn}$;

4. $\det\mathbf{I}_n = 1$; 5. $\det\mathbf{A}^{T} = \det\mathbf{A}$;

6. $\det\mathbf{A} = 0$, если в матрице А есть две одинаковые строки (два одинаковых столбца);

7. инверсия (обмен местами) двух строк (столбцов) матрицы **A** приводит к изменению знака её определителя;

8. значение определителя матрицы **A** не изменится, если к любой её строке (столбцу) добавить линейную комбинацию других строк (столбцов);

9. разложение определителя по элементам строки (или столбца): $\det\mathbf{A} = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} \det\mathbf{A}_{ij}$, где $\mathbf{A}_{ij}$ – $(n-1)\times(n-1)$ – матрица, получающаяся из матрицы **A** вычеркиванием из нее *i*-й строки и *j*-го столбца. Величина $\det\mathbf{A}_{ij}$ называется минором матрицы **A**, а величина $(-1)^{i+j}\det\mathbf{A}_{ij}$ – алгебраическим дополнением элемента a_{ij} в матрице **A**.

Если квадратная матрица имеет отличный от нуля определитель, то она называется невырожденной.

ДЕТЕРМИНИСТСКАЯ МОДЕЛЬ

(от лат. determino – определять, modus – образ) – модель «строго определённая», предполагающая однозначное и предсказуемое поведение моделируемой системы в данных условиях. Параметры и переменные Д.м. являются неслучайными величинами. Для её построения требуется знание поведения и взаимодействия всех элементов моделируемой системы, что возможно лишь в относительно простых условиях. В отсутствие неопределённости наглядность и простота использования Д.м. делает её привлекательной для анализа, и с её помощью можно получать достаточную информацию для принятия решения. При многовариантности поведения элементов системы и внешних условий возможна вариация конструкции модели и исходных данных, позволяющая изучать устойчивость и надежность получаемых решений. В реальных технических, биологических, экономических, социальных и других сложных системах присутствует значительная неопределенность, что делает практически невозможным адекватное представление этих систем Д.м., поэтому исследователи значительно чаще прибегают к стохастическим моделям.

ДИСКРИМИНАНТНАЯ ФУНКЦИЯ

линейная функция, используемая в *дискриминантном анализе* для оптимального разделения наблюдений в рассматрива-емые группы. Канонической Д.ф. называется линейная функция: $d_{km} = \beta_0 + \beta_1 x_{1km} + \beta_2 x_{2km} + ... + \beta_p x_{pkm}$, где: d_{km} – значение канонической Д.ф. для *m*-го объекта в группе *k* ($m = 1, ..., n$, $k = 1, ..., g$); x_{ikm} – значение дискриминантной переменной x_i для *m*-го объекта в группе *k*; $\beta_0, ..., \beta_p$ – коэффициенты Д.ф. С геометрической точки зрения Д.ф. определяют гиперповерхности в *p*-мерном пространстве. В частном случае при $p=2$ она является прямой, а при $p=3$ – плоскостью. Коэффициенты β_i первой канонической Д.ф. выбираются т.о., чтобы центроиды (средние значения) различных групп как можно больше отличались друг от друга. Коэффициенты второй группы выбираются также, т.е. соответствующие средние значения должны максимально отличаться по классам, при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой. Аналогично третья функция должна быть некоррелирована с первыми двумя и т.д. Отсюда следует, что любая каноническая Д.ф. d имеет нулевую внутригрупповую корреляцию с $d_1, d_2, ..., d_{g-1}$. Если

число групп равно g, то число канонических Д.ф. будет на единицу меньше числа групп. Классификация переменных будет осуществляться тем лучше, чем меньше расстояние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп. Следует отметить, что большая внутригрупповая вариация нежелательна, так как в этом случае любое заданное расстояние между двумя средними тем менее значимо в статистическом смысле, чем больше вариация распределений, соответствующих этим средним. Один из методов поиска лучшей дискриминации данных заключается в нахождении такой канонической Д.ф. d_{km} , которая бы максимизировала отношение межгрупповой вариации к внутригрупповой:

$$\lambda = \frac{B(d)}{W(d)},$$

где B – межгрупповая и W – внутригрупповая матрицы рассеяния наблюдаемых переменных от средних.

ДИСКРИМИНАНТНАЯ ФУНКЦИЯ ЛИНЕЙНАЯ (ФИШЕРА)

осуществляет преобразование исходного множества переменных, в одномерную величину. *Дискриминантный анализ* для двух групп также называется линейным дискриминантным анализом Фишера – (ЛДА). ЛДА – метод поиска линейной комбинации переменных, наилучшим образом разделяющих два или более класса. Сам по себе он не является алгоритмом классификации, хотя и работает с информацией о принадлежности объекта к одному из классов. Однако чаще всего результат работы линейного дискриминантного анализа используется как часть линейного классификатора. Другое возможное применение – снижение размерности входных данных перед применением нелинейных алгоритмов классификации. Р.Э.Фишер предложил применять линейную комбинацию, которая максимизирует различия между классами, но минимизирует дисперсию внутри классов. Для этого необходимо определить линейную комбинацию для каждого класса, называемую классифицирующей функцией. Она имеет следующий вид: $d_{ik} = b_{k0} + b_{k1}x_{i1} + \ldots + b_{kp}x_{ip} + \ln q_k$, k = 1,…, g, где d_{ik} – значение функции для класса k; b_{kp} – коэффициенты, которые необходимо определить; q_k – априорная вероятность того, что объект принадлежит к группе k. Объект $x_i = (x_{i1} \ldots x_{ip})$ относится к классу с наибольшим значением d_{ik}. Коэффициенты для классифицирующих функций определяются с помощью вычислений:

$$b_{ki} = (n - g) \sum_{j=1}^{p} (\omega^{-1})_{ij} \overline{x_{jk}}, \quad k = 1,\ldots, g,$$

где b_{ki} – коэффициент для переменной i в выражении, соответствующем классу k, а $(\omega^{-1})ij$ – элемент матрицы, обратной внутригрупповой матрице сумм попарных произведений W. Константа определяется:

$$b_{k0} = \frac{1}{2} \sum_{j=1}^{p} b_{kj} \overline{x_{jk}}, \quad k = 1,\ldots, g.$$

На практике чаще всего применяют линейный дискриминантный анализ. В этом случае дискриминантная функция представляет собой либо прямую, либо плоскость (гиперплоскость) разделяющие совокупности на *классы*. Простая Д.ф.л. осуществляет преобразование исходного множества переменных в одномерную величину. Эта преобразованная переменная, определяет положение объекта на прямой, определённой дискриминантной функцией. Поэтому можно представлять дискриминантную функцию как способ преобразования многомерной задачи в одномерную.

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

совокупность статистических методов классификации многомерных наблюдений, используемых в ситуации, когда исследователь обладает, т.н., обучающими выборками.

Решается задача классификации n объектов, каждый из которых характеризуется значениями k показателей $x_1, x_2, \ldots x_k$ на p однородных в определенном смысле групп. Причём число классов p нам заранее известно. Напр., n пртий отрасли по показателям финансовой деятельности нужно разбить на три группы $(p = 3)$: успешные, средние и не успешные.

Среди n рассматриваемых пр-тий эксперты определили n_1 пр-тие, относящиеся к первой группе, n_2 – ко второй и n_3 – к третьей группе, причём $n > (n_1 + n_2 + n_3)$. В основе Д.а. лежит матрица наблюдений размерности $n \times k$:

$$X = \begin{pmatrix} x_{11} & x_{1j} & x_{1k} \\ x_{i1} & x_{ij} & x_{ik} \\ x_{n1} & x_{nj} & x_{nk} \end{pmatrix},$$

где x_{ij} – значение j-го показателя $(j = 1,2,...,k)$ для i-го наблюдения $(i = 1,2,...,n)$. Напр., x_{ij} – число работающих на i-м пр-тии . Тогда i-я строка матрицы содержит значения *k* признаков, характеризирующих i-й объект. Имеется также *p* обучающих выборок $X^{(l)} = (X_1^{(l)},..., X_i^{(l)} ,..., X_{n_l}^{(l)})^T$, где $l = 1,2,..., p$. Т.о. *l*-я обучающая выборка представляет собой матрицу наблюдений размерности $n_l \times k$. При этом число *p* обучающих выборок равно общему числу всех возможных классов. В Д.а. под однородной группой, классом понимается *ген. совокупность*, описываемая одномодальным, наиболее часто k-мерным нормальным законом распределения с функцией плотности $f(x;\theta_l)$, где θ_l – вектор неизвестных параметров распределения, оценку которого $\hat{\theta}_l$ находят по обучающей выборке $X^{(l)}$ объёмом n_l для l-го класса.

Получив оценку вектора параметров $\hat{\theta}_l$, находят оценку плотности распределения $\hat{f}(X_i;\hat{\theta}_l)$ для i - го наблюдения, где $i = 1,2,...,n$ и $l = 1,2,..., p$. Наблюдение X_i относят к той совокупности l_0, которой соответствует наибольшее значение плотности, т.е. если $\hat{f}(X_i;\hat{\theta}_{l_0}) => \hat{f}(X_i;\hat{\theta}_l)$, для всех $l = 1,2,..., p$ Классифицируемые наблюдения $X_1, X_2,..., X_n$ интерпретируются в данной задаче как выборка из ген. совокупности, описываемой так называемой смесью *p* классов (одномодальных ген. совокупностей) с плотностью вероятности:

$$f_{(X)} = \sum_{l=1}^{p} \pi_l \cdot f_l(X),$$

где π_l – априорная вероятность появления в выборке элемента из *l*-го класса с плотностью $f_l(X) = f(X,\theta_l)$, т.е. доля объектов l-го класса в общей совокупности. Процедуру классификации называют оптимальной, если среди всех других процедур она обладает наименьшими потерями от ошибочной классификации (отнесения объекта m-го класса к l-му).

При выполнении условия нормальности распределения $X^{(l)}$ и $\pi_l = \pi$ для всех $l = 1,2,..., p$ правило дискриминации сводится к следующему: наблюдение X_i следует отнести к той совокупности $X^{(l)}$, *расстояние Махаланобиса* до центра которой $\overline{X}^{(l)}$, описываемого вектором средних, минимально, а вероятность принадлежности к $X^{(l)}$- максимальна. Здесь $\overline{X}^{(l)}$ вектор средних значений k показателей $\overline{x}_1^{(l)}$, $\overline{x}_2^{(l)},..., \overline{x}_k^{(l)}$, полученных по n_l наблюдениям $l-й$ обучающей выборки.

Е

ЕВКЛИДОВО РАССТОЯНИЕ

способ (мера, метрика) нахождения расстояния между объектами в задачах *кластерного анализа*. Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается *k*-признаками, то он может быть представлен как точка в *k*-мерном пространстве, и сходство с другими объектами будет определяться как соответствующее расстояние. В кластерном анализе используются различные меры расстояния между объектами, но наиболее используемым является Е.р.:

$$d_E(X_i, X_j) = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2},$$

где $d_E(X_i, X_j)$ – расстояние между *i*-м и *j*-м объектами; X_{ik}, X_{jk} – значение *k-й* переменной у *i*-го и *j*-го объекта ($k = 1,2,..., m$; $i, j = 1,2,..., n$). Использование этого расстояния оправдано в следующих случаях: наблюдения берутся из ген. совокупности, имеющей многомерное нормальное распределение с ковариационной матрицей вида $\sigma^2 E_k$, т.е. компоненты вектора X взаимно независимы и имеют одну и ту же дисперсию, где E_k – единичная матрица; компоненты вектора наблюдений X однородны по физическому смыслу и одинаково важны для классификации. Предполагается, что признаковое пространство

совпадает с геометрическим пространством, и понятие близости объектов соответственно совпадает с понятием геометрической близости в этом пространстве. Естественно с содержательной точки зрения Е.р. может оказаться бессмысленным, если его признаки имеют разные единицы измерения. Для приведения признаков к одинаковым единицам прибегают к нормировке каждого признака путём деления центрированной величины на среднее квадратическое отклонение и переходят от матрицы X к нормированной матрице Z с элементами:

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{s_i},$$

где x_{ij} – значение *i*-го признака у *j*-го объекта; $\overline{x_j}$ – *среднее арифметическое значение* *j*-го признака; $s_i = \sqrt{\frac{1}{n}\sum_i (x_{ij} - \overline{x_j})^2}$ – *отклонение среднеквадратическое* *j*-го признака. В случае тесной взаимозависимости признаков переходят от исходных признаков к *гл. компонентам*.

З

ЗНАЧИМОСТЬ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

неравенство нулю ген. *коэффициента корреляции*. Проверяется с помощью различных *критериев: Стьюдента, Фишера – Иейтса или Фишера – Снедекора*. Значимость *частных* и *парных коэффициентов корреляции* сводится к проверке гипотезы $H_0 : \rho = 0$, с помощью t – критерия Стьюдента. Наблюдаемое значение критерия находится по формуле:

$$t_{набл} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-l-2},$$

где r – соответственно оценка частного или парного коэффициента корреляции; l – порядок частного коэффициента корреляции, т.е. число фиксируемых факторов. Для парного коэффициента корреляции l = 0. Проверяемый коэффициент корреляции считается значимым, т.е. гипотеза $H_0 : \rho = 0$ отвергается с вероятностью ошибки α, если $t_{набл}$ по модулю будет больше, чем $t_{кр}$, определяемого по таблицам t – распределение Стьюдента для заданного α и $\nu = n - l - 2$. Если же $|t_{набл}| < t_{кр,}$ то гипотеза H_0 не отвергается, т.е. гипотеза об отсутствии зависимости между признаками не противоречит наблюдениям. Значимость частных и парных коэффициентов корреляции можно проверить также с помощью таблиц Фишера–Иейтса. В этом случае гипотеза H_0 отвергается с вероятностью ошибки α, если полученное значение r коэффициента корреляции по модулю окажется больше табличного значения $r_{кр}$, найденного по таблице Фишера–Иейтса при заданном α и числе степеней свободы $\nu = n - l - 2$, из этого следует, что зависимость между величинами имеет место. В противном случае | r | < гкр гипотеза $H_0 : \rho = 0$ не отвергается. Значимость множественного коэффициента корреляции (или его квадрата – коэффициента детерминации) проверяется по F – критерию. Напр., для *множественного коэффициента корреляции* $r_{1(2,\ldots k)}$ (МКК) проверка значимости сводится к проверке гипотезы, что ген. множественный коэффициент корреляции равен нулю, т.е. $H_0 : \rho_{1(2,\ldots k)} = 0$, а наблюдаемое значение статистики находится по формуле:

$$F_{набл} = \frac{\frac{1}{k-1} r^2_{1(2,\ldots,k)}}{\frac{1}{n-k}\left(1 - r^2_{1(2,\ldots k)}\right)}.$$

МКК считается значимым, т.е. имеет место линейная статистическая зависимость, между x_1 и остальными факторами $x_2, \ldots, x_k$, если $F_{набл} > F_{кр}(\alpha, k-l, n-k)$, где $F_{кр}$ определяется по таблице F– распределения Фишера – Снедекора для заданных $\alpha, \nu_1 = k - l, \nu_2 = n - k$.

И

ИЕРАРХИЧЕСКИЕ ПРОЦЕДУРЫ КЛАСТЕРНОГО АНАЛИЗА

наиболее распространенные методы разбиения n многомерных наблюдений $x_1, x_2, \ldots, x_n$ на заранее неизвестное число p однородных групп, кластеров. Иерархические кластер – процедуры могут быть агломеративные и дивизимные. Принцип работы иерархичесих агломеративных (дивизимных) процедур состоит в последовательном объединении (разделении) групп элементов сначала самых близких (далёких), а затем все более отдаленных друг от друга (при-

ближённых друг к другу). Сущность агломеративных методов заключается в том, что на первом шаге каждое наблюдение *Xi (i=1,2,...,n)* рассматривается как отдельный кластер. На основании матрицы расстояний или матрицы сходства объединяются наиболее близкие объекты (кластеры). Если матрица расстояния первоначально имела размерность *n x n*, то полностью процесс кластеризации завершится за *n -1* шагов, в итоге все объекты будут объединены в один кластер. Важным для агломеративных иерархических процедур является выбор вида расстояния между отдельными элементами и кластерами. Алгоритм иерархической классификации предусматривает графическое представление результатов классификации в виде *дендограммы*.

ИНФОРМАЦИОННЫЙ КРИТЕРИЙ АКАИКЕ

(Akaike's information criterion (AIC))

наряду с *информационным критерием Шварца*, один из самых простых и распространённых способов выбора наилучшей модели, основанный на принципе снижения остаточной суммы квадратов при добавлении значимого фактора.

При использовании этого критерия, линейной модели с *p* объясняющими переменными, оцененной по *n* наблюдениям, сопоставляется значение

$$AIC = ln\left(\frac{RSS_p}{n}\right) + \frac{2p}{n} + 1 + ln\,2\pi\,,$$

где RSS_p – остаточная сумма квадратов, полученная при оценивании коэффициентов модели *методом наименьших квадратов*. При увеличении количества объясняющих переменных первое слагаемое в правой части уменьшается, а второе увеличивается. Среди нескольких альтернативных моделей (полной и редуцированных) предпочтение отдается спецификации модели с наименьшим значением AIC, в которой достигается определенный компромисс между величиной остаточной суммы квадратов и количеством объясняющих переменных. Для авторегриссионных процессов, в отличие от информационного критерия Шварца, AIC переоценивает порядок модели и, следовательно, оценка порядка модели на основании этого критерия несостоятельна.

ИНФОРМАЦИОННЫЙ КРИТЕРИЙ ШВАРЦА

(Schwarz's information criterion (SIC)) – наряду с *информационным критерием Акаике*, один из самых простых и распространённых способов выбора наилучшей модели из некоторого набора альтернативных моделей, основанный на принципе снижения остаточной суммы квадратов при добавлении значимого фактора. При использовании этого критерия, для линейной модели с *p* объясняющими переменными, оцененной по *n* наблюдениям, сопоставляется значение:

$$SIC = ln\left(\frac{RSS_p}{n}\right) + \frac{p\,ln\,n}{n} + 1 + ln\,2\pi\,,$$

где RSS_p – остаточная сумма квадратов, полученная при оценивании коэффициентов модели *методом наименьших квадратов*. При увеличении количества объясняющих переменных первое слагаемое в правой части уменьшается, а второе увеличивается. Среди нескольких альтернативных моделей (полной и редуцированных) предпочтение отдается модели с наименьшим значением SIC. SIC всегда выбирает наилучшую модель с числом параметров, не превышающим число параметров в модели, которая была выбрана по критерию Акаике. Кроме того, SIC является асимптотически состоятельным, в то время как информационный критерий Акаике смещён в сторону выбора перепараметризованной модели.

ИНФОРМАЦИОННАЯ МАТРИЦА ФИШЕРА

матрица $I(\Theta,X)$ размерности $k \times k$, является характеристикой, на основании которой измеряют величину изменения функции правдоподобия $L(X^*,\Theta)$ при изменении значения k-мерного параметра $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$, при $k \geq 2$ и определяется как:

$$I_{ij}(\Theta,X) = E\left[\left(\frac{\partial \ln L}{\partial \theta_i}\right)\times\left(\frac{\partial \ln L}{\partial \theta_j}\right)\right].$$

Функция правдоподобия $L = L(X^*,\Theta)$, определяемая равенством $L(x_1^*,x_2^*,\ldots,x_n^*;\theta) = f(x_1^*;\theta)\cdot f(x_2^*;\theta)\cdot\ldots\cdot f(x_n^*;\theta)$, задаёт вероятность получения, при извлечении конкретной выборки объёма n из независимо и одинаково распределённой случайной величины $X = (x_1, x_2,\ldots,x_n)$, именно наблюдений $X^* = (x_1^*,x_2^*,\ldots,x_n^*)$ (или величину, пропорциональную вероятности получения выборочных значений в непосредственной близости от точки X^* в непрерывном случае). Функция правдоподобия рассматривается как функция неизвестного k-мерного параметра $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ (при заданных фиксированных наблюдениях $x_1^*,x_2^*,\ldots,x_n^*$). Чем резче проявляется зависимость вероятности $L(X^*,\Theta)$ от параметра Θ, тем больше информации заключено в конкретных значениях величин X и Θ друг о друге. Если по наблюденному значению X^* случайной величины X можно с вероятностью единица точно восстановить значение параметра Θ, то это значит, что случайная величина X (или её наблюдение X^*) содержит максимально возможную информацию о параметре Θ. И наоборот если распределение $L(x_1^*,x_2^*,\ldots,x_n^*;\theta) = f(x_1^*;\theta)\cdot f(x_2^*;\theta)\cdot\ldots\cdot f(x_n^*;\theta)$ случайной величины X одно и тоже при всех значениях параметра Θ, то нет оснований делать какие-либо заключения о Θ по результатам наблюдений этой случайной величины.

ИНФОРМАЦИОННОЕ РАССТОЯНИЕ КАЛЛБЭКА

используется в теоретико-вероятностной схеме кластер-анализа для измерения расстояния между нормальными классами S_l и S_m и определяется формулой:

$$\rho^2(S_l,S_m) = \frac{1}{2}(a(l)-a(m))^{\mathrm{T}}(\Sigma^{-1}(l)+\Sigma^{-1}(m))(a(l)-a(m)) + \frac{1}{2}tr\left\{(\Sigma(l)-\Sigma(m))(\Sigma^{-1}(l)-\Sigma^{-1}(m))\right\},$$

где: $a(l)$, $a(m)$ – вектора средних значений l-го и m-го нормальных *классов*, а $\Sigma(l)$ и $\Sigma(m)$ – *ковариационные матрицы* этих классов.

В данной схеме анализируемая ген. совокупность интерпретируется как смесь унимодальных ген. совокупностей, каждая из которых и представляет один из искомых классов.

В статистической практике приведённая формула используется для вычисления расстояний между классами и при отклонении распределения наблюдений внутри классов от нормального с заменой теоретических характеристик a(j) и Σ(j) их оценками $\hat{a}(j)$ и $\hat{\Sigma}(j)$, построенными по наблюдениям, составляющим класс с номером j (j= l,m).

ИНФОРМАЦИОННАЯ ХАРАКТЕРИСТИКА СВЯЗИ

характеристика степени тесноты статистической связи Y^2 классификационных переменных x_1 и x_2, определяемая соотношением:

$$Y^2 = 2\sum_{i=1}^{m_1}\sum_{j=2}^{m_2} n_{ij}\ln\left(\frac{n_{ij}}{n_i.n._j/n}\right),$$

где n_{ij} – число наблюдении, имеющих градацию i по переменной x_1, где $i=1,2,\ldots,m_1$, и градацию j по переменной x_2, где $j=1,2,\ldots, m_2$;

$n_{i.}$ – число наблюдений, имеющих градацию i по переменной x_1 $n_{j.} = \sum n_{ij}$

Соответственно $n_{.j}$ – число наблюдений, имеющей градацию j по переменной x_2;

$n = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} n_{ij}$ – общее число наблюдений

Значения Y^2 варьируют от 0 (в случае статистической независимости x_1 и x_2) до $+\infty$, и в условиях справедливости гипотезы о статистической независимости анализируемых признаков она приблизительно подчиняется *закону распределения вероятностей* χ^2 с $(m_1\text{-}1)(m_2\text{-}1)$ степенями свободы.

Для проверки гипотезы о статистически значимом отличии от нуля Y^2 (т.е. гипотеза о наличии связи между x_1 и x_2), необходимо убедиться в выполнении неравенства $Y^2 > \chi^2_\alpha (m_1\text{-}1)(m_2\text{-}1)$, где χ^2_α – 100α%-я точка χ^2 распределения с m степенями свободы, а α – заданный уровень значимости критерия, т.е. вероятность принять ложное решение о наличии статистической связи между анализируемыми переменными.

К

КЛАСС

группа однородных в некотором смысле объектов; множество объектов, сходных по природе и признакам; *ген. совокупность*, описываемая одномодальной *функцией плотности* f(X) (или одномодальным *полигоном* вероятностей в случае дискретных признаков X). Выбор того или иного определения К. зависит от постановки задачи классификации.

Для пояснения общей идеи, заложенной в основу построения всех вероятностно-статистических методов классификации, рассмотрим пример решения задачи отнесения наблюдения к одной из двух гипотетических нормальных совокупностей, т.е. к одному из К., различающихся между собой средними значениями. Решение принимается в пользу К., в котором данное наблюдение выглядит более правдоподобно, т.к. ему соответствует наибольшее значение плотности именно в этом К. Этот принцип положен в основу вероятностных методов классификации: наблюдение будет относиться к тому К. (т.е. к той ген. совокупности), в рамках которого (которой) он выглядит более правдоподобным. Этот принцип может корректироваться с учётом удельных весов К. и специфики так называемой «функции потерь» с (j|i), определяющей «стоимость» потерь от отнесения объекта *i*-о К. к К. с номером *j*. Для того чтобы этот принцип практически реализовать, необходимо располагать полным описанием гипотетических К., т.е. знанием функций $f_1(X)$, $f_2(X)$, …, $f_k(X)$, задающих закон распределения вероятностей соответственно для 1-о, 2-о, …, k-о К. Последнее затруднение обходят с помощью обучающих выборок в случае классификации с обучением, и с помощью модели смеси распределений в случае классификации без обучения.

В задачах *кластерного анализа* под К. (кластером) понимают группу однородных объектов расстояние d между которыми в к-мерном пространстве меньше критического значения $d_{кр}$, т.е. $d < d_{кр}$. При этом исходят предположения, что из геометрической близости следует физическая близость объектов.

КЛАССИФИКАЦИОННЫЕ ПЕРЕМЕННЫЕ

переменные, которые позволяют разбить совокупность наблюдений на непересекающиеся множества, которые трудно или невозможно упорядочить по какому-либо признаку.

Типичные примеры таких переменных – качество жилья («плохое», «удовлетворительное», «хорошее», «отличное»), пол особи (мужской, женский), вид и род в биологии и т.д. Если хотя бы одна из переменных является количественной, такие данные исследуются методами *дисперсионного анализа*. В общем случае осн. инструменты исследования зависимостей между К.п. – *таблицы сопряжённости.*

Пусть имеется двумерная случайная величина Z=(X,Y), где случайная величина X принимает значения (признаки) A_1, A_2, …, A_s, а случайная величина Y – значения (признаки) B_1, B_2, …, B_r, при этом n_{ij} – количество выборочных значений, имеющих признаки B_i и A_j. Напр., необходимо проверить, есть ли зависимость между цветом глаз и цветом волос у людей. Если случайная величина X – «цвет глаз», а величина Y – «цвет волос», тогда A_1 – «карие глаза», A_2 – «синие глаза» и т.д., B_1 – «блондин(ка)», B_2 – «брюнет(ка)» и т.д. Каждый индивидуум, информация о котором включена в исследуемую выборку, характеризуется двумя признаками: A_j и B_i, где *j* – номер цвета глаз, *i* – номер цвета волос.

Статистический анализ парных связей между к.п. X и Y производится на базе исходных данных, представленных в виде т.н. двухвходовых табл. сопряжённости (см. табл.).

Таблица

	A_1	A_2	…	A_s	Всего
B_1	n_{11}	n_{12}	…	n_{1s}	$n_{1.}=\sum_{i=1}^{s} n_{1i}$
B_2	n_{21}	n_{11}	…	n_{2s}	$n_{2.}=\sum_{i=1}^{s} n_{2i}$
…	…	…	…	…	…
B_r	n_{r1}	n_{r2}	…	n_{rs}	$n_{rs.}=\sum_{i=1}^{s} n_{ri}$
Всего	$n_{.1}=\sum_{i=1}^{r} n_{i1}$	$n_{.2}=\sum_{i=1}^{r} n_{i2}$	…	$n_{.s}=\sum_{i=1}^{r} n_{is}$	$n=\sum_{i=1}^{s} n_{.i}=\sum_{i=1}^{r} n_{i.}$

Для проверки гипотезы о независимости случайных величин X и Y вычисляется критическая статистика

$$\hat{T}=n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-n_{i}.n_{.j}\right)^2}{n_{i}.n_{.j}}=n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i}.n_{.j}}-1\right).$$

Её значение может меняться от нуля (при статистической независимости переменных X и Y) до +∞. Проверка гипотезы H0: T=0 о статистической независимости переменных X и Y осуществляется с использованием критической статистики $\hat{T}$, которая приближенно имеет значение χ^2 со степенью свободы, равной (r-1)(s-1). Поэтому, если оказалось, что $\hat{T} > \chi_{\alpha}^2$ (r-1)(s-1), то нулевая гипотеза H0: T=0 отвергается, т.е. делается вывод (с вероятностью ошибки α), что К.п. X и Y не являются статистическими независимыми.

На основе статистики $\hat{T}$ разработано несколько показателей меры зависимости между К.п.:

коэффициент сопряжённости $C=\sqrt{\frac{T}{T+n}}$;

мера связи Чупрова $K=\sqrt{\frac{T}{n\sqrt{(r-1)(s-1)}}}$;

коэффициент $\varphi=\sqrt{\frac{T}{n}}$.

Эти коэффициенты используются в различных ситуациях и имеют свои недостатки и преимущества.

КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ

разделение рассматриваемой совокупности объектов или явлений на однородные, в определённом смысле, группы по наиболее характерным или агрегированным признакам, определяющим принадлежность или точность отнесения объекта к к.-л. группе, либо отнесение каждого из заданного множества объектов к одному из заранее известных классов.

В научных исследованиях по совр. теории классификации можно выделить два относительно самостоятельных направления. Одно из них опирается на опыт таких наук, как биология, география, геология, и таких прикладных областей, как ведение классификаторов продукции и библиотечное дело. Типичные объекты рассмотрения – классификация химических элементов (табл. Д.И. Менделеева), биологическая систематика, универсальная десятичная классификация публикаций (УДК), классификатор товаров на основе штрих-кодов.

Другое направление опирается на опыт технических исследований, экономики, маркетинговых исследований, социологии, медицины. Типичные задачи, напр., техническая и медицинская диагностика, а также, разбиение на группы видов экономической деятельности, тесно связанных между собой, выделение групп однородной продукции и т.д.

Развитие электронно-вычислительной техники как средства обработки больших массивов дан-

ных стимулировало проведение в последние годы широких комплексных исследований сложных социально-экономических, технических, медицинских и других процессов и систем, таких, как образ и уровень жизни нас., совершенствование организационных систем, региональная дифференциация социально-экономического развития, планирование и прогнозирование экономических систем и т.д. В связи с многоплановостью и сложностью этих объектов и процессов данные о них носят многомерный, разнотипный характер и до анализа обычно бывает неясно, насколько существенно то, или иное свойство для конкретной цели. В 60-х гг. 20 в. внутри прикладной статистики достаточно чётко оформилась область, посвящённая методам К.м.н. В зависимости от используемых методов и вида априорной информации о классах различают следующие разновидности классификации: группировка, *кластерный анализ, дискриминантный анализ,* расщепление смеси вероятностных распределений, экспертный метод.

Задача кластерного анализа состоит в выяснении по эмпирическим данным, насколько элементы «группируются» или распадаются на изолированные «скопления», «кластеры» (от cluster (англ.) – гроздь, скопление). Иными словами, задача К.м.н. заключается в выявление естественного разбиения на классы, свободного от субъективизма исследователя с целью выделение групп однородных объектов, сходных между собой, при резком отличии групп друг от друга.

Задачи кластеризации и группировки принципиально различны, хотя для их решения могут применяться одни и те же алгоритмы. Важная для практической деятельности проблема состоит в том, чтобы понять, разрешима ли задача кластер-анализа для конкретных данных или возможна только их группировка, поскольку они достаточно однородны и не разбиваются на резко разделяющиеся между собой кластеры.

Традиционно задача группировки решается следующим образом. Из множества признаков, описывающих объект, отбирается один, наиболее информативный с точки зрения исследователя, и производится группировка в соответствии со значениями данного признака. Если требуется провести классификацию по нескольким признакам, ранжированным между собой по степени важности, то сначала производится классификация по первому признаку, затем каждый из полученных классов разбивается на подклассы по второму признаку и т.д. Подобным образом строится большинство комбинационных статистических группировок.

В тех случаях, когда не представляется возможным упорядочить классификационные признаки, применяется наиболее простой метод многомерной группировки – создание интегрального показателя (индекса), функционально зависящего от исходных признаков, с последующей классификацией по этому показателю. Развитием этого подхода является вариант классификации по нескольким обобщающим показателям (гл. компонентам), полученным с помощью методов *факторного* или *компонентного анализа*.

Наиболее трудным и наименее формализованным в задаче классификации является определение понятия однородности объектов. В общем случае понятие однородности объектов задается либо введение правила вычисления расстояний $d(x_i,x_j)$ между любой парой исследуемых объектов (x_1, x_2, ... , x_n), либо заданием некоторой функции $d(x_i,x_j)$, характеризующей степень близости i-го и j-го объектов.

Выбор метрики или меры близости узловой момент исследования, от которого в основном зависит окончательный вариант разбиения объектов на классы при данном алгоритме разбиения. В каждом конкретном случае этот выбор должен производиться в зависимости от целей исследования, физической и статистической природы вектора наблюдений X, априорных сведений о характере вероятностного распределения X.

Наиболее широко в задачах классификации используются следующие расстояния: обычное и взвешенное евклидово, Махаланобиса, Хемминогово. Меры близости двух групп объектов: расстояние, измеряемое по принципу «бли-

жайшего соседа», «дальнего соседа», «средней связи», «центра тяжести» и другие.

Проверить качество разбиения можно с помощью целевой функции, значения которой позволят сопоставить различные схемы классификации. В экономических исследованиях целевая функция, как правило, должна минимизировать некоторый параметр, определённый на множестве объектов (напр., целью классифицирования оборудования может явиться группировка, минимизирующая совокупность затрат времени и средств на ремонтные работы). В случаях, когда формализовать цель задачи не удается, критерием качества классификации может служить возможность содержательной интерпретации найденных групп.

В качестве критерия естественности классификации следует рассматривать устойчивость относительно выбора алгоритма кластерного анализа. Проверить устойчивость разбиения можно, применив к данным несколько подходов. Если полученные результаты содержательно близки, то они адекватны действительности. В противном случае следует предположить, что естественной классификации не существует, задача кластерного анализа не имеет решения.

В дискриминантном анализе (синонимы: распознавание образов, классификация с обучением) кластеры предполагаются заданными унимодальными плотностями вероятностей или обучающими выборками. Наблюдение X_i следует отнести к одной из k анализируемых совокупностей, в рамках которой оно выглядит наиболее правдоподобным. Другими словами, если дано точное описание в виде функций $f_1(X),\ldots,f_k(X)$ плотности в непрерывном случае или полигонов вероятностей в дискретном конкурирующих ген. совокупностей, то следует поочередно вычислить значения функций плотности для данного наблюдения X_i в рамках каждой из рассматриваемых ген. совокупностей (т.е. вычислить значения $f_1(X_i), f_2(X_i) \ldots f_k(X_i)$) и отнести X_i к тому классу, функция плотности которого максимальна. Если известен лишь общий вид функций $f_1(X;\Theta_1),\ f_2(X;\Theta_2),\ldots,\ f_k(X,\Theta_k)$, описывающих анализируемые классы, но неизвестны значения многомерных параметров $\Theta_1,\Theta_2,\ldots,\Theta_k$, и если при этом располагают обучающими выборками, то данный случай относится к параметрической схеме дискриминантного анализа и порядок действий следующий: по j-ой обучающей выборке оценивают параметр $\Theta_j\ (j=1,2,\ldots,k)$, затем производят классификацию наблюдений, руководствуясь принципом максимальной плотности, как и в случае известных функций $f_j(X)$.

Методы расщепления смесей вероятностных распределений оказываются полезными в том случае, когда каждый класс интерпретируется как параметрически заданная одномодальная ген. совокупность $f_j(X;\Theta_1),\ j=1,2,\ldots,k$ при неизвестном векторе значений параметров Θ_j и соответственно каждое из классифицируемых наблюдений X_i считается извлеченным из одной из этих (но неизвестно, из какой именно) ген. совокупностей.

В схеме автоматической классификации, опирающейся на модель смеси распределений, как и в схеме параметрического дискриминантного анализа, задающие искомые классы функций $f_1(X;\Theta_1),\ f_2(X;\Theta_2),\ldots$ известны лишь с точностью до значений параметров. Но в схеме автоматической классификации неизвестные значения параметров $\Theta_1,\Theta_2,\ldots,\Theta_k$, число k компонентов смеси и их удельные веса $\pi_1,\pi_2,\ldots,\pi_k$ в общей совокупности оцениваются не по обучающим выборкам, а по классифицируемым наблюдениям $X_1,X_2,\ldots,X_n$ с помощью метода максимального правдоподобия. Начиная с момента, когда по выборке $X_1,X_2,\ldots,X_n$ получены оценки $\hat{k},\hat{\pi}_1,\hat{\pi}_2,\ldots,\hat{\pi}_k,\hat{\Theta}_1,\ldots,\hat{\Theta}_k$ неизвестных параметров $k,\pi_1,\pi_2,\ldots,\pi_k,\Theta_1,\ldots,\Theta_k$ модели $f(X)=\sum_{j=1}^{k}\pi_j f(X;\Theta(j))$ приходят к схеме дискриминантного анализа и процесс классификации наблюдений производят по схеме параметрического дискриминантного анализа (т.е относят наблюдение X_i к классу с номером j_0, если $\hat{\pi}_{j0} f_{j0}(X_i;\hat{\Theta}_{j0}) = \max_{1\le j\le k}\{\hat{\pi}_j f_j(X_i;\hat{\Theta}_j)\}$.

Статистические методы многомерной классификации не всегда могут быть использованы, и в первую очередь, из-за отсутствия информа-

ции. В таких случаях наиболее часто используются методы экспертных оценок, базирующиеся на опыте и интуиции специалистов. При организации экспертизы должен использоваться системный подход, т.е.: определены цель и задачи экспертного опроса; выбрана методика его проведения; определены объекты, факторы и показатели; установлены методы обработки результатов. Достоверность результатов зависит от правильной организации экспертизы.

КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ БЕЗ ОБУЧЕНИЯ

разделение рассматриваемой совокупности объектов или явлений на однородные, в определённом смысле, группы. Состоит в нахождении естественного разделения объектов на однородные группы так, чтобы объекты, попавшие в одну группу, были как можно более близки (схожи) друг с другом, а объекты из разных групп (классов) были по возможности несхожими. При этом характеристики кластеров, а также обучающие выборки заранее не известны.

Рассматривается случай классификации многомерных наблюдений, когда искомая информация о классифицируемых объектах $O_1, O_2, ..., O_n$ представлена либо в форме матрицы X «объект-свойство», либо в форме матрицы попарных сравнений объектов, где величина $\gamma_{ij} = \rho_{ij}$ характеризует взаимную отдаленность (или близость) объектов O_i и O_j. В этом случае для классификации используют методы *кластерного анализа*. Переход от формы исходных данных типа «объект-свойство» к форме матрицы попарных расстояний осуществляется посредством задания способа вычисления расстояния (близости) между парой объектов, когда известны координаты (значения признаков) каждого из них.

При наличии априорных сведений о виде закона распределения искомых классов следует обратиться либо к классификации с использованием методов расщепления смесей вероятностных распределений, которые оказываются полезными в том случае, когда каждый класс интерпретируется как параметрически заданная унимодальная ген. совокупность $f_j(X;\Theta_1),\ j=1,2,...,k$ при неизвестном значении определяющего её векторного значения параметра Θ_j и соответственно каждое из классифицируемых наблюдений X_i считается извлечённым из одной из этих (но неизвестно, из какой именно) ген. совокупностей.

КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ С ОБУЧЕНИЕМ

(распознавание образов) – классификация объектов на основе различных характеристик, т.е. отнесение объекта некоторым оптимальным способом к одному из заранее сформированных *классов*. Задача состоит в том, чтобы построить решающее правило, позволяющее по результатам измерений параметров объекта указать группу, к которой он принадлежит. Один из наиболее распространённых методов К.м.н. с о. – *дискриминантный анализ*.

Базовая идея, лежащая в основе принятия решения, к какой из k – анализируемых совокупностей отнести данное классифицируемое наблюдение X_i, состоит в том, что наблюдение следует отнести к той ген. совокупности, в рамках которой оно выглядит наиболее правдоподобным. Другими словами, если дано точное описание (напр., в виде функций $f_1(X),\ldots,f_k(X)$ плотности в непрерывном случае или полигонов вероятностей в дискретном) конкурирующих ген. совокупностей, то следует поочередно вычислить значения функций плотности для данного наблюдения X_i в рамках каждой из рассматриваемых ген. совокупностей (т.е. вычислить значения $f_1(X),f_2(X),...,f_k(X)$) и отнести X_i к тому классу, значение функции плотности для которого максимальна. Если известен лишь общий вид функций $f_1(X;\Theta_1),\ f_2(X;\Theta_2),...,\ f_k(X,\Theta_k)$, описывающих анализируемые классы, но неизвестны значения многомерных параметров $\Theta_1,\Theta_2,...,\Theta_k$, и если при этом располагают обучающими выборками, то данный случай относится к параметрической схеме дискриминантного анализа и порядок действий следующий: по j-ой обучающей выборке оценивают параметр $\Theta_j\ (j=1,2,...,k)$, затем производят

классификацию наблюдений, руководствуясь принципом макс. правдоподобия, как и в случае известных функций $f_j(X)$.

КЛАСТЕР

см. в ст. Класс

КЛАСТЕРНЫЙ АНАЛИЗ

группа методов статистического анализа, позволяющая разбивать анализируемую совокупность точек (наблюдений) на сравнительно небольшое число (заранее известное или нет) *классов* т.о., чтобы объекты, принадлежащие одному классу, находились бы на сравнительно небольших расстояниях друг от друга. Полученные в результате разбиения классы часто называют кластерами (таксонами, образами).

К.а. позволяет определить естественное расслоение исходных наблюдений на чётко выраженные кластеры, лежащие друг от друга на некотором расстоянии, но не разбивающиеся на столь же удаленные части. Такая задача не всегда имеет решение: может оказаться, что множество исходных наблюдений не обнаруживает естественного расслоения на кластеры.

С точки зрения априорной информации об окончательном числе кластеров, на которую требуется разбить исследуемую совокупность объектов, задачи К.а подразделяются на три осн. типа: число кластеров априори задано; число кластеров неизвестно и подлежит определению (оценке); число кластеров неизвестно, но его определение и не входит в условие задачи, требуется определить иерархию исследуемой совокупности.

Соответственно типам задач выделяются три типа процедур К.а.: процедуры иерархические (агломеративные и дивизимные) предназначены для решения задач 2-го и 3-го типа, объём совокупности классифицируемых наблюдений должен быть сравнительно небольшим (как правило, не более чем несколько десятков наблюдений). Иерархические методы могут иногда применяться и для решения задач 1-го и 2-го типа с большим объёмом совокупности наблюдений (несколько сотен и тыс.); процедуры параллельные предназначены для решения задач 1-го и 2-го типа с небольшим объёмом совокупности наблюдений. Они реализуются с помощью итерационных алгоритмов, на каждом шаге которых одновременно (параллельно) используются все имеющиеся наблюдения; процедуры последовательные предназначены в основном для решения задач 1-го и 2-го типа с большим объёмом совокупности наблюдений. Реализуются с помощью итерационных алгоритмов, на каждом шаге которых используются лишь небольшая часть наблюдений, а также результат разбиения на предыдущем шаге.

В задачах К.а. обычной формой представления исходных данных служит прямоугольная матрица, каждая строка которой представляет результат измерения k рассматриваемых признаков на одном из обследованных объектов:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

В конкретных ситуациях может представлять интерес как группировка объектов, так и группировка признаков. Числовые значения, входящие в матрицу X, соответствуют трём типам: *количественным переменным*, ранговым (порядковым) переменным и качественным (номинальным). Желательно, чтобы табл. исходных данных соответствовала одному типу переменных. Для этого используются предварительные преобразования, переводящие, напр., количественные переменные в ранговые, разбивая области значений количественных переменных на интервалы, которые затем нумеруются числами натурального ряда.

Матрица X не единственный способ представления исходных данных в задачах К.а. Исходная информация задаётся в виде квадратной матрицы: $R = (r_{ij}); i, j = 1,2,\dots,k$,

где r_{ij} – степень близости *i*-го объекта к *j*-му.

Наиболее трудное и наименее формализованное в задаче классификации – определение понятия однородности объектов. Оно может быть задано введением правила вычисления расстояния $\rho(X_i, X_j)$ между любой парой исследуе-

мых объектов $(X_1, X_2, \ldots, X_n)$. Если задана функция $\rho(X_i, X_j)$ то близкие с точки зрения этой метрики объекты считаются однородными, принадлежащими к одному классу. При этом расчётное значение функции должно сопоставляться с некоторым пороговым значением, определяемым в каждом конкретном случае по-своему. Наиболее часто в К.а. для определения расстояния между объектами используется обычное *евклидово расстояние, расстояние Махалонобиса* и Хемингово расстояние.

Для сравнения качества различных способов разбиения заданной совокупности элементов на классы вводится понятие функционала качества разбиения $Q(S)$, определенного на множестве всех возможных разбиений. Наилучшее разбиение S^* определяет такое разбиение, при котором достигается экстремум выбранного функционала качества. За функционалы качества часто берутся следующие характеристики: сумма внутриклассовых дисперсий, сумма попарных внутриклассовых расстояний между элементами и обобщенная внутриклассовая дисперсия.

Рассмотрим осн. типы кластер-процедур. Иерархические (деревообразные) процедуры бывают двух типов: агломеративные и дивизимные. Принцип работы агломеративных процедур заключается в последовательном объединении отдельных элементов в группы сначала самых близких, а затем все более отдаленных друг от друга; в дивизимных процедурах – наоборот, происходит процесс разъединения все совокупности наблюдений.

В агломеративных иерархических процедурах К.а. используют понятие расстояния между группами объектов. Наиболее употребительные расстояния и меры близости между классами – расстояния, измеряемое по принципу «*ближайшего соседа*», «*дальнего соседа*», «*центрам тяжести*» групп и по принципу «*средней связи*».

Параллельные кластер-процедуры предусматривают одновременный обсчёт всех исходных наблюдений на каждом шаге алгоритма. К такому типу процедур относятся алгоритмы, связанные с функционалами качества разбиения, такие, как алгоритмы «последовательного переноса точек из класса в класс». Эти алгоритмы отправляются от некоего начального разбиения, полученного произвольно или с помощью методов предварительной обработки, вычисляется значение принятого качества разбиения, затем каждое наблюдение поочередно перемещается во все кластеры и остается в том положении, которое соответствует наилучшему значению функционала качества.

Последовательные кластер-процедуры используют итерационные алгоритмы, на каждом шаге которых последовательно обсчитывается лишь небольшая часть исходных наблюдений или даже одно из них.

Наиболее распространённый метод, относящийся к последовательным кластер-процедурам – метод k-средних. Суть описываемого алгоритма заключается в последовательном уточнении k эталонных точек, которые характеризуют центры кластеров таким образом, чтобы каждое из наблюдений относилось к одному из k кластеров и центр каждого кластера совпадал с центром тяжести относящихся к нему наблюдений. Проведённые исследования свойств метода k-средних говорят о том, что в достаточно общих ситуациях при больших объёмах выборочных совокупностей (от нескольких сотен и более) этот алгоритм строит разбиение, близкое к наилучшему в смысле функционала.

КОВАРИАЦИОННАЯ МАТРИЦА

определяется как *математическое ожидание* произведения центрированного случайного вектора на этот же транспонированный вектор, т.е. $\sum[(\xi-a)(\xi-a)^T]$.

К.м. Σ многомерной случайной величины $\xi = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})^T$ есть (p × p)-матрица, составленная из ковариаций $\sigma_{jk} = M[(\xi^{(j)} - a^{(j)})(\xi^{(k)} - a^{(k)})]$, где j, k = 1, 2, …, p, а $a^{(i)}=M(\xi^{(i)})$ – математическое ожидание случайной величины $\xi^{(j)}$; имеет вид:

$$\Sigma=\begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1j} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2j} & \dots & \sigma_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{i1} & \sigma_{i2} & \dots & \sigma_{ij} & \dots & \sigma_{ip} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pj} & \dots & \sigma_{pp} \end{pmatrix}.$$

К.м. содержит сведения о степени случайного разброса анализируемых переменных, а также о характере и структуре статистических взаимосвязей между ними. По гл. диагонали К.м., когда i=k, находятся дисперсии элементов $\xi^{(j)}$ вектора $\xi=(\xi^{(1)}, \xi^{(2)},\dots, \xi^{(p)})^{\tau}$. Из определения следует, что она является симметричной и неотрицательно-определённой.

В статистическом анализе используется её статистическая оценка, или выборочная К.м. $\Sigma' = (\sigma'_{jk})$, где

$$\sigma'_{jk}=\frac{1}{n}\sum_{i=1}^{n}(x_i^{(j)}-\bar{x}_i^{(j)})(x_i^{(k)}-\bar{x}_i^{(k)}),$$

$\bar{x}^{(l)}=\frac{1}{n}\sum_{i=1}^{n}x_i^{(l)}$ – выборочные средние значения случайных величин $\xi^{(l)}$, l = 1, 2, …, p.

КОВАРИАЦИОННЫЙ АНАЛИЗ

совокупность методов *математической статистики*, относящихся к анализу моделей зависимости среднего значения некоторой случайной величины Y от набора неколичественных факторов F и одновременно от набора количественных факторов X. По отношению к Y переменные X называют сопутствующими; факторы F задают сочетания условий качественной природы, при которых получены наблюдения Y и X, и описываются с помощью т.н. индикаторных переменных; среди сопутствующих и индикаторных переменных могут быть как случайные, так и неслучайные (контролируемые в эксперименте). Если случайная величина Y является вектором, то говорят о многомерном К.а. Осн. теоретические и прикладные проблемы К.а. относятся к линейным моделям. В частности, если анализируются n наблюдений $Y_1, \dots, Y_n$ с p сопутствующими переменными ($X = (x^{(1)}, \dots x^{(p)})$), k возможными типами условий эксперимента ($F = (f_1, \dots f_k)$), то линейная модель соответствующего К.а. задается уравнением:

$$Y_i=\sum_{j=1}^{k}f_{ij}\Theta_j+\sum_{s=1}^{p}\beta_s(f_i)x_i^{(s)}+\varepsilon_i(f_i), \qquad (1)$$

где i = 1,..., n, индикаторные переменные f_{ij} равны 1, если j-е условие эксперимента имело место при наблюдении Y_i, и равны 0 в ином случае. Коэффициенты θ_j определяют эффект влияния j-го условия; x_i^{s} – значение сопутствующей переменной $x^{(s)}$, при котором получено наблюдение Y_i, i = 1,...,n; s = 1, ..., p; $\beta_s(f_i)$ – значения соответствующих коэффициентов регрессии Y по $x^{(s)}$, зависящие от конкретного сочетания условий эксперимента, т.е. от вектора $f_i = (f_{i1}, \dots, f_{ik})$; $\varepsilon_i(f_i)$ – случайные ошибки, имеющие нулевые средние значения.

Осн. назначение К.а. – использование в построении статистических оценок $\theta_1, \theta_2, \dots, \theta_k, \beta_1, \beta_2, \dots, \beta_p$ и статистических критериев для проверки различных гипотез относительно значений этих параметров. Если в модели (1) постулировать априори $\beta_1 = \beta_2 = \dots = \beta_p = 0$, то получится модель *дисперсионного анализа*; если из (1) исключить влияние неколичественных факторов (положить $\theta_1 = \theta_2 = \dots = \theta_k = 0$), то получится модель *регрессионного анализа*.

Своим названием К.а. обязан тому обстоятельству, что в его вычислениях используются разбиения ковариации величин Y и X точно так, же как в дисперсионном анализе используются разбиения суммы квадратов отклонений.

КОМПОЗИЦИЯ РАСПРЕДЕЛЕНИЙ

сумма распределений независимых *случайных величин*.

Пусть независимые случайные величины ξ_1 и ξ_2 имеют плотности вероятности соответственно $f_{\xi_1}(x)$ и $f_{\xi_2}(y)$. Композиция этих плотностей есть плотность распределения случайной величины $\eta=\xi_1+\xi_2$. По существу, следует рассматривать совместное двумерное распределение $f_{\xi_1,\xi_2}(x,y)$ и для определения функции распределения случайной величины η найти в плоскости xOy область возможных

значений (ξ_1, ξ_2), соответствующих событию $\{\eta < z\}$. Получается:

$$F_\eta(z) = P\{\eta < z\} = P\{\xi_1 + \xi_2 < z\} =$$
$$= \iint\limits_a f_{(\xi_1,\xi_2)}(x,y)\,dxdy = \iint\limits_a f_{(\xi_1)}(x) f_{(\xi_2)}(y)\,dxdy =$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_{(\xi_1)}(x) f_{(\xi_2)}(y)\,dxdy = \int_{-\infty}^{\infty} f_{(\xi_1)}(z)\left(\int_{-\infty}^{z-x} f_{(\xi_2)}(y)\,dx\right)$$

Т.о., формула композиции двух распределений, называемая формулой свёртки, имеет вид:

$$f_\eta = \int_{-\infty}^{\infty} f_{\xi_1}(x) f_{\xi_2}(z-x)dx$$
.

Для обозначения композиции (свёртки) законов распределения часто применяют символическую запись: $f_\eta = f_{\xi_1} \cdot f_{\xi_2}$.

В частности, формула плотности суммы $\eta_n = \xi_1 + \ldots + \xi_n$ равномерно распредёлённых на отрезке [0, 1] величин ξ_i, i = 1, 2, …, n., имеет вид:

$$f_{\eta n} = \begin{cases} \dfrac{1}{(n-1)!} x^{n-1} & при\ 0 \le x \le 1; \\ \dfrac{1}{(n-1)!}\left[x^{n-1} - C_n^1 (x-1)^{n-1}\right] & при\ 1 < x \le 2; \\ \dfrac{1}{(n-1)!}\left[x^{n-1} - C_n^1 (x-1)^{n-1} + C_n^2 (x-2)^{n-1}\right] & при\ 2 < x \le 3; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ \dfrac{1}{(n-1)!}\left[x^{n-1} - C_n^1 (x-1)^{n-1} + \right. & \\ \left. \ldots + (-1)^{n-1} C_n^{n-1} (x-(n-1))^{n-1}\right] & при\ n-1 < x \le n \end{cases}$$

КОМПОНЕНТНЫЙ АНАЛИЗ

см. в ст. *Метод главных компонент*

КОРРЕЛЯЦИЯ

зависимость между случайными величинами, не имеющая строго функционального характера, при которой изменение одной из случайных величин приводит в случае корреляционной зависимости к изменению только математического ожидания другой, а в общем случае стохастической зависимости – к изменению закона распределения последней.

Характеристику степени корреляционной зависимости выбирают в зависимости шкалы измерения анализируемых величин. Если все признаки количественные, то степень линейной зависимости определяют с помощью парных, частных и множественных коэффициентов корреляции или их квадратов – коэффициента детерминации. Адекватность коэффициентов корреляций тем выше, чем точнее выполняется условие линейности связей. Обычно это сводится к требованию, чтобы совместное распределение анализируемых признаков подчинялось многомерному нормальному закону.

Степень нелинейной корреляционной зависимости между двумя количественными признаками определяют с помощью корреляционного отношения. Величина распределения между корреляционными отношениями и коэффициентом детерминации свидетельствует о нелинейности связи.

Для тесноты связи между двумя признаками, измеренными в шкале порядка, используют

ранговые коэффициенты корреляции Спирмена и Кендалла. В качестве измерителя тесноты связи между несколькими порядковыми переменными используют коэффициент конкордации (согласованности).

Тесноту связи между двумя номинальными (атрибутивными) признаками, значения которых можно классифицировать, но не ранжировать (например, профессия работающего), используют коэффициент квадратичной сопряженности или информационную меру связи.

КОРРЕЛЯЦИОННАЯ МАТРИЦА

матрица, составленная из парных коэффициентов корреляции, имеет вид:

$$\mathrm{P} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ \rho_{l1} & \dots & \rho_{lj} & \dots & \rho_{lk} \\ \dots & \dots & \dots & \dots & \dots \\ \rho_{k1} & \dots & \dots & \rho_{k,k-1} & 1 \end{pmatrix}$$

где $\rho_{i,j}$ – *коэффициент корреляции* или *парный коэффициент корреляции*, который представляет собой *коэффициент ковариации* нормированных *случайных величии*:

$$\rho_{i,j} = M\left(\frac{\xi_i - M\xi_i}{\sigma_i}\frac{\xi_j - M\xi_j}{\sigma_j}\right) = \frac{\sigma_{ij}}{\sigma_i\sigma_j},$$

где σ_i, σ_j – средние квадратические отклонения величин ξ_i и ξ_j.

Оценку К.м. можно получить по формуле:

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ r_{21} & 1 & r_{23} & \dots & r_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{l1} & \dots & r_{lj} & \dots & r_{lk} \\ \dots & \dots & \dots & \dots & \dots \\ r_{k1} & \dots & \dots & r_{k,k-1} & 1 \end{pmatrix},$$

где элементы $r_{i,j}$ получают из элементов *ковариационной матрицы* с помощью нормировки:

$$r_{lj} = \frac{S_{lj}}{\sqrt{S_{ll}S_{jj}}},$$

где $S_{lj} = \frac{1}{n}\sum_{i=1}^{n}(x_{il} - \bar{x}_l)(x_{ij} - \bar{x}_j); l, j = 1, k$ –

коэффициент ковариации,

$S_{ll} = \frac{1}{n}\sum_{i=1}^{n}(x_{ll} - \bar{x}_l)^2$ – *выборочная дисперсия*.

Или:

$$R = \frac{1}{n}Z^T Z, \quad Z = \begin{pmatrix} 1 & z_{12} & z_{13} & \dots & z_{1k} \\ z_{21} & 1 & z_{23} & \dots & z_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ z_{l1} & \dots & z_{lj} & \dots & z_{lk} \\ \dots & \dots & \dots & \dots & \dots \\ z_{k1} & \dots & \dots & z_{k,k-1} & 1 \end{pmatrix},$$

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}}.$$

К.м. описывает характер и структуру статистических взаимосвязей, существующих между компонентами анализируемого многомерного признака. Матрица служит основой для вычисления частных и *множественных коэффициентов корреляции*. Для получения адекватных результатов требуется, чтобы все признаки были количественными и подчинялись совместному *многомерному нормальному закону распределения*.

К.м. симметричная (совпадает со своей транспонированной) и положительно определена (её определитель и все гл. миноры – положительны). Используется в многомерном статистическом анализе.

КОРРЕЛЯЦИОННОЕ ОТНОШЕНИЕ

мера причинной связи, применяемая для измерения тесноты корреляционной связи между признаками как при линейной, так и нелинейной связи. К.о. показывает, какую часть общей колеблемости результативного признака вызывает изучаемый фактор. Соответственно, этот показатель может рассчитываться на основе отношения межгрупповой дисперсии, вызванной влиянием признака – фактора (факторной дисперсии) δ^2 к общей дисперсии результативного признака σ^2_y. Формула эмпирического К.о.:

$$\eta = \sqrt{\frac{\delta^2}{\sigma_y^2}}.$$

Вычисление К.о. требует достаточно большого объёма информации, которая может быть представлена в форме табл., в которой группировка данных проведена по признаку – фактору. Группы образуются значениями *дискретной* (номинальной, порядковой, количественной) независимой *переменной* либо интервалами, в которые сгруппированы значения непрерывной независимой переменной.

К.о. определяется также на основе *уравнения регрессии*. Теоретическое К.о. рассчитывается по формуле:

$$\eta = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}},$$

где $\hat{y}_i$ – предсказанные значения переменной *y*, полученные по уравнению регрессии, $\bar{y}$ – среднее значение переменной *y*, y_i – исходные (фактические) значения переменной *y*. К.о. изменяется в интервале [0; 1]: чем ближе показатель к 1, тем теснее связь, и наоборот.

КОРРЕЛЯЦИОННОЕ ПОЛЕ

вспомогательное средство при анализе выборочных данных. С помощью К.п. статистические данные, характеризуемые двумя признаками могут быть представлены графически. При построении К.п. на оси абсцисс откладывается значение факторного признака, а по оси ординат – результирующего. Каждая точка характеризует выборочный объект наблюдения. По характеру расположения точек поля можно составить предварительное мнение о виде статистической зависимости между случайными величинами (см. рис. A, B, C, D).

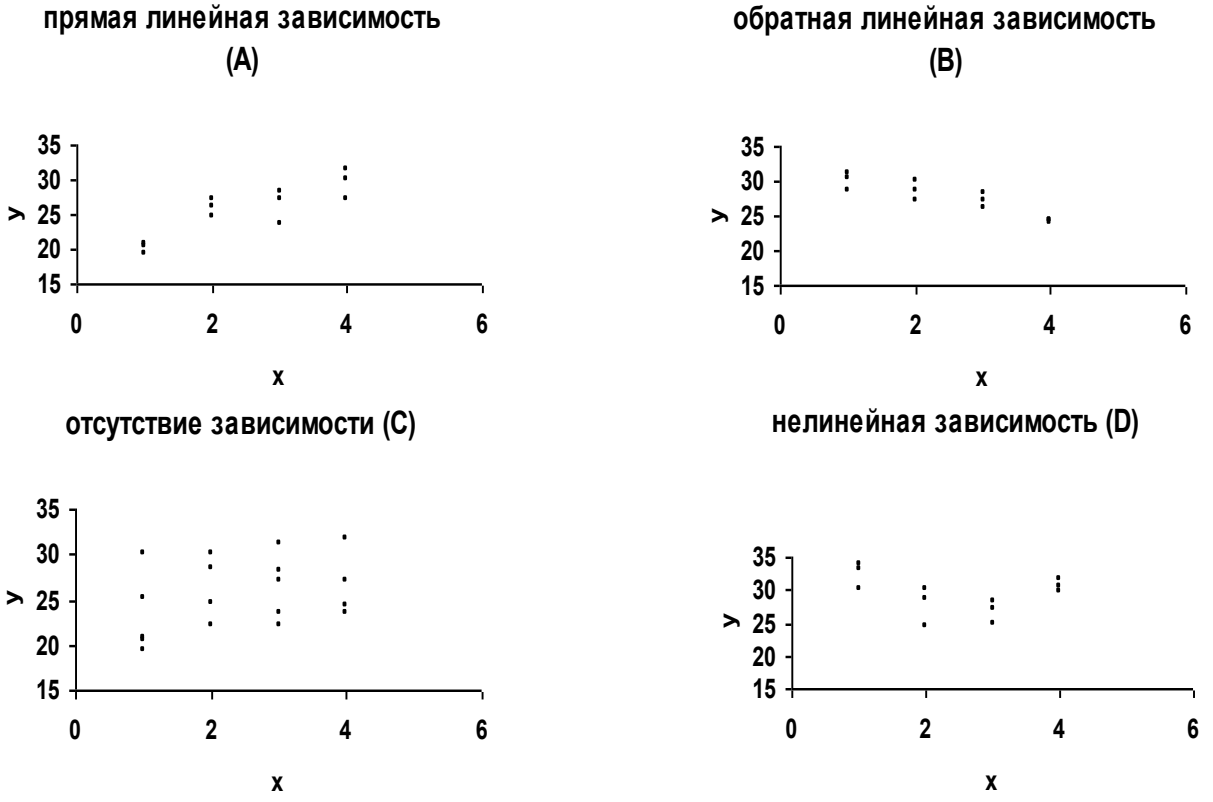


График на рисунке А – пример прямой линейной зависимости: при увеличении факторного признака x также увеличиваются среднее значения результирующего признака *y*, причем линейно. График В показывает пример обратной линейной зависимости, на котором при увеличении факторного показателя x линейно уменьшается среднее значение y. На графике С

представлен пример отсутствия корреляционной зависимости между *x* и *y*. Наконец, график D – случай нелинейной статистической зависимости между факторной и результирующей переменными. По мере увеличения *x* среднее значение результирующей переменной *y* сначала уменьшается, а затем увеличивается.

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

совокупность статистических методов исследования тесноты и структуры взаимосвязи между случайными величинами.

К.а. многомерной ген. совокупности рассматривает вопросы: выбора с учётом специфики и природы анализируемых переменных подходящего измерителя тесноты статистической связи (*коэффициент корреляции, корреляционное отношение, коэффициент ранговый корреляции* и т.д.); построения *точечной оценки* и *оценки интервальной* его числового значения по имеющимся выборочным данным; проверки гипотезы о том, что полученное числовое значение анализируемого измерителя тесноты связи действительно свидетельствует о наличии статистической связи, т.е. проверки исследуемой корреляционной характеристики на статистически значимое отличие от нуля; определение структуры связей между компонентами исследуемого многомерного признака, когда каждой паре компонент ставится в соответствие ответ: связь есть или нет.

В ряде случаев решают задачи сравнения исследуемого коэффициента корреляции со стандартом или сравнения коэффициентов корреляции между определёнными признаками различных совокупностей. Иногда приходится находить обобщённый (для нескольких выборок) коэффициент корреляции.

Характеристики статистической связи, рассматриваемые в К.а. используются в качестве «входной» информации при решении других осн. задач *эконометрики* и *многомерного статистического анализа*: определение вида зависимостей (*регрессионный анализ*); снижение размерности анализируемого признакого пространства (*факторный* и компонентный *анализ*); классификация объектов и признаков (методы многомерной классификации).

Поэтому с К.а. и начинаются все многомерные статистические исследования.

КОРРЕЛЯЦИЯ ЛОЖНАЯ

состоит в следующем эффекте. Анализ исходных данных указывает на существование значимой связи между исследуемыми признаками. Но по своему содержательному смыслу эти признаки не должны коррелировать между собой. При этом может искажаться не только теснота, но и направление связи. Обычно это вызвано влиянием на изучаемую связь других признаков, не учтённых в модели. Т.е. причина возникновения этого эффекта в не вполне правильной формализации решаемой содержательной задачи. Выявляется сравнением соответствующих *парных и частных коэффициентов корреляции*, что позволяет охарактеризовать влияние зафиксированных признаков на связь двух осн. Если парный и частный коэффициенты существенно различаются, то делается вывод о сильном влиянии зафиксированных признаков на связь двух осн. Может быть поставлена и решена задача поиска набора признаков, наиболее существенно влияющих на связь этих признаков. Эта идея получила развитие в виде методов снижения размерности (*метод гл. компонент и факторный анализ*). Термин введён Пирсоном при исследовании связи между относительными показателями (индексами). Признаки: X_1, X_2, X_3 могут быть взаимно независимыми, но при этом новые признаки: $\frac{X_1}{X_3}$ и $\frac{X_2}{X_3}$ могут сильно коррелировать между собой.

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

квадрат соответствующего *коэффициента корреляции – парного, частного или множественного* – указывает долю *дисперсии* одного (результативного) признака, которую можно объяснить изменением другого признака (других, факторных) при линейной связи признаков; при этих условиях совпадает с квадратом *корреляционного отношения*; меняется в преде-

лах от 0 до 1; используется в *критерии Стьюдента* и Фишера – Снедекора при проверке значимости коэффициента корреляции. Различие характеризует степень нелинейности связи. В множественной регрессионной модели изменение множественного (исправленного) *коэффициента детерминации* характеризует информативность регрессора, вновь включённого в модель или исключенного из неё, на очередном шаге.

КОЭФФИЦИЕНТ СОПРЯЖЁННОСТИ (ПИРСОНА)

математическая мера связи двух признаков. Как правило, К.с. используется для признаков, измеренных в номинальной шкале. Значение коэффициента основано на статистике χ^2 и рассчитывается по данным *табл. сопряжённости* в соответствии с формулой:

$$P = \sqrt{\frac{\chi^2_{набл}}{\chi^2_{набл} + n_{**}}} = \sqrt{\frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i*}n_{*j}} - 1}{\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i*}n_{*j}}}},$$

где n_{ij} – наблюдаемая частота в ячейке табл. сопряжённости, соответствующей i–й категории первого признака и j–й категории второго,

n_{i*} – маргинальная частота i–й категории первого признака: $n_{i*} = \sum_{j=1}^{s} n_{ij}$,

n_{*j} – маргинальная частота j–й категории второго признака: $n_{*j} = \sum_{i=1}^{r} n_{ij}$,

n_{**} – общий объём выборки: $n_{**} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$.

С точки зрения анализа взаимосвязи между признаками К.с. удобнее в сравнении со статистикой χ^2, т.к. он является нормированной величиной и принимает значения в интервале от 0 до 1, что упрощает его интерпретацию: если К.с. равен 0, это означает, что рассматриваемые признаки статистически независимы; чем больше значение К.с., тем сильнее связь между признаками.

К недостаткам данного коэффициента следует отнести, во-первых, тот факт, что он никак не характеризует направление связи между признаками. Во-вторых, значение К.с. не достигает 1 даже при полной связи признаков. В-третьих, макс. значение К.с. зависит от размера табл. сопряжённости, что ограничивает его возможности по сравнению тесноты взаимосвязи между признаками для табл. разного размера.

Л

ЛАТЕНТНО-СТРУКТУРНЫЙ АНАЛИЗ

метод вероятно-статистического моделирования, идея которого основана на предположении, что наблюдаемое поведение (напр., ответы индивидов на вопросы теста или анкеты) есть внешнее проявление некоторой скрытой (латентной) характеристики, присущей индивидам. Задача метода заключается в том, чтобы, изучив наблюдаемое поведение индивидов, вывести эту скрытую характеристику и разделить (классифицировать) индивидов по сходству (равенству) ее значений. Метод возник в кон. 40-х – нач. 50-х гг. 20 в. Логические и математические основания метода были изложены в работах американского социолога П. Лазарсфельда.

При такой постановке задачи исходной координатной системы не существует, а подлежащее статистическому анализу и моделированию данные представлены в виде матрицы парных сравнений в статическом варианте, т.е. описывающей ситуацию в один какой-то фиксированный момент времени. Ставится задача: для заданной, сравнительно невысокой размерности p`, матрицы определить вспомогательные условные координатные оси и способ сопоставления каждому объекту его координат в этой системе т.о., чтобы попарные отношения (попарные взаимные расстояния) между объектами, вычисленные на базе этих условных координат, в определённом смысле минимально отличались от заданных величин. Построенные т.о. условные переменные поддаются содержательной интерпретации и могут рассматриваться в качестве латентных характеристик определенных свойств анализируемых объектов. Снижение размерности происходит здесь в том

смысле, что от исходного массива информации размерности n x n переходят к матрице типа «объект – свойство» размерности p`, где p` < n.

ЛАТИНСКИЙ КВАДРАТ

квадратная матрица порядка n, каждая строка и каждый столбец которой являются перестановками элементов конечного множества S, состоящего из n элементов. Говорят, что Л.к. построен на множестве S. Число n называют порядком Л.к.

Л.к. существует для любого n. Напр., A = $||a_{ij}||$, где $a_{ij} = i + j - 1 (\mathrm{mod}\ n)$, i, j = 1, 2, …, n, есть Л.к. Каждый Л.к. можно рассматривать как таблицу умножения квазигруппы, верно и обратное: таблица умножения конечной квазигруппы есть Л.к. Для числа L_n Л.к. порядка n верна оценка снизу: $L_n \geq n!\,(n-1)!\,\ldots\,2!\,1!$.

Два Л.к., построенные на одном и том же множестве S, называются эквивалентными, если один из другого получается перестановкой строк, столбцов и переименованием элементов.

Два Л.к. A = $||a_{ij}||$ и B = $||b_{ij}||$ порядка n называются ортогональными, если $(a_{ij}, b_{ij}) \neq (a_{kl}, b_{kl})$ при $(i, j) \neq (k, l)$, i, j, k, l ∈ S = (1, …, n), где (a_{ij}, b_{ij}) – элемент на пересечении i-ой строки и j-го столбца матрицы, полученной наложением двух Л.к. A и B. Для всех n > 2, n ≠ 6, имеются примеры пар ортогональных Л.к., а для n = 6 путём перебора всех возможностей доказано, что таких пар нет. Несколько Л.к. одного порядка называются попарно ортогональными, если любые два из них ортогональны. Если N(n) – максимально возможное число попарно ортогональных Л.к., то $N(n) \leq n - 1$. Получены также оценки снизу для N(n) (см. табл.):

Таблица

n	≥	7	52	53	63	90
N(n)	≥	2	3	4	5	6

Кроме того, $N(12) \geq 5$, $N(33) \geq 3$, $N(35) \geq 4$, $N(40) \geq 4$, $N(45) \geq 4$. Доказано, что $N(n) \to \infty$, при $n \to \infty$.

Множество из n - 1 попарно ортогональных Л.к. порядка n называется полным. Полные множества попарно ортогональных Л.к. находят применение в планировании экспериментов, при построении симметричных блок-схем; они могут интерпретироваться и как конечные проективные плоскости. Существуют много методов построения ортогональных Л.к. Все они созданы с целью получения как можно большего множества попарно ортогональных Л.к. порядка n. Приложения ортогональных Л.к. в статистике, теории информации и планировании экспериментов требуют построения ортогональных Л.к. специального вида и перенесения понятия ортогональности на другие объекты.

М

МАТЕМАТИКО-СТАТИЧЕСКИЕ МЕТОДЫ

методы обработки и анализа статистических данных (т.е. сведений о числе объектов, обладающих определёнными признаками, в какой-либо более или менее обширной совокупности). Они строятся безотносительно к тому, какие статистические данные обрабатываются (физические, экономические и др.), однако обращение с ними требует обязательного понимания сущности явления, изучаемого с их помощью. В общем случае анализ данных М.-с.м. позволяет сделать два вывода: либо вынести искомое суждение о характере и свойствах этих данных или взаимосвязей между ними, либо доказать, что собранных данных недостаточно для такого суждения. Причём выводы могут делаться не из сплошного рассмотрения всей совокупности данных, а из её выборки, как правило, случайной (последнее означает, что каждая единица, включенная в выборку, могла быть с равными

шансами, т.е. с равной вероятностью, заменена любой другой).

Центральное понятие М.-с.м. – *случайная величина*– всякая наблюдаемая величина, изменяющаяся при повторениях общего комплекса условий, в которых она возникает. Если сам по себе набор, перечень значений этой величины неудобен для их изучения (поскольку их много), М.-с.м. дают возможность получить необходимые сведения о случайной величине, зная существенно меньшее количество её значений. Это объясняется тем, что статистические данные подчиняются таким законам распределения (или приводятся к ним порой искусственными приемами), которые характеризуются всего лишь несколькими параметрами, т. е. характеристиками. Зная их, можно получить столь же полное представление о значениях случайной величины, какое даётся их подробным перечислением в очень длинной табл. Характеристики распределения – среднее, *медиана*, *мода* и т.п.

М.-с.м. – широкий круг одно- и многомерных методов и правил обработки статистических данных: от простых приёмов статистического описания (выведение средней, а также степени и характера разброса исследуемых признаков вокруг неё, группировка данных по классам и сопоставление их характеристик и т. д.), правил отбора фактов при выборочном их рассмотрении до сложных методов исследования зависимостей между случайными величинами: выявление связей между ними – *корреляционный анализ;* оценка величины случайной переменной, если величина другой или других известна – *регрессионный анализ*; выявление наиболее важных скрытых факторов, влияющих на изучаемые величины – *факторный анализ*; определение степени влияния отдельных неколичественных факторов на общие результаты их действия (напр., в научном эксперименте) – *дисперсионный анализ*. Кроме того, ещё *дискриминантный анализ*, *кластерный анализ* и другие М.-с.м., как правило, не опирающиеся на предпосылку о вероятностном характере исследуемых зависимостей. В частности, дискриминантный анализ предназначен для решения задач, связанных с разделением совокупностей наблюдений (элементарных данных). Если у исследователя имеется по одной выборке из каждой неизвестной ему ген. совокупности (такую выборку называют “обучающей”), то с помощью методов дискриминантного анализа удается приписать некоторый новый элемент (наблюдение x) к своей ген. совокупности. Кластер-анализ позволяет разбивать исследуемую совокупность элементов (координаты которых известны) т.о., чтобы элементы одного класса находились на небольшом расстоянии друг от друга, в то время как разные классы были бы на достаточном удалении друг от друга и не разбивались бы на столь же взаимоудалённые части. Для экономических исследований большое значение имеет также анализ стохастических процессов, в т.ч. “марковских процессов”.

Задачи, решаемые с помощью М.-с.м. в экономике можно разделить на пять осн. типов: 1. оценка статистических данных; 2. сравнение этих данных с каким-то стандартом и между собой (оно применяется при эксперименте или, напр., в контроле качества на пр-тиях); 3. формирование групп данных и исследование связей между статистическими данными и их группами. Эти три типа позволяют вынести суждение описательного характера об изучаемых явлениях, подверженных по каким-то причинам искажающим случайным воздействиям. Следующий четвертый, тип задач связан с нахождением наилучшего варианта измерения изучаемых данных. И наконец, пятый тип задач связан с проблемами предвидения и развития: здесь важное место занимают задачи анализа временных рядов.

Для экономики особенно ценно то, что М.-с.м. позволяют на основании анализа течения событий в прошлом, т.е. изучения выбранных на определённые даты сведений о характерных чертах системы, предсказать вероятное развитие изучаемого явления в будущем (если внешние или внутренние условия существенноне изменятся). Различные М.-с.м. применяются в управлении хозяйственными и производственными процессами. На них основаны многие методы исследования операций: методы теории массового обслуживания, позволяющие наиболее эффективно организовывать ряд процессов

производства и обслуживания нас.; методы теории расписаний, предназначенной для выработки оптимальной последовательности производственных, транспортных и других операций; методы теории решений, теории управления запасами, а также теории планирования эксперимента и выборочного контроля качества продукции; сетевые методы планирования и управления.

В эконометрических исследованиях на основе математико-статистической обработки данных строятся *модели экономико-математические* (*экономико-статистические*) экономических процессов, производятся экономические и технико-экономические прогнозы. Экономико-математические модели – осн. средство модельного исследования экономики. Модель описывает либо внутреннюю структуру объекта, либо (если структура неизвестна) его поведение, т.е. реакцию на воздействие известных факторов (принцип «чёрного ящика»). Один и тот же объект может быть описан различными моделями в зависимости от исследовательской или практической потребности, возможностей математического аппарата и т.п. Поэтому всегда необходима оценка модели и области, в которой выводы из её изучения могут быть достоверны. Во всех случаях необходимо, чтобы модель содержала достаточно детальное описание объекта, позволяющее, в частности, осуществлять измерение экономических величин и их взаимосвязей, чтобы были выделены факторы, воздействующие на исследуемые показатели. Кроме того, полезно записать условия, в которых она действительна, т. е. ограничения модели. Большое значение в экономике имеют оптимизационные или оптимальные модели – системы уравнений, равенств и неравенств, которые кроме ограничений (условий) включают также особого рода уравнение, называемое функционалом, или критерием оптимальности.

Широкое распространение М.-с.м. в общественном произ-ве, а также в других областях социально-экономической жизни общества (здравоохранение, экология, естественные науки) опирается на развитие электронно-вычислительной техники. Для решения типовых задач математико-статистической обработки данных созданы и применяются многочисленные стандартные прикладные компьютерные программы

МАТРИЦА

прямоугольная табл. чисел

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mj} & \dots & a_{mn} \end{pmatrix}$$

такая табл., содержащая *m* строк и *n* столбцов, называется прямоугольной матрицей размера m × n, или (m × n)-матрицей, с элементами aij (i − номер строки, j − номер столбца). Наряду с круглыми скобками используются и другие обозначения М.: [ ], || ||. Сокращённо М. обозначается $A_{m \times n} = (a_{ij})$, или $A = (a_{ij})$,i = 1, 2,…, m; j = 1, 2,…, n.

М. $(a_{i1}\ a_{i2}\ \dots\ a_{in})$, состоящая из одной строки (m = 1), называется М. (вектором) – строкой или просто строкой, а М. $\begin{pmatrix} a_{1j} \\ a_{2j} \\ \dots \\ a_{mj} \end{pmatrix}$ состоящая из одного столбца (*n* = 1) – М. (вектором) – столбцом или просто столбцом. Если m = n, то такая матрица наз. *матрицей квадратной,* а число *n* – её порядком. Если в М. все элементы равны нулю, то такая матрица наз. нулевой, или нуль – М. (обозначается символом 0). Если в М. A поменять местами строки и столбцы (с сохранением их порядка), то полученная М. называется транспонированной (обозначается A' или A^T). Если A = A′, то матрица A называется *М. симметрической* (симметричной). Две М. одного размера $A_{m \times n} = (a_{ij})$ и $B_{m \times n} = (b_{ij})$ считаются равными, если их соответствующие элементы равны, т.е. $a_{ij} = b_{ij}$, i =1, 2,…, m; j = 1, 2,…, n.

Операции над М.

1. Произведение М. $A_{m \times n} = (a_{ij})$ на число λ – М. того же размера $B_{m \times n} = (b_{ij})$, элементы которой $b_{ij} = a_{ij}\lambda$, i = 1, 2,…, m; j = 1, 2,…, n.

2. Сумма двух М. одного размера $A_{m\times n} = (a_{ij})$ и $B_{m\times n} = (b_{ij})$ есть М. того же размера $C_{m\times n} = (c_{ij})$, элементы которой $c_{ij} = a_{ij} + b_{ij}$, i = 1, 2,…, m; j = 1, 2,…, n.

3. Произведение М. $A_{m\times n} = (a_{ij})$ на М. у и $B_{n\times p} = (b_{ij})$ есть М. $C_{m\times p} = (c_{ij})$, каждый элемент которой

$$c_{ij} = \sum_{i=1}^{n} a_{ik} b_{kj}\ \text{j, i} = 1, 2,\ldots, \text{m; j} = 1, 2,\ldots, \text{n.}$$

Следовательно, для умножения М. А и В должны быть согласованы, т.е. число столбцов М. А должно быть равно числу строк М. В. Напр.,

$$\begin{pmatrix} 2 & 7 & 8 \\ 9 & 2 & 0 \end{pmatrix}\begin{pmatrix} 3 & 4 \\ 1 & 2 \\ 5 & 6 \end{pmatrix}\begin{pmatrix} 2\cdot 3+7\cdot 1+8\cdot 5 & 2\cdot 4+7\cdot 2+8\cdot 6 \\ 9\cdot 3+2\cdot 1+0\cdot 5 & 9\cdot 4+2\cdot 2+0\cdot 6 \end{pmatrix}=\begin{pmatrix} 53 & 70 \\ 29 & 40 \end{pmatrix}.$$

В частном случае, AE = EA = E, где E – единичная М. Операции с М. обладают свойствами:

1) A + B = B + A;

2) (A + B) + C = A + (B + C);

3) λ(A + B) = λA + λB;

4) A(B + C) = AB + AC;

5) (A + B)C = AC + BC;

6) λ(AB) = (λA)B = A(λB);

7) A(BC) = (AB)C;

8) (λA)′ = λA′;

9) (A+B)′ = A′ +B′; 10)(AB)′ = B′A′.

К особенностям операций с М. следует отнести то, что в общем случае коммутативный (переместительный) закон умножения М. не выполняется т.е., вообще говоря, AB ≠ BA, а, например, из равенств AB = 0, AB = AC и т.п. вовсе не обязательно следует, что A(или B) = 0 и B = C. *М. квадратные*, для которых коммутативный закон умножения все же выполняется, называются перестановочными.

Числовые функции М.

1. Определитель (детерминант) квадратной М. n-го порядка обозначается $|A_n|$ (или Δ_n, det A). При n = 1 $|A_1| = a_{11}$, при

n = 2 $|A_2| = a_{11}a_{22} - a_{21}a_{12}$, и т.д. Напр.,

$$|A_2| = \begin{vmatrix} 2 & 7 \\ 5 & 9 \end{vmatrix} = 2\cdot 9 - 5\cdot 7 = -17.$$

Алгебраическим дополнением элемента a_{ij} М. А = (a_{ij}) n-го порядка называется взятый со знаком $(-1)^{i+j}$ определитель М. (n−1)-го порядка, полученной из М. А вычеркиванием i-ой строки и j-го столбца.

2. Если в М. А размера m × n выделить какие-либо k строк и k столбцов (k≤ min (m,n)), то определитель подматрицы из получаемых на пересечении этих строк и столбцов элементов называется минором k -го порядка М. *Рангом М.* (обозначается rang (A), или rank (A), или r(A)) называется наивысший порядок её миноров, отличных от нуля. Напр., ранг М.

$A = \begin{pmatrix} 4 & 0 & 2 & 0 \\ 2 & 0 & 1 & 0 \\ 6 & 0 & 3 & 0 \end{pmatrix}$ равен 1, т.е. r (*A*) = 1, т.к. все миноры второго и третьего порядков: $\begin{vmatrix} 4 & 0 \\ 2 & 0 \end{vmatrix}, \begin{vmatrix} 0 & 0 \\ 0 & 0 \end{vmatrix}, \begin{vmatrix} 4 & 2 \\ 2 & 1 \end{vmatrix}$ и т.д. равны нулю.

М. А размера m × n называется М. полного ранга, если rang (A)= min (m, n). Ранг М. равен макс. числу её линейно независимых строк (или столбцов). Ранг М. не меняется при элементарных преобразованиях (при отбрасывании нулевой строки (столбца), перестановках строк или столбцов, умножении строки или столбца на отличное от нуля число, при сложении строк или столбцов), Свойства ранга М.:

1) 0 ≤ r (A) ≤ min (m, n); 2) r (A) = 0, если A = 0; 3) r (A_n), = n, если $|A_n| \neq 0$; 4) | r (A) − r (B)| ≤ r (A + B) ≤ r (A) + +r (B); 5) r (AB) ≤ min (r (A), r (B)); 6) r (AB) = r (A), если | B | ≠ 0;)); 7) r (A) = r (AA′) = r (A′A) (причём М. AA′ и A′A − квадратные М. соответственно m-го и n-го порядков).

3. Следом квадратной М. А n-го порядка (обозначается tr (A) − от англ. «trace» − след или sp (A) − от немец. «spur» − след) называется сумма её диагональных элементов:

$$\text{tr (A)} = \sum_{i-1}^{n} a_{ii}\ .$$

Свойства следа М.:

1) tr (E_n) = n; 2) tr (λA) = λtr (A); 3) tr (A′) = tr (A); 4) tr (A + B)) = tr (A) + tr (B); 5) tr (AB) = tr (BA).

Все приведённые здесь определения и утверждения дословно повторяются для М., элементами которых являются комплексные числа.

Используемые на практике *модели экономико-статистические* часто предназначены для описания взаимосвязи экономических структур, их динамики во времени, зависимости от ряда факторов и т.п. Матричный анализ – один из наиболее компактных способов описания и исследования таких структур. Операции с М. отличаются краткостью и простотой. Применение М. не только позволяет «экономно» формализовать поставленную проблему, но и использовать в статистических расчётах преимущества матричной алгебры. Особенно привлекателен почти универсальный язык матричных выражений, позволяющий применять одни и те же методы анализа независимо от величины массива исходных данных.

Впервые М. как математическое понятие появилось в работах У. Гамильтона, А. Кали и Дж. Сильвестра в середине 19 в. Основы теории М. созданы К.Вейерштрассом и Г. Фробениусом во 2-й пол. 19 в. и нач. 20 в. В совр. статистических исследованиях, в связи с развитием компьютерных технологий, позволяющих проводить обработку больших массивов исходных данных, роль методов матричной алгебры возрастает.

МАТРИЦА ИДЕМПОТЕНТНАЯ

матрица A, совпадающая со своим квадратом, т.е. $A = A^2$. Обычно в статистических приложениях полагают, что матрица A является также м*атрицей симметрической*. Свойства М.и. A: 1) $A^k = A$, $k \in N$; 2) собственные значения A λ = 0 или λ = 1; 3) A > 0; 4) r (A) = tr (A) = k (где k – число собственных значений матрицы A, равных единице).

МАТРИЦА КВАДРАТНАЯ

Если в М.к. все элементы a_{ij}($i>j$ ($i<j$)), расположенные под (над) её гл. диагональю, равны нулю, то такая называется верхней (нижней) треугольной матрицей. Если в М.к. все элементы a_{ij} ($i\neq j$), расположенные вне гл. диагонали, равны нулю, то такая матрица D наз. диагональной: D= diag ($a_{11}, a_{22}, \ldots, a_{nn}$). Если в матрице диагональной все элементы a_{ii} на гл. диагонали равны, то полученная матрица называется скалярной, если – равны единице, то – матрицей единичной,

напр.,

$$E_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

– единичная матрица второго порядка. Если в М.к. A определитель |A| ≠ 0, то матрица *A* наз. *матрицей невырожденной* (неособенной). В противном случае (при |A| = 0) A – матрица вырожденная (особенная). Матрица A^{-1} называется *матрицей обратной* по отношению к К.м. A, если $A^{-1}A = AA^{-1} = E$. Для существования матрицы обратной A^{-1} необходимо и достаточно, чтобы |A| ≠ 0, т.е. матрица A была невырожденной. Матрица обратная может быть найдена по формуле $A^{-1} = \breve{A} / |A|$, где $\breve{A}$ – матрица присоединённая, элементы которой равны алгебраическим дополнениям элементов матрицы A′, транспонированной к A.

Свойства обратной матрицы:

1) $|A^{-1}| = 1/|A|$; 2) $(A^{-1})^{-1} = A$; 3) $(A^m)^{-1} = (A^{-1})^m$; 4) $(AB)^{-1} = B^{-1}A^{-1}$; 5) $(A^{-1})' = (A')^{-1}$.

С помощью обратной матрицы получается решение системы линейных алгебраических уравнений (СЛАУ), широко используемых в линейных экономико-статистических моделях. В матричном виде СЛАУ имеет вид: AX = B, где A – матрица коэффициентов при переменных (матрица системы), X – столбец переменных, B – столбец свободных членов. Если матрица A невырожденная (|A| ≠ 0), то неизвестный столбец переменных находится по формуле $X = A^{-1}B$.

М.к. A и B одного порядка называются подобными, если существует невырожденная матрица S того же порядка, что $B = S^{-1}AS$. Одна из

задач матричного анализа состоит в замене матрицы A подобной ей матрицей B, имеющей возможно более простой вид.

Вектор $x = (x_1, x_2,\ldots, x_n)$ называется собственным вектором М.к. A, если найдётся такое число λ, что $Ax = \lambda x$. Число λ называется собственным значением (или собственным числом) матрицы A, соответствующим вектору x. Собственные значения матрицы A находятся из решения характеристического уравнения матрицы $|A - \lambda E| = 0$.

МАТРИЦА КОВАРИАЦИОННАЯ

см. ст. *Ковариационная матрица*

МАТРИЦА КОРРЕЛЯЦИОННАЯ

см. ст. Корреляционная матрица

МАТРИЦА НАГРУЗОК

матрица коэффициентов линейного преобразования, М.н. общих факторов на исследуемых признаки в *факторном анализе.*

Пусть имеются центрированные наблюдения $X_1, X_{2,\ldots,} X_n$, получаемые от исходных наблюдений с помощью переноса начала координат в центр исходного множества наблюдений. Наиболее распространенная в практике исследований линейная модель факторного анализа: $X = QF + U$, где X – исследуемый вектор наблюдений; Q – матрица нагрузок размера $(p \times p')$, F – вектор ненаблюдаемых (скрытых) общих факторов; U – вектор остаточных специфических факторов (определяющий ту часть каждого из исследуемых признаков, которая не может быть объяснена общими факторами); p – число анализируемых признаков, p' – число общих (скрытых) факторов $(p' < p)$.

В *методе гл. компонент* под М.н. понимается матрица $A = (a_{ij})$, i,j = 1,2,…, p гл. компонент на исходные признаки. Если анализируемые переменные процентрированы и пронормированы, то элементы М.н. a_{ij} определяют одновременно степень тесноты парной линейной связи между i-ой переменной и j-ой главной компонентой (т.е. парный коэффициент корреляции) и удельный вес влияния пронормированной j-ой гл. компоненты на i-ый признак. М.н. A определяется соотношением $A = L' \Lambda^{1/2}$, где L – матрица линейного преобразования,

$\Lambda^{1/2} = \mathrm{diag}\,(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_h})$, λ_j (j=1,2,…, p) – собственные значения *матрицы ковариационной* вектора гл. компонент. Отметим свойства М.н. A:

$$\sum_{j=1}^{p} a_{ij}^2 = 1,\ i=1,2,\ldots, p;\quad \sum_{i=1}^{p} a_{ij}^2 = \lambda_j,\ j=1,2,\ldots, p.$$

МАТРИЦА НЕВЫРОЖДЕННАЯ

см. в ст. *Матрица*

МАТРИЦА НЕОТРИЦАТЕЛЬНО ОПРЕДЕЛЁННАЯ

см. в ст. *Матрица положительно определённая*

МАТРИЦА ОБРАТНАЯ

см. в ст. *Матрица*.

МАТРИЦА «ОБЪЕКТ − СВОЙСТВО»

матрица результатов $X = (X_1, X_2, \ldots, X_n)$, статистического обследования объектов $(O_1, O_2, \ldots, O_n)$, в *многомерном статистическом анализе*. В этой матрице $X_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$ − вектор значений анализируемых признаков (свойств), зарегистрированных на обследованном *объекте* O_i *(i=1,2,…,n).* Результаты статистического обследования объектов могут даваться также в виде *матрицы парных сравнений* вида $A = (a_{ij})$, где элемент a_{ij} определяет результат сопоставления объектов O_i и O_j, напр., может выражать меру сходства или различия объектов O_i и O_j, меру их связи или взаимодействия в каком-либо процессе, геометрическое расстояние между объектами, отношение предпочтения ($a_{ij} = 1$, если объект O_i не хуже объекта O_j, и $a_{ij} = 0$ в противном случае) и т.д. При этом предполагается, что существует небольшое (в сравнении с p), число определяющих (типообразующих) факторов, с помощью которых могут быть достаточно точно описаны наблюдаемые характеристики анализируемых объектов (т.е. элементы матриц X и A) и характер связи между ни-

ми, а также искомая классификация самих объектов,. т.е. разделение рассматриваемой совокупности объектов на однородные (в определённом смысле) группы. При этом отмеченные определяющие факторы могут находиться среди статистически обследованных характеристик, а могут быть латентными (скрытыми), т.е. непосредственно статистически ненаблюдаемыми, но восстанавливаемыми по исходным данным в виде М.«о.-с.» и парных сравнений.

МАТРИЦА ОРТОГОНАЛЬНАЯ

матрица квадратная невырожденная A называется ортогональной, если транспонированная к ней *матрица* совпадает с обратной, т.е. $A' = A^{-1}$. Свойства ортогональной матрицы A: 1) $A'A = E$; 2)$|A| = 1$ или $|A| = -1$; 3) $A'BA = \Lambda$ и $B = A\Lambda A'$ (где B – некоторая симметрическая матрица; $\Lambda = \text{diag}\ (\lambda_1, \lambda_2, \ldots, \lambda_n)$; $\lambda_1, \lambda_2, \ldots, \lambda_n$ − собственные значения матрицы A).

МАТРИЦА ПАРНЫХ СРАВНЕНИЙ

см. в ст. *Матрица «объект – свойство».*

МАТРИЦА ПЕРЕХОДНЫХ ВЕРОЯТНОСТЕЙ

матрица квадратная $P=(p_{ij})$, элементами которой являются переходные вероятности однородной *марковской цепи* с конечным или счётным множеством состояний. *Случайный процесс*, протекающий в некоторой системе *S* с возможными состояниями $S_1, S_2, \ldots, S_n, \ldots$, называется марковским процессом, или случайным процессом без последействия, если для любого момента времени t_0 вероятностные характеристики процесса в будущем (при $t > t_0$) зависят только от его состояния в данный момент t_0 и не зависят от того, когда и как система пришла в это состояние; т.е. не зависит от ее поведения в прошлом (при $t < t_0$). Марковским случайным процессом с дискретными состояниями и дискретным временем, или марковской цепью, называется марковский процесс, в котором его возможные состояния происходят мгновенно (скачком), но только в определённые моменты времени $t_0, t_1, t_2, \ldots$, называемые шагами процесса. Если переходные вероятности p_{ij} случайного процесса (системы S) из состояния i в состояние j не зависят от шага процесса, то такая марковская цепь называется однородной. М.п.в. P_n из состояния в состояние однородной марковской цепи за n шагов определяется по формуле $P_n = P^n$, где P – М.п.в. из состояния в состояние за один шаг.

МАТРИЦА ПОЛНОГО РАНГА

см. в ст. *Матрица*

МАТРИЦА ПОЛОЖИТЕЛЬНО (НЕОТРИЦАТЕЛЬНО) ОПРЕДЕЛЁННАЯ

матрица симметрическая n-го порядка, если для любого ненулевого вектора $x = (x_1, x_2, \ldots, x_n)'$ выполняется неравенство $x'Ax > 0$ ($x'Ax \geq 0$). Неотрицательно определённая матрица называется также положительно полуопределённой. Для М.п.о. используется запись $A > 0$ ($A \geq 0$). Соотношение $A > B$ ($A \geq B$) означает, что матрица $A - B$ положительно (неотрицательно) определена. Свойства М.п.о.:

1) $A > B \Rightarrow a_{ii} > b_{ii}$, I = 1, 2,…, n; 2) $A > B$, $C \geq 0 \Rightarrow A + C > B$; 3) $A > B \Rightarrow B^{-1} > A^{-1}(|A| \neq 0, (|B| \neq 0)$;

4) $A > 0$ ($A \geq 0$) $\Rightarrow$ все собственные значения матрицы A $\lambda_i > 0$, i = 1,2,…, n.

МАТРИЦА СИММЕТРИЧЕСКАЯ

см. в ст. *Матрица*

МАТРИЦА ТРАНСПОНИРОВАННАЯ

см. в ст. *Матрица*

МЕРА БЛИЗОСТИ

функция, по значению которой определяется степень «похожести», близости между объектами (или группами объектов) и между признаками (двумя или группами признаков). М.б. – важный инструмент анализа данных. Вычисление М.б. – составная часть многих математиче-

ских методов, напр., *кластерного* и *дискриминантного анализа.* М.б. между признаками традиционно называется коэффициентами связи, между объектами – мерами сходства. Принято считать, что принципиальное отличие этих двух видов М.б. заключается в том, что объекты представляют собой элементы некоторой однородной совокупности, а природа признаков различна. В частности, связь между признаками может иметь отрицательный характер, что не имеет смысла при измерении близости между объектами. При использовании М.б. типа корреляции, признаки считаются тем более похожими, чем больше связь между ними.

При использовании М.б. типа расстояния (функции расстояния), объекты считаются тем более похожими, чем меньше расстояние между ними. Данные меры с помощью элементарных преобразований могут быть сведены к М.б. типа расстояния.

Применяются М.б., в зависимости от принадлежности параметров, описывающих объекты в различных шкалах измерения, для количественных признаков: *евклидово расстояние*, манхеттенское расстояние, супремум-норма, *расстояние Махаланобиса*, расстояние Пирсона; для порядковых признаков: расстояние Спирмэна, расстояние Кендалла; для номинальных признаков: расстояние Жаккара, расстояние Рассела-Рао, расстояние Бравайса, расстояние Юла, для смешанных и произвольных данных – расстояние отношений.

МЕРА ТОЧНОСТИ

характеристика рассеяния значений случайной величины. М.т. h связана с *отклонением средне-квадратическим* σ формулой

$$h = \frac{1}{\sigma\sqrt{2}}.$$

При наличии систематической ошибки М.т. выражается отношением:

$$h = \frac{1}{\sqrt{2(b^2 + \sigma^2)}},$$

где b – *математическое ожидание* систематической ошибки. Т.о., отсутствие систематической ошибки означает, что b = 0.

Этот способ измерения рассеяния объясняется тем, что в случае нормального распределения плотность вероятности случайной величины с М.т. h и математическим ожиданием a записывается формулой:

$$f(x) = \frac{h}{\sqrt{\pi}} e^{-h^2(x-a)^2}.$$

М.т. пользуются как характеристикой рассеяния гл. обр. в теории стрельбы и теории ошибок.

МЕТОД «БЛИЖАЙШЕГО СОСЕДА» (ОДИНОЧНОЙ СВЯЗИ)

метод *кластерного анализа*, при котором расстояние между кластерами рассчитывается как миним. расстояние из расстояний между всевозможными парами представителей этих кластеров. Расстояние, измеряемое по данному методу, находят используя формулу:

$$d_{\min}(s_l, s_m) = \min_{x_i \in s_l, x_j \in s_m} d(x_i, x_j).$$

Графически это можно проиллюстрировать (см. рис.):

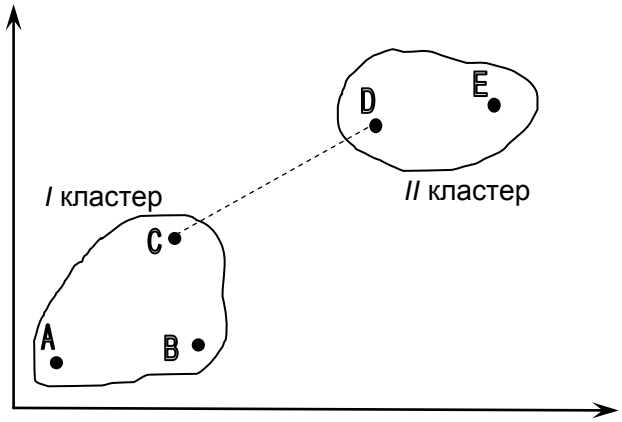


Рис. **Графическая иллюстрация принципа «ближнего соседа»**

Если в 1 кластер вошли объекты (A, B, C), а во 2 кластер вошли объекты (D, E), то используя М. «д.с.» для определения расстояния между двумя кластерами получим, что расстояние между кластером {A, B, C} и кластером {D, E} равно длине отрезка CD.

Академиком А.Н. Колмогоровым предложено «обобщённое расстояние» между кластерами.

Расстояние между классом s_l и классом $s_{(m,\ q)}$, являющимся объединением двух других классов s_m и s_q, можно определить по формуле:

$$d_{l,(m,q)} = d(s_l, s_{(m,q)}) = \alpha d_{lm} + \beta d_{lq} + \gamma d_{mq} + \delta \left| d_{lm} - d_{lq} \right|,$$

где d_{lm}, d_{lq}, d_{mq} – расстояния между соответствующими кластерами;

Для М.«б.с.» параметры, учитывающие особенности алгоритма кластеризации данных равны $\alpha = 1/2, \beta = 1/2, \gamma = 0, \delta = -1/2$.

МЕТОД «ДАЛЬНЕГО СОСЕДА» (ПОЛНЫХ СВЯЗЕЙ)

метод *кластерного анализа*, при котором расстояние между кластерами рассчитывается как макс. расстояние между любыми двумя объектами в различных кластерах (т.е. «наиболее удалёнными соседями»). Расстояние, измеряемое по данному методу находят, используя формулу:

$$d_{\max}(s_l, s_m) = \max_{x_i \in s_l, x_j \in s_m} d(x_i, x_j).$$

Графически это можно проиллюстрировать (см. рис. 1):

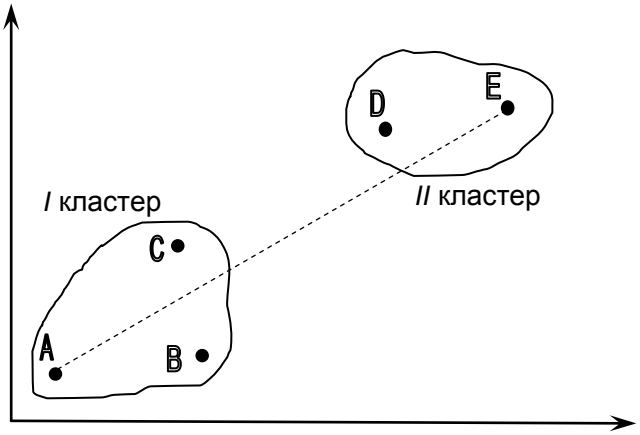


Рис. 1. **Графическая иллюстрация метода «дальнего соседа».**

Если в 1 кластер вошли объекты (A, B, C), а во 2 кластер вошли объекты (D, E), то используя М.«д.с.» для определения расстояния между двумя кластерами получим, что расстояние между кластером {A, B, C} и кластером {D, E} равно длине отрезка AE.

А.Н. Колмогоровым предложено «обобщённое расстояние» между кластерами, согласно которому, расстояние между классом s_l и классом $s_{(m,\ q)}$, являющимся объединением двух других классов s_m и s_q, можно определить по формуле:

$$d_{l,(m,q)} = d(s_l, s_{(m,q)}) = \alpha d_{lm} + \beta d_{lq} + \gamma d_{mq} + \delta \left| d_{lm} - d_{lq} \right|$$

где d_{lm}, d_{lq}, d_{mq} – расстояния между соответствующими кластерами;

Для М.д.с. $\alpha = 1/2, \beta = 1/2, \gamma = 0, \delta = 1/2$ – параметры, учитывающие особенности алгоритма кластеризации данных.

МЕТОД К-СРЕДНИХ (МЕТОД МАК-КУИНА)

итеративный метод *кластерного анализа*. В отличие от иерархических процедур метод k-средних не требует вычисления и хранения матрицы расстояний или сходств между объектами. Действие алгоритма таково, что он стремится минимизировать дисперсию на точках каждого кластера:

$$V = \sum_{i=1}^{k} \sum_{x_i \in Si} (x_i - \mu_i)^2 ,$$

где k – число кластеров, S_i – полученные кластеры, $i=1,2,...,k$ и μ_i – центры тяжести векторов $x_i \in S_i$.

Пусть имеется n наблюдений, каждое из которых характеризуется p признаками $X_1, X_2, ..., X_p$. Эти наблюдения необходимо разбить на k кластеров. Для начала из n точек исследуемой совокупности отбираются случайным образом или задаются исследователем исходя из каких-либо

априорных соображений k – точек (объектов). Эти точки принимаются за эталоны. Каждому эталону присваивается порядковый номер, который одновременно является и номером кластера. На первом шаге из оставшихся (n-k) объектов извлекается точка X_i с координатами (x_{i1},x_{i2}, ..., x_{ip}) и проверяется, к какому из эталонов (центров) она находится ближе всего. Для этого используется одна из метрик, напр., *евклидово расстояние*.

Проверяемый объект присоединяется к тому центру (эталону), которому соответствует min d_{il} (l = 1,..., k). Эталон заменяется новым, пересчитанным с учётом присоединенной точки, и вес его (количество объектов, входящих в данный кластер) увеличивается на единицу. Если встречаются два или более миним. расстояния, то i-й объект присоединяют к центру с наименьшим порядковым номером. На следующем шаге выбираем точку X_{i+1} и для неё повторяются все процедуры. Т.о., через (n-k) шагов все точки (объекты) совокупности окажутся отнесенными к одному из k кластеров, но на этом процесс разбиения не заканчивается. Для того чтобы добиться устойчивости разбиения по тому же правилу, все точки X_1, X_2, ..., X_n опять подсоединяются к полученным кластерам, при этом веса продолжают накапливаться. Новое разбиение сравнивается с предыдущим. Если они совпадают, то работа алгоритма завершается. В противном случае цикл повторяется. Окончательное разбиение имеет центры тяжести, которые не совпадают с эталонами, их можно обозначить C_1, C_2,..., $C_к$. При этом каждая точка X_i = (1, 2, ..., *n*) будет относиться к такому кластеру (классу) *l*, для которого

$$d(x_j, c_l) = \min_{1 \le j \le R} d(x_j, C_j).$$

Возможны две модификации M.k.-c. Первая предполагает пересчёт центра тяжести кластера после каждого изменения его состава, а вторая – лишь после того, как будет завершён просмотр всех данных. В обоих случаях итеративный алгоритм этого метода минимизирует дисперсию внутри каждого кластера, хотя в явном виде такой критерий оптимизации не используется. M.k-c. допускает в качестве исходного разбиения использовать группировку, полученную одним из методов иерархического кластерного анализа. Такой подход можно рекомендовать для сокращения времени обработки в том случае, когда совокупность объектов достаточно велика и пользователь затрудняется указать количество образуемых кластеров.

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

один из осн. способов уменьшить размерность данных, потеряв наименьшее количество информации; применён К. Пирсоном в 1901; используется во многих областях, таких как распознавание образов, сжатие данных и т.п. М.г.к. – один из методов *факторного анализа,* который был предложен Харманом в 1972; Вычисление *гл. компонент* сводится к вычислению собственных векторов и собственных значений *ковариационной (корреляционной) матрицы* исходных данных.

М.г.к. позволяет решить четыре осн. задачи: отыскать скрытые, но объективно существующие закономерности, которые определяются воздействием внутренних и внешних причин на изучаемый процесс; сжать информацию путём описания процесса при помощи общих факторов или главных компонент, число которых значительно меньше количества первоначально взятых признаков; выявить и изучить статистические связи признаков с главными компонентами; спрогнозировать ход развития процесса на основе уравнения регрессии на главных компонентах. Модель М.г.к. имеет вид:

$$z_{ij} = \sum_{\nu=1}^{k} a_{j\nu} f_{i\nu},$$

где z_{ij} – нормированное значение j-го признака у i-го объекта исходных данных матрицы X = (x_{ij}); $a_{j\nu}$ – вес ν-й гл. компоненты на j-ой переменной; $f_{i\nu}$ – значение ν-й главной компоненты для i-го объекта; j, ν = 1, 2, …, k; i = 1, 2, …, n.

$$\sum_{\nu=1}^{k} a_{j\nu}^2 = 1.$$

Вектор A=($a_{1\nu}$, $a_{2\nu}$, ..., $a_{k\nu}$) лежит на единичной сфере в n-мерном пространстве. В М.г.к. сначала находят направление макс. разброса, т.е. та-

кой A, при котором достигает макс. дисперсия случайной величины z_{ij}. Тогда вектор A задает первую гл. компоненту, а величина z_{ij} является проекцией вектора данных на ось первой гл. компоненты. Затем, рассматривают гиперплоскость в n-мерном пространстве, перпендикулярную первой гл. компоненте, и на эту гиперплоскость проецируются все элементы выборки. Размерность гиперплоскости на единицу меньше, чем размерность исходного пространства. В рассматриваемой гиперплоскости процедура повторяется. В ней находят направление наибольшего разброса, т.е. вторую гл. компоненту. Затем выделяется гиперплоскость, перпендикулярная первым двум гл. компонентам. Её размерность на 2 меньше, чем размерность исходного пространства. Далее – следующая итерация. С точки зрения линейной алгебры, речь идёт о построении нового базиса в n-мерном пространстве, ортами которого служат гл. компоненты. Дисперсия, соответствующая каждой новой гл. компоненте, меньше, чем для предыдущей. Обычно останавливаются, когда она меньше заданного порога. Т.о., первая гл. компонента вносит наибольший вклад в суммарную дисперсию, а последняя k-я – наименьший.

Для визуального анализа данных часто используют проекции исходных векторов на плоскость первых двух гл. компонент. Множество гл. компонент – удобная система координат, а их вклад в общую дисперсию характеризует статистические свойства гл. компонент. Из общего числа гл. компонент для исследования, как правило, оставляют m (m<k) наиболее весомых, т.е. вносящих макс. вклад в объясняемую часть общей дисперсии. Опыт исследований показывает, что m (0,1 ÷ 0,3)k. Для смысловой интерпретации полученных результатов самыми наглядными являются случаи, когда m = 1, 2, 3, 4. Т.е., несмотря на то, что в М.г.к. для точного воспроизведения корреляции и дисперсий между переменными надо найти все компоненты, большая доля дисперсии объясняется небольшим числом гл. компонент. Кроме того, можно по признакам описать факторы, а по факторам (гл. компонентам) описать признаки.

Одно из основополагающих условий М.г.к. связано с представлением корреляционной матрицы R через матрицу факторных нагрузок A:

$$R=\frac{1}{n}Z^TZ=\frac{1}{n}\left(FA^T\right)^TFA^T=A\left(\frac{1}{n}F^TF\right)A^T,\ R=AA^T.$$

Общий вклад всех гл. компонент в суммарную дисперсию равен k. Тогда удельный вклад ν-й гл. компоненты определяется по формуле:

$$\lambda_\nu/k\cdot100\%,$$

где λ_ν – собственное значение гл. компоненты ν.

Суммарный вклад m первых гл. компонент определяется из выражения:

$$\sum_{\nu=1}^{m}\lambda_\nu\Big/k\cdot100\%.$$

Обычно для анализа используют m первых гл. компонент, суммарный вклад которых превышает 60-70%.

Матрица факторных нагрузок A используется для экономической интерпретации гл. компонент, которые представляют собой линейные функции исходных признаков. Для экономической интерпретации используются лишь те f_ν, для которых

$$\left|a_{j\nu}\right|>0{,}5$$

(в зависимости от поставленной задачи неравенство можно «ужесточать», напр.,

$$\left|a_{j\nu}\right|>0{,}6$$

и т.д.).

Результаты применения М.г.к. представляются либо данными матрицы отображения A, либо данными матрицы значений гл. компонент для всех объектов наблюдения F. Матрица F используется для построения регрессионного уравнения на гл. компонентах, классификации наблюдений в пространстве гл. компонент.

МЕТОД КОРРЕЛЯЦИОННЫХ ПЛЕЯД

эвристический метод систематизации признаков; предназначен для нахождения таких групп признаков – «плеяд», когда корреляционная

связь, т.е. сумма модулей *коэффициентов корреляции* между параметрами одной группы (внутриплеядная связь) достаточно велика, а связь между параметрами из разных групп (межплеядная) – мала. По определённому правилу по *корреляционной матрице* признаков образуют чертёж – граф, который затем с помощью различных приёмов разбивают на подграфы. Элементы, соответствующие каждому из подграфов, и образуют плеяду. М.к.п. разработан П.В. Терентьевым в 1959.

Рассмотрим корреляционную матрицу R = (r_{ij}), $i,j = 1,2, ..., p$, исходных признаков. Нарисуем p кружков; внутри каждого кружка напишем номер одного из признаков. Каждый кружок соединяется линиями со всеми остальными кружками; над линией, соединяющей *i*-й и *j*-й элементы (ребром графа), ставится значение модуля коэффициента корреляции $|r_{ij}|$. Полученный таким образом чертеж рассматриваем как исходный граф. Задавшись (произвольным образом или на основании предварительного изучения корреляционной матрицы) некоторыми пороговыми значениями коэффициента корреляции r_0, исключаем из графа все ребра, которые соответствуют коэффициентам корреляции, по модулю меньшим r_0. Затем задаем некоторое $r_1 > r_0$ и относительно него повторяем описанную процедуру. При некотором достаточно большом г граф распадается на несколько подграфов, т.е. таких групп кружков, что связи (ребра графа) между кружками различных групп отсутствуют. Очевидно, что для полученных т.о. плеяд внутриплеядные коэффициенты корреляции будут больше *r*, а межплеядные – меньше *r*.

В другом варианте корреляционных плеяд предлагается упорядочивать признаки и рассматривать только те коэффициенты корреляции, которые соответствуют связям между элементами в упорядоченной системе. Упорядочение производится на основании принципа макс. корреляционного пути: все p признаков связываются при помощи (p – 1) линий (ребер) так, чтобы сумма модулей коэффициентов корреляции была макс.

Различные варианты М.к.п., исторически раньше возникшие, – в действительности несколько упрощенные эвристические версии более совершенных в математическом плане алгоритмов исследования структуры связей между компонентами многомерного признака, использующими графы-деревья и стохастические сети.

МЕТОД ЦЕНТРОИДНЫЙ

см. в ст. *Центроидный метод*

МНОГОМЕРНОЕ НАБЛЮДЕНИЕ

реализация *случайной величины многомерной.* Это n-мерный вектор $x=(x_1,x_2,...,x_n)$, составленный из первичных признаков одномерных x_1, x_2, ..., x_n. Набор значений этих признаков называется значением М.н. Совокупность значений М.н. обычно рассматривается как массив многомерных данных, матрица данных. Такие данные нередко являются исходными для многомерного статистического анализа. При этом М.н. чаще всего интерпретируется как многомерная случайная величина. Часто используется геометрическая интерпретация объектов как точек признакового пространства, координатные оси которого соответствуют рассматриваемым одномерным признакам. А координаты каждого изучаемого объекта это отвечающие ему значения этих признаков.

МНОГОМЕ́РНОЕ ШКАЛИРОВАНИЕ

совокупность математических методов, позволяющих по заданной информации о мерах различия (близости) между объектами рассматриваемой совокупности приписывать каждому из этих объектов вектор характеризующих его количественных показателей; при этом размерность искомого координатного пространства задаётся заранее, а «погружение» в него анализируемых объектов производится т.о., чтобы структура взаимных различий (близостей) между ними, измеренных с помощью приписываемых им вспомогательных координат, в

среднем наименее отличалась бы от заданной в смысле того или иного функционала качества.

Процедуры М.ш. применяются тогда, когда данные заданы в виде матрицы попарных расстояний между объектами или их порядковых отношений. В первом случае используются методы т.н. *многомерного шкалирования метрического*, а во втором – *шкалирования неметрического*.

МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ МЕТРИЧЕСКОЕ

вид *многомерного шкалирования*; цель М.ш.м. – отображение информации о конфигурации исходных многомерных данных, заданную матрицей расстояний, в виде геометрической конфигурации *n* точек в соответствующем многомерном пространстве. Предполагается, что исходная информация об объектах задана в форме матрицы попарных сравнений, элементы которой интерпретируются как *евклидово расстояние* между объектом O_i и объектом O_j (i,j = 1,2,…,n). При этом элементы матрицы расстояний получены по интервальным шкалам.

МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ НЕМЕТРИЧЕСКОЕ

в М.ш.н. предполагается, что различия (близости) γ_{ij} измерены в ординальной шкале, так что важен только ранговый порядок различий, а не сами их численные значения. Процедуры М.ш.н. стремятся построить такую геометрическую конфигурацию точек в пространстве заданной размерности, чтобы ранговый порядок попарных расстояний между ними, по возможности, минимально отличался от того порядка, который задан матрицей расстояний γ. Одна процедура М.ш.н. отличается от другой выбором вида критерия различия двух разных упорядочений.

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

наиболее полное описание случайного вектора (многомерной случайной величины) $X = (X_1, X_2,\ldots, X_n)$. При конечном множестве значений дискретного случайного вектора такой закон может быть задан в форме табл. (матрицы), содержащей всевозможные сочетания значений каждой из одномерных *случайных величин* (составляющих вектора) и соответствующие их вероятности. Так, если рассматривается двумерный случайный вектор (*X, Y*) с дискретными составляющими, то его распределение можно представить в виде матрицы $P = (p_{ij})$, I = 1, 2, …, m; j = 1,…, n, для которой

$$\sum_{i=1}^{m}\sum_{j-1}^{n} p_{ij} = 1.$$

Столбец и строка, получаемые суммированием всех столбцов и строк матрицы, представляют распределения одномерных составляющих (x_i, p_i) и (y_j, p_j).

Функция распределения $F(x_1, x_2,\ldots, x_n)$ случайного вектора $X = (X_1, X_2,\ldots, X_n)$, выражающая вероятность совместного выполнения n неравенств $X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n$, т.е. $F = (x, x_2,\ldots, x_n) = P(X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n)$, представляет универсальное описание многомерного случайного вектора, как с дискретными, так и непрерывными составляющими. В двумерном случае для случайного вектора (X, Y) функция распределения $F(x, y) = P(X < x, Y < y)$. Геометрически функция распределения F(x, y) означает вероятность попадания случайного вектора (X, Y) в бесконечный квадрант, лежащий левее и ниже точки M(x, y) координатной плоскости. Свойства функции распределения $F(x_1, x_2,\ldots, x_n)$: 1) $0 \le F(x_1, x_2,\ldots, x_n) \le 1$; 2) $F(x, x_2,\ldots, x_n)$ – неубывающая функция каждого их своих аргументов; 3) Если хотя бы один из аргументов $x_1, x_2,\ldots, x_n$ обращается в $-\infty$, функция распределения равна нулю; 4) Функция распределения одной из случайных величин (составляющих) X_k (k = 1,2,…, n) получается из функции распределения $F(x_1, x_2,\ldots, x_n)$, если все аргументы, кроме x_k, считать равными $+\infty$. В частности для двумерного случайного вектора: $0 \le F(x, y) \le 1$; при $x_2 > x_1$ $F(x_2, y) \ge F(x_1, y)$, при $y_2 > y_1$ $F(x; y_2) \ge F(x; y_1)$; $F(x, -\infty) = F(-\infty, y) = F(-\infty, -\infty) = 0$; 4) $F(x, +\infty) = F_1(x)$, $F(+\infty, y) = F_2(y)$, где $F_1(x)$ и $F_2(y)$ − функции распределения случайных величин X и Y.

Если случайные составляюшие $X_1, X_2,\ldots, X_n$ вектора X независимы (т.е. закон распределения каждой из них не зависит от того, какие значения приняли другие), то функция распределения вектора $X = (X_1, X_2,\ldots, X_n)$ равна произведению функций распределения его составляющих: $F(x_1, x_2,\ldots, x_n) = F_1(x_1)\, F_2(x_2) \ldots F_n(x_n)$.

В отличие от одномерного случая, многомерная функция распределения $F(x, x_2,\ldots, x_n)$, перестает быть исчерпывающей (по информативности) формой задания изучаемого закона распределения. В этом смысле предпочтительнее *плотность распределения*.

Плотностью распределения (плотностью вероятности или совместной плотностью) $f(x_1, x_2,\ldots, x_n)$ случайного вектора (непрерывной многомерной случайной величины) $X = (X_1, X_2,\ldots, X_n)$ наз. смешанная частная производная его функции распределения, взятая один раз по каждому аргументу:

$$f(x_1, x_2,\ldots, x_n) = \frac{\partial^n F(x_1, x_2,\ldots, x_n)}{\partial x_1 \partial x_2 \ldots \partial x_n}.$$

Для двумерного случайного вектора (X, Y)

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

Элемент вероятности $f(x_1, x_2,\ldots, x_n)\, dx_1 dx_{2\ldots}\, dx_n$ показывает приближённо вероятность попадания случайного вектора $X = (X_1, X_2,\ldots, X_n)$ в элементарную область n-мерного пространства с размерами $dx_1 dx_2 \ldots dx_n$, примыкающую к точке $(x_1, x_2,\ldots, x_n)$: $f(x_1, x_2,\ldots, x_n)\, dx_1 dx_{2\ldots}\, dx_n \approx P[X_1 \in (x_1, x_1 + dx_1), X_2 \in (x_2, x_2 + dx_2), \ldots, X_n \in (x_n, x_n + dx_n)]$.

Вероятность попадания случайного вектора $X = (X_1, X_2,\ldots, X_n)$ в произвольную область D n-мерного пространства выражается n-кратным интегралом по области D:

$$P[(X_1, X_2,\ldots, X_n) \in D] = \int\!\!\int_D \ldots \int f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots . dx_n.$$

Плотность распределения составляющей X_k находится аналогично (n−1) кратным интегрированием плотности вероятности $f(x_1, x_2,\ldots, x_n)$ по всем аргументам, кроме x_k.

Отметим свойства плотности вероятности f (x, y) двумерного случайного вектора (X, Y):

1) $f(x, y) \geq 0$;

2) 2) $P[(X,Y) \in D] = \iint_D f(x,y)\, dxdy$;

3) $F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x, y)\, dxdy$;

4) $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)\, dxdy = 1$.

Частной маржинальной плотностью $f_{1,2,\ldots,k}(x_1, x_2,\ldots, x_k)$ распределения любого подвектора $(X_1, X_{2,\ldots,} X_k)$ случайного вектора $(X_1, X_{2,\ldots,} X_{k,\ldots,} X_n)$, $k<n$, называется плотность распределения этого подвектора, задаваемая соотношением:

$$f_{1,2,\ldots,k}(x_1, x_2,\ldots, x_k) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty} f(x_1, x_2, \ldots, x_n)\, dx_{k+1} dx_{k+2}, \ldots, dx_n$$

(интегрирование проводится (n−k) раз по аргументам $(x_{k+1,}\ x_{k+2,\ \ldots,}\ x_n)$, относящимся к остальным случайным составляющим).

В частном случае плотности распределения f(x), f(y) одномерных составляющих X и Y находятся по формулам:

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, y)\, dy;\ f_2(y) = \int_{-\infty}^{+\infty} f(x, y)\, dx.$$

Условной плотностью распределения $f_{1,2,\ldots,k}(x_1, x_2,\ldots, x_{k\,|}\ x_{k+1,}\ x_{k+2,\ \ldots,}\ x_n)$ любого подвектора $(X_1, X_{2,\ \ldots,} X_k)$ случайного вектора $(X_1, X_{2,\ \ldots,} X_{k,\ldots,} X_n)$, $k<n$, называется плотность распределения этого подвектора, вычисленная в предположении, что составляющие другого подвектора $(X_{k+1,}\ X_{k+2,\ \ldots,}\ X_n)$ приняли определенные значения $(X_{k+1} = x_{k+1,}\ X_{k+2} = x_{k+2,\ \ldots,} X_n = x_n)$:

$$f_{1,2,\ldots,k}(x_1, x_2,\ldots, x_{k\,|}\ x_{k+1,}\ x_{k+2,\ \ldots,}\ x_n) = \frac{f(x_1, x_2, \ldots, x_n)}{f_{k+1,\ldots,n}(x_{k+1}, \ldots, x_n)}.$$

В случае двумерного случайного вектора (X,Y) условные плотности вероятности f(x|y) и f(y|x) определяются по формулам.

$$f(x|y) = \frac{f(x, y)}{f_2(y)} = \frac{f(x, y)}{\int_{=\infty}^{+\infty} f(x, y)\, dx};$$

$$f(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dy}.$$

Плотности частная (маржинальная) $f_{1,2,\ldots,k}$ (x_1, x_2,…, x_k) и условная $f_{1,2,\ldots,k}$ (x_1, x_2,…, x_k | x_{k+1}, x_{k+2}, …, x_n) описывают распределение одного и того же подвектора (X_1, X_2, …, X_k); при этом вторая плотность, в отличие от первой, зависит от того, какие значения принимают остальные составляющие (X_{k+1},X_{k+2}, …,X_n). Законы распределения многомерных случайных величин являются функциями многих переменных, чаще всего неудобны в практическом применении и к тому же для своего определения требуют значительного объема статистических данных. В большинстве случаев в статистических приложениях вместо законов распределения случайного вектора рассматриваются его важнейшие числовые характеристики: вектор *математических ожиданий* MX = (MX_1, MX_2,…,MX_n), *матрицу ковариационную* $\sum = (\sigma_{ij})$, i,j = 1,2,…,n с элементами σ_{ij} = Cov (X_i , X_j) = M[(X_i – MX_i) (X_j –MX_j)], на главной диагонали которой находятся *дисперсии* переменных σ_{ii} = D (X_i), i = 1, 2, …, n.

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ ГАММА НОРМАЛЬНЫЙ

закон распределения параметров – среднего M и меры точности *R* повторной выборки из нормально распределённой совокупности.

Пусть X_1,X_2 „…., X_n – повторная выборка из нормально распределённой совокупности с неизвестным значением среднего *M* и неизвестным значением меры точности *R*, априорное совместное распределение *M и R* таково, что *условное распределение* M при *R*=*r* (*r*>0)−нормальное со средним μ и мерой точности τ*r* (*τ*>0), а *распределение маргинальное* R есть *гамма распределение* с параметрами α и β (α>0 и β>0). Тогда апостериорное совместное распределение *M* и *R* при $X_i = x_i$ (*i*=1,2,…, *n*) имеет вид: условное распределение *M* при *R* = *r* – нормальное со средним μ′ и мерой точности (τ+*n*)*r*, где μ′ = =(τμ+*n* $\bar{x}$)/(τ+*n*), $\bar{x}$ – среднее значение выборочное, а маргинальное распределение *R* есть гамма распределение с параметрами α+*n*/2 и β′, где

$$\beta' = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\tau n(\bar{x} - \mu)^2}{2(\tau + n)}.$$

В этом случае *плотность распределения* параметров *M* и *R*

$$f(m, r) \sim \sqrt{r}\; e^{-\frac{\tau r}{2}(m+\mu)^2} r^{\alpha-1} e^{-\beta r}$$

(символ ~ означает знак пропорциональности) – плотность гамма-нормального закона распределения. При этом условное распределение M для любого заданного значения *R* = *r* будет нормальным, но маргинальное распределение M таковым не будет.

Планирование и анализ эксперимента упрощаются, если повторная выборка извлекается из распределения, принадлежащего к сопряжённому семейству распределений, и если априорное распределение некоторого параметра W принадлежит этому семейству, то при любом объёме выборки n и любых значениях наблюдений в выборке апостериорное распределение W также принадлежит этому семейству. Особенность совместного распределения M и R состоит в том, что при любом гамма-нормальном распределении из этого семейства величины M и R зависимы. Другой особенностью является то, что для каждого распределения из сопряженного семейства мера точности условного распределения M при *R* = *r* пропорциональна r.

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ ПОЛИНОМИАЛЬНЫЙ

мультиномиальный, закон распределения *k*-мерного случайного вектора (k-мерной случайной величины) X= (X_1, X_2,…, X_k) с дискретными составляющими X_i (i = 1, 2,…, k), принимающими целые неотрицательные значения m_1, m_2,…, m_k с вероятностями:

$$P(X_1=m_1, X_2=m_2,\ldots, X_k=m_k) = \frac{n!}{m_1!m_2!,,,m_k!} p_1^{m_1} p_2^{m_2},,, p_k^{m_k},$$

где

$$\sum_{i=1}^{k} m_i = n,\ 0<p_i<1,\ \sum_{i=1}^{k} p_i = 1.$$

Эти вероятности могут быть получены при разложении полинома $(p_1 + p_2 + \ldots + p_k)^n$; отсюда и название распределения – «полиномиальное». Если в каждом испытании может произойти одно из k исключающих друг друга событий $A_1, A_2, \ldots, A_k$ соответственно с вероятностями $p_1, p_2, \ldots, p_k$, то $P(X_1=m_1, X_2=m_2, \ldots, X_k=m_k)$ выражает вероятность того, что в n независимых испытаниях событие A_1 произойдет m_1 раз, A_2 – m_2, и т.д., событие A_k – m_k раз ($m_1+m_2+\ldots+m_k = n$). Полиномиальный закон обобщает закон *распределения биномиального* и совпадает с последним при k=2.

Случайный вектор $(X_1, X_2, \ldots, X_k)$ имеет вектор математическое ожидание $(np_1, np_2, \ldots, np_k)$ и ковариационную матрицу $\sum = (\sigma_{ij})$, где $\sigma_{ij} = np_i(1-np_i)$, если i= j и $\sigma_{ij} = -np_ip_j$, если $i \neq j$ (i,j=1,2,…,k). При $n \to \infty$ распределение вектора $(X_1, X_2, \ldots, X_n)$ с нормированными компонентами $x_i = (X_i - np_i)/\sqrt{np_i(1-np_i)}$ стремится к некоторому *многомерному закону распределении вероятностей* нормальному, а распределение суммы

$$\sum_{i=1}^{k} (1-p_i)x_i^2$$

стремится к χ^2 – *распределению* с $k-1$ степенями свободы.

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ СТЬЮДЕНТА

закон распределения n-мерного случайного вектора (n-мерной *случайной величины*) $X = (X_1, X_2, \ldots, X_n)'$ с непрерывными составляющими. Для случайного вектора $X = (X_1, X_2, \ldots, X_n)'$ плотность n-мерного распределения Стьюдента (t-распределения) с k степенями свободы, вектором сдвига M и матрицей точности T имеет вид:

$$f(X)= f(x_1,x_2,\ldots,x_n)=\frac{\Gamma\left(\frac{n+k}{2}\right)\sqrt{|T|}}{\Gamma\left(\frac{k}{2}\right)\sqrt{(2\pi)^n}}\ [1+\frac{1}{k}(X-M)'T(X-M)]^{-\frac{n+k}{2}},$$

где T − симметрическая положительно определенная матрица, $\Gamma(x)$ – гамма функция.

Если случайный вектор $Y = (Y_1, Y_2, \ldots, Y_n)'$ имеет *многомерный закон распределения вероятностей* нормальный с нулевым вектором *математических ожиданий* $M = M(Y_i) = 0$ и невырожденной *матрицей ковариационной* $\sum = T^{-1}$ ($\sum = (\sigma_{ij})$, $i,j = 1,2,\ldots, n$), а Z имеет χ^2 - *распределение* с k степенями свободы, то вектор $X = (X_1, X_2, \ldots, X_n)'$, где

$$X_i = \frac{\sqrt{k}}{\sqrt{Z}} Y_i + M_i,$$

имеет n-мерное распределение Стьюдента с k степенями свободы, вектором сдвига M и матрицей точности T. Если случайный вектор X = $(X_1, X_2, \ldots, X_n)'$ имеет n-мерное распределение Стьюдента с k степенями свободы, вектором сдвига M и матрицей точности T, то случайная величина

$$Z = \frac{1}{n}(X-M)'T(X-M)$$

имеет F-распределение с n и k степенями свободы.

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ СТЬЮДЕНТА (ОБОБЩЁННЫЙ)

закон распределения pq элементов случайной матрицы $T = (t_{ij})$

размера $p \times q$, если они имеют плотность

$$P(T) = k\frac{|Q|^{(n-p)/2}|P|^{q/2}}{|Q+T'PT|^{n/2}},$$

где

$$k^{-1} = \pi^{pq/2}\prod_{i=1}^{q}\Gamma\{(n-p-i+1)/2\}/\prod_{i=1}^{q}\Gamma(n-i+1)/2,\ n > p+q-1,\ P \text{ и } Q -$$

положительно определенные симметрические матрицы размеров $p \times p$ и $q \times q$ соответственно; $\Gamma(x)$ – гамма функция. Указанную плотность распределения в литературе иногда наз. матричной плотностью распределения вероятностей *t*-распределения Стьюдента (обобщённого).

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ УИШАРТА

закон распределения случайной *матрицы*, многомерный аналог *χ^2-распределения*. Пусть X – симметрическая матрица k-го порядка, либо, что эквивалентно, $\frac{n(n+1)}{2}$ – мерный вектор, а Y – симметрическая положительно определённая матрица n-го порядка. Случайная матрица X имеет невырожденное распределение Уишарта с k-степенями свободы и матрицей точности T ($|T| \neq 0$), если её плотность распределения имеет вид:

$$F(X)= \frac{|T|^{\frac{k}{2}} |Y|^{\frac{k-n-1}{2}}}{2^{\frac{nn}{2}} \pi^{\frac{n(n-1)}{4}} \prod_{j=1}^{n} \Gamma\left(\frac{k+1-j}{2}\right)} e^{-\frac{1}{2}tr(TY)},$$

где *tr (TY)* – след матрицы *TY*, т.е. сумма её диагональных элементов.

Если $X^{(1)}$, $X^{(2)}$,…, $X^{(n)}$ – независимые нормально распределённые векторы с нулевыми математическими ожиданиями и невырожденной *матрицей ковариационной* $\sum$, то случайная матрица

$$X= \sum_{p=1}^{n} X^{(p)} (X^{(p)})'^{h}$$

имеет распределение Уишарта с n – степенями свободы и матрицей точности $T=(\sum)^{-1}$.

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

раздел *математической статистики*, посвящённый математическим методам построения оптимальных планов сбора, систематизации, обработки и интерпретации многомерных статистических данных, нацеленным, в первую очередь, на выявление характера и структуры взаимосвязей между компонентами исследуемого многомерного признака и предназначенным для получения научных и практических выводов. Под многомерным признаком понимается p-мерный вектор $x=(x_1, x_2, \ldots, x_p)'$ показателей (признаков, переменных) $x_1, x_2, \ldots, x_p$, среди которых: количественные, т.е. скалярно измеряющие в определённой шкале степень проявления изучаемого свойства объекта; порядковые (или ординальные), т.е. позволяющие упорядочивать анализируемые объекты по степени проявления в них изучаемого свойства; классификационные (или номинальные), т.е. позволяющие разбивать исследуемую совокупность объектов на не поддающиеся упорядочиванию однородные (по анализируемому свойству) классы. Результаты измерения этих показателей:

$$\{x_i\}_1^n = \left\{(x_{1i}, x_{2i}, \ldots, x_{pi})'\right\}_1^n, \qquad (1)$$

на каждом из *n* объектов исследуемой совокупности образуют последовательность многомерных наблюдений, или исходный массив многомерных данных для проведения М.с.а. Значительная часть М.с.а. обслуживает ситуации, в которых исследуемый многомерный признак интерпретируется как многомерная случайная величина и соответственно последовательность многомерных наблюдений (1) – как выборка из генеральной совокупности. В этом случае выбор методов обработки исходных статистических данных и анализ их свойств производится на основе тех или иных допущений относительно природы многомерного (совместного) *закона распределения вероятностей* (ЗРВ).

По содержанию М.с.а., может быть условно разбит на три осн. подраздела; М.с.а. многомерных распределений и их основных характеристик; М.с.а. характера и структуры взаимосвязей между компонентами исследуемого многомерного признака; М.с.а. геометрической структуры исследуемой совокупности многомерных наблюдений.

М.с.а. многомерных распределений и их осн. характеристик охватывает лишь ситуации, в которых обрабатываемые наблюдения (1) имеют вероятностную природу, т.е. интерпретиру-

ются как выборка из соответствующей ген. совокупности. К осн. задачам этого подраздела относятся: статистическое оценивание исследуемых многомерных распределений, их главных числовых характеристик и параметров; исследование свойств используемых статистических оценок; исследование распределений вероятностей для ряда статистик, с помощью которых строятся статистические критерии проверки различных гипотез о вероятностной природе анализируемых многомерных данных. Основные результаты относятся к частному случаю, когда исследуемый признак *x* подчинён многомерному нормальному закону распределения $N_p(\mu,V)$, функция плотности которого $f(x\,|\,\mu,V)$ задаётся соотношением:

$$f(x\,|\,\mu,V)=\frac{1}{(2\pi)^{\frac{p}{2}}\,|V|^{\frac{1}{2}}}\times\exp\left\{-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)\right\}, \qquad (2)$$

где $\mu=(\mu_1,\ldots,\mu_p)'$ – вектор математических ожиданий компонент случайной величины *x,* т.е. $\mu_l=Ex_l,\ l=1,2,\ldots,p,$ а $V=\|v_{ij}\|i,\ j=1,$ ковариационная матрица случайного вектора *x,* т.е. $v_{ij}=E(x_i-\mu_i)(x_j-\mu_j)$ – ковариации компонент вектора x (рассматривается невырожденный случай, когда ранг $V=p$; в противном случае, т.е. при ранге $V=p'<p$, все результаты остаются справедливыми, но применительно к подпространству меньшей размерности p', в которой оказывается сосредоточенным распределение вероятностей исследуемого случайного вектора *x*).

Так, если (1) – последовательность независимых наблюдений, образующих случайную выборку из $N_p(\mu,V)$, то оценками макс. правдоподобия для параметров μ и V, участвующих в (2), являются соответственно статистики:

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

и

$$\hat{V}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})',$$

причём случайный вектор $\hat{\mu}$ подчиняется p-мерному нормальному закону

$$N_p(\mu,\frac{1}{n}V)$$

и не зависит от $\hat{V}$, а совместное распределение элементов матрицы $\hat{Q}=n\hat{V}$ описывается т.н. распределением Уишарта, плотность которого:

$$w(\hat{Q}\,|\,V;n)=\begin{cases}\dfrac{|\hat{Q}|^{(n-p-2)/2}\exp\left\{-\frac{1}{2}\operatorname{tr}(V^{-1}\hat{Q})\right\}}{2^{(n-1)p/2}\pi^{p(p-1)/4}\,|V|^{(n-1)/2}\prod_{j=1}^{p}\Gamma\left\{\frac{n-j}{2}\right\}},\\ \text{если } \hat{Q}-\text{положительно определена;}\\ 0-\text{в противном случае.}\end{cases}$$

В рамках этой же схемы исследованы распределения и моменты таких выборочных характеристик многомерной случайной величины, как коэффициенты *парной, частной* и *множественной корреляции, обобщённая дисперсия* (т.е. статистика $|\hat{V}|$), обобщённая T^2 – статистика Хотеллинга. В частности, если определить в качестве выборочной ковариационной матрицы S_n подправленную «на несмещённость» оценку $\hat{V}$, а именно:

$$S_n=\frac{n}{n-1}\hat{V},$$

то распределение случайной величины $\sqrt{n}(|S_n|/|V|-1)$ стремится к $N_1(0,2p)$ при $n\to\infty$, а случайные величины:

$$\frac{n-p}{p(n-1)}T^2=\frac{n-p}{p(n-1)}n(\hat{\mu}-\mu)'S_n^{-1}(\hat{\mu}-\mu) \qquad (6)$$

и

$$\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}\tilde{T}^2=\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}\cdot\frac{n_1n_2}{n_1+n_2}(\hat{\mu}_{n_1}-\hat{\mu}_{n_2})'S_{n_1+n_2}^{-1}(\hat{\mu}_{n_1}-\hat{\mu}_{n_2}) \qquad (7)$$

подчиняются *F*-распределениям с числами степеней свободы соответственно $(p,n-p)$ и (p,n_1+n_2-p-1). В соотношении (7) n_1 и n_2 – объёмы двух независимых выборок вида (1), извлечённых из одной и той же ген. совокупности $N_p(\mu,V)$, μ_{n_i} и S_{n_i} – оценки вида (3) и (4) – (5), построенные по *i*-й выборке, а

$$S_{n_1+n_2}=\frac{1}{n_1+n_2-2}\left[(n_1-1)S_{n_1}+(n_2-1)S_{n_2}\right]$$

– общая выборочная ковариационная матрица, построенная по оценкам S_{n_1} и S_{n_2}.

М.с.а. характера и структуры взаимосвязей компонент исследуемого многомерного признака объединяет в себе понятия и результаты, обслуживающие такие методы и модели М.с.а. как множественная регрессия, многомерный *дисперсионный анализ* и *ковариационный анализ*, *факторный анализ* и *метод гл. компонент*, анализ канонических корреляций, анализ многомерных временных рядов. Результаты, составляющие содержание этого подраздела, условно разделяются на два осн. типа.

1. Построение наилучших (в определённом смысле) статистических оценок для параметров упомянутых моделей и анализ их свойств (точности, а в вероятностной постановке – законов их распределения, доверительных областей и т.п.). Так, пусть исследуемый многомерный признак *x* интерпретируется как векторная случайная величина, подчинённая *p*-мерному нормальному распределению $N_p(\mu,V)$, и расчленён на два подвектора-столбца $x^{(1)}$ и $x^{(2)}$ размерности *q* и *p*-*q* соответственно. Это определяет и соответствующее расчленение вектора математических ожиданий μ, теоретической и выборочной ковариационных матриц *V* и $\hat{V}$, а именно:

$$\mu=\begin{pmatrix}\mu^{(1)}\\ \mu^{(2)}\end{pmatrix},\ V=\begin{pmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix}\ и\ \hat{V}=\begin{pmatrix}\hat{V}_{11} & \hat{V}_{12}\\ \hat{V}_{21} & \hat{V}_{22}\end{pmatrix}.$$

Тогда *условное распределение* подвектора $x^{(1)}$ (при условии, что второй подвектор принял фиксированное значение $x^{(2)}$) будет также нормальным $N_q(\mu^{(1)}+B(x^{(2)}-\mu^{(2)}),\Sigma)$. При этом оценками макс. правдоподобия $\hat{B}$ и $\hat{\Sigma}$ для матриц регрессионных коэффициентов *B* и ковариаций Σ этой классической многомерной модели множественной регрессии:

$$E(x^{(1)}\mid x^{(2)})=\mu^{(1)}+B(x^{(2)}-\mu^{(2)}) \qquad (8)$$

будут взаимно независимые статистики соответственно:

$\hat{B}=\hat{V}_{12}\hat{V}_{22}^{-1}$ и $\hat{\Sigma}=\hat{V}_{11}-\hat{V}_{12}\hat{V}_{22}^{-1}\hat{V}_{21}$;

здесь распределение оценки $\hat{B}$ подчинено нормальному закону $N_{q(p-q)}(B,V_B)$, а оценки $n\hat{\Sigma}$ – закону Уишарта с параметрами Σ и $n-(p-q)$ (элементы ковариационной матрицы V_B выражаются в терминах элементов матрицы *V*).

Осн. результаты по построению оценок параметров и исследованию их свойств в моделях факторного анализа, гл. компонент и канонических корреляций относятся к анализу вероятностно-статистических свойств собственных (характеристических) значений и векторов различных выборочных ковариационных матриц.

В схемах, не укладывающихся в рамки классической нормальной модели, и тем более в рамки какой-либо вероятностной модели, осн. результаты относятся к построению алгоритмов (и исследованию их свойств) вычисления оценок параметров, наилучших с точки зрения некоторого экзогенно заданного функционала качества (или адекватности) модели.

2. Построение статистических критериев для проверки различных гипотез о структуре исследуемых взаимосвязей. В рамках многомерной нормальной модели (последовательности наблюдений вида (1) интерпретируются как случайные выборки из соответствующих; многомерных нормальных ген. совокупностей) по-

строены, напр., статистические критерии для проверки гипотез.

I. Гипотезы $\mu = \mu^*$ о равенстве вектора математических ожиданий исследуемых показателей заданному конкретному вектору μ^*; проверяются с помощью T^2-статистики Хотеллинга с подстановкой в формулу (6) $\mu = \mu^*$.

II. Гипотезы $\mu^{(1)} = \mu^{(2)}$ о равенстве векторов математических ожиданий в двух генеральных совокупностях (с одинаковыми, но неизвестными ковариационными матрицами), представленных двумя выборками; проверяются с помощью статистики $\tilde{T}^2$.

III. Гипотезы $\mu^{(1)} = \mu^{(2)} = \ldots = \mu^{(k)} = \mu_0$ о равенстве векторов математических ожиданий в нескольких ген. совокупностях (с одинаковыми, но неизвестными ковариационными матрицами), представленных своими выборками, проверяются с помощью статистики:

$$U_{p,k-1,n-k} = \frac{\left|\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{.i}^{(j)} - \hat{\mu}^{(j)}\right)\left(x_{.i}^{(j)} - \mu^{(j)}\right)'\right|}{\left|\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{.i}^{(j)} - \hat{\mu}\right)\left(x_{.i}^{(j)} - \hat{\mu}\right)'\right|},$$

в которой $x_{.i}^{(j)}$ есть i-е p-мерное наблюдение в выборке объёма n_j, представляющей j-ю генеральную совокупность, а $\hat{\mu}^{(j)}$ и $\hat{\mu}$ – оценки вида (3), построенные соответственно отдельно по каждой из выборок и по объединённой выборке объёма $n = n_1 + \ldots + n_k$.

IV. Гипотезы $\mu^{(1)} = \mu^{(2)} = \ldots = \mu^{(k)} = \mu$ и $V_1 = \ldots = V_k = V$ об эквивалентности нескольких нормальных ген. совокупностей, представленных своими выборками

$$\{x_{.i}^{(j)}\}_{i=1}^{n_j},\ j = 1,2,\ldots,k,$$

проверяются с помощью статистики:

$$\lambda = \frac{\prod_{j=1}^{k} |\, n_j \hat{V}_j \,|^{(n_j-1)/2}}{\left|\sum_{j=1}^{k}\sum_{i=j}^{n_j}(x_{.i}^{(j)} - \mu)'\right|^{(n-k)/2}},$$

в которой $\hat{V}_{j.}$ – оценка вида (4), построенная отдельно по наблюдениям j-й выборки, $j = 1,2,\ldots,k$.

V. Гипотезы о взаимной независимости подвекторов-столбцов $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ размерностей соответственно $p_1, p_2, \ldots, p_m$, на которые расчленён исходный p-мерный вектор исследуемых показателей x, $p_1 + p_2 + \ldots + p_m = p$; проверяются с помощью статистики:

$$\Psi = \frac{\left|n\hat{V}\right|}{\prod_{i=1}^{m}\left|n_i\hat{V}_i\right|},$$

в которой V и $\hat{V}_i$ – выборочные ковариационные матрицы вида (4) для всего вектора x и для его подвектора $x^{(i)}$ соответственно.

М.с.а. геометрической структуры исследуемой совокупности многомерных наблюдений объединяет в себе понятия и результаты таких моделей и схем, как дискриминантный анализ, смеси вероятностных распределений, кластер-анализ и таксономия, *многомерное шкалирование*. Узловым во всех этих схемах является понятие расстояния (меры близости, меры сходства) между анализируемыми элементами. При этом анализируемыми могут быть как реальные объекты, на каждом из которых фиксируются значения показателей x, – тогда геометрическим образом i-го обследованного объекта будет точка $x_i = (x_{1i}, \ldots, x_{pi})'$ в соответствующем p-мерном пространстве, так и сами показатели $x_{l.}$, $l = 1,2,\ldots,p$ – тогда геометрическим образом l-го показателя будет точка $x_{l.} = (x_{l1}, x_{l2}, \ldots, x_{ln})'$ в соответствующем n-мерном пространстве.

Методы и результаты дискриминантного анализа направлены на решение заданной задачи. Известно о существовании определённого числа $k \geq 2$ ген. совокупностей и имеется по одной выборке из каждой совокупности («обучающие выборки»). Требуется построить основанное на имеющихся обучающих выборках наилучшее в определённом смысле классифицирующее правило, позволяющее приписать некоторый новый элемент (наблюдение x) к своей ген. совокупности в ситуации, когда заранее неизвестно, к какой из совокупностей

этот элемент принадлежит. Обычно под классифицирующим правилом понимается последовательность действий: по вычислению скалярной функции от исследуемых показателей, по значениям которой принимается решение об отнесении элемента к одному из классов (построение дискриминантной функции); по упорядочению самих показателей по степени их информативности с точки зрения правильного отнесения элементов к классам; по вычислению соответствующих вероятностей ошибочной классификации.

Задача анализа смесей *распределений вероятностей* чаще всего (но не всегда) возникает также в связи с исследованием «геометрической структуры» рассматриваемой совокупности. При этом понятие r-го однородного класса формализуется с помощью ген. совокупности, описываемой некоторым (как правило, унимодальным) законом распределения $P(x \mid \theta_r)$, так что распределение общей ген. совокупности, из которой извлечена выборка (1), описывается смесью распределений вида:

$$P(x) = \sum_{r=1}^{k} \pi_r P(x \mid \theta_r),$$

где π_r – априорная вероятность (удельный вес элементов) r-го класса в общей генеральной совокупности. Задача состоит в «хорошем» статистическом оценивании (по выборке $\{x_i\}_1^n$) неизвестных параметров $\theta_r, \pi_r,$ *а* иногда и k. Это, в частности, позволяет свести задачу классификации элементов к схеме дискриминантного анализа, хотя в данном случае отсутствовали обучающие выборки.

Методы и результаты кластер-анализа (классификация, таксономии, распознавания образов «без учителя») направлены на решение следующей задачи. Геометрическая структура анализируемой совокупности элементов задана либо координатами соответствующих точек (т.е. матрицей $\| x_{ij} \|,\ i = 1,\ldots, p;\ \ j = 1,\ldots, n$), либо набором геометрических характеристик их взаимного расположения, напр., матрицей попарных расстояний

$$\| \rho_{ij} \|_{i,j=1}^{n}.$$

Требуется разбить исследуемую совокупность элементов на сравнительно небольшое (заранее известное или нет) число классов так, чтобы элементы одного класса находились на небольшом расстоянии друг от друга, в то время как разные классы были бы по возможности достаточно взаимоудалены один от другого и не разбивались бы на столь удалённые друг от друга части.

Задача многомерного шкалирования относится к ситуации, когда исследуемая совокупность элементов задана с помощью матрицы попарных расстояний $\| \rho_{ij} \|_{i,j=1}^{n}$, и заключается в приписывании каждому из элементов заданного числа (*p*) координат т.о., чтобы структура попарных взаимных расстояний между элементами, измеренных с помощью этих вспомогательных координат, в среднем наименее отличались бы от заданной. Следует заметить, что осн. результаты и методы кластер-анализа и многомерного шкалирования развиваются обычно без каких-либо допущений о вероятностной природе исходных данных.

Прикладное значение М.с.а. выражается в решении проблем: статистического исследования зависимостей между анализируемыми показателями. Предполагая, что исследуемый набор статистически регистрируемых показателей *x* разбит, исходя из содержательного смысла этих показателей и окончательных целей исследования, на *q*-мерный подвектор $x^{(1)}$ предсказываемых (зависимых) переменных и $(p-q)$-мерый подвектор $x^{(2)}$ предсказывающих (независимых) переменных, можно сказать, что проблема состоит в определении на основании выборки (1) такой *q*-мерной векторной функции $f(x^{(2)})$ из класса допустимых решений *F*, которая давала бы наилучшую, в определённом смысле, аппроксимацию поведения подвектора показателей $x^{(1)}$. В зависимости от конкретного вида функционала качества аппроксимации и природы анализируемых показателей приходят к тем или иным схемам множественной регрессии, дисперсионного, ковариационного или конфлюентного анализа; классификации элементов (объектов или показателей) в общей (нестрогой) постановке заключается в том, чтобы всю анализируемую совокупность элемен-

тов, статистически представленную в виде матрицы $\| x_{ij} \|, i = 1,\ldots, p;\ j = 1,\ldots, n,$ или матрицы $\| \rho_{ij} \|, i, j = 1,\ldots, n,$ разбить на сравнительно небольшое число однородных, в определённом смысле, групп. В зависимости от природы априорной информации и конкретного вида функционала, задающего критерий качества классификации, приходят к тем или иным схемам дискриминантного анализа, кластер-анализа (таксономии, распознавания образов «без учителя»), расщепления смесей распределений; снижения размерности исследуемого факторного пространства и отбора наиболее информативных показателей заключается в определении такого набора сравнительно небольшого числа $m << p$ показателей $z = (z_1, z_2, \ldots, z_m)$, найденного в классе допустимых преобразований $Z(x)$ исходных показателей $x = (x_1, x_2, \ldots, x_p)$, на котором достигается верхняя грань некоторой экзогенно заданной меры информативности m-мерной системы признаков. Конкретизация функционала, задающего меру автоинформативности (т.е. нацеленного на макс. сохранение информации, содержащейся в статистическом массиве (1) относительно самих исходных признаков), приводит, в частности, к различным схемам *факторного анализа* и *гл. компонент*, к методам экстремальной группировки признаков. Функционалы, задающие меру внешней информативности, т.е. нацеленные на извлечение из (1) макс. информации относительно некоторых других, не содержащихся непосредственно в х показателей или явлений, приводят к различным методам отбора наиболее информативных показателей в схемах статистического исследования зависимостей и дискриминантного анализа.

Осн. математический инструментарий М.с.а. составляют специальные методы теории систем линейных уравнений и теории матриц (методы решения простой и обобщённой задачи о собственных значениях и векторах; простое обращение и псевдообращение матриц; процедуры диагонализации матриц и т.д.) и некоторые оптимизационные алгоритмы (методы покоординатного спуска, сопряжённых градиентов, ветвей и границ, различные версии случайного поиска и стохастической и т. д.).

М.с.а. – гл. поставщик инструментария для статистических и эконометрических исследований.

МНОГОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

статистический метод анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния. Изучение влияния факторов по их дисперсиям называется дисперсионным анализом. М.д.а., в том смысле как он обычно понимается и используется, был развит в значительной мере Р. Фишером.

С помощью М.д.а. изучается степень влияния нескольких факторных признаков на результативный признак, то есть решается задача аналогичная множественному корреляционному анализу. Отличие М.д.а. состоит в том, что в ходе его изучается вариация лишь одного признака – результативного, а простейшим показателем вариации служит дисперсия, которая, по правилу сложения дисперсий, может быть разложена на межгрупповую и внутригрупповую.

Межгрупповая дисперсия возникает под действием каких-либо факторов, которые приводят к изменению средних в отдельных группах и отражает вариацию этих средних. Внутригрупповая дисперсия возникает под действием прочих, не учтенных в модели факторов, которые можно назвать случайными и отражает вариацию внутри групп. При разложении общей дисперсии на составляющие можно выделить те её части, которые обусловлены действием одного, другого, третьего и т.д. факторов в общей дисперсии, то есть получить возможность измерять их действие.

Суть метода состоит в разложении общей вариации изучаемого показателя на части, соответствующие раздельному и совместному влиянию факторов, в статистическом изучении этих частей с целью проверки гипотез о существовании этих влияний. Модели дисперсионного анализа в

зависимости от числа факторов классифицируются на однофакторные, двухфакторные и т.д. комплексы. По природе факторов модели дисперсионного анализа принято подразделять на детерминированные (М1), случайные (М2) и смешанные (М3) в зависимости от того, являются ли все уровни факторных признаков фиксированными или случайными, а также когда часть факторов имеют фиксированные уровни, а часть – случайные.

В М.д.а измерение влияния факторов ведётся с помощью дисперсий. При этом нужно различать разложение общей дисперсии, представляющей собой сумму квадратов отклонений всех вариант от средней, от разложения ее числа степеней сумму чисел степеней свободы её слагаемых.

Общая дисперсия признака (дисперсия всего комплекса) – сумма квадратов отклонений всех вариант результативного признака от общей его средней:

$$D_y^2 = \sum\sum (y_{ij} - \bar{y})^2 ,$$

где y_{ij} – отдельные значения результативного признака; $\bar{y}$ – общее среднее значение. Общая дисперсия всегда больше дисперсии, показывающей влияние изучаемых факторов, поскольку в одном исследовании невозможно освободиться от действия множества других факторов, влияющих на результативный признак. При разложении общей дисперсии выделяются факторная дисперсия, идентичная межгрупповой, и случайная, или остаточная, дисперсия, вызванная не учитываемыми в данном опыте факторами, идентичная внутригрупповой.

Н

НАГРУЗКА (В ФАКТОРНОМ АНАЛИЗЕ)

см. в ст. *Матрица нагрузок*

НЕПАРАМЕТРИЧЕСКИЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

раздел *многомерного статистического анализа*, посвящённый получению правил классификации, наблюдений (объектов) в один из нескольких описанных классов (групп, категорий, популяций). В отличие от *дискриминантного анализа* параметрического, непараметрические методы дискриминации не требуют знаний о точном функциональном виде распределений (напр., имеются лишь некоторые общие предположения о законе распределения исследуемого вектора: гладкость, сосредоточенность внутри ограниченной области и т.п.) и позволяют решать задачи дискриминации на основе незначительной априорной информации о совокупностях, что особенно ценно при практическом применении.

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ СТАТИСТИКИ

методы статистического анализа социально-экономических явлений, где используются различные условные оценки, напр., ранги, а взаимосвязь между отдельными признаками измеряются с помощью непараметрических коэффициентов связи, которые исчисляются при условии, что исследуемые признаки подчиняются различным законам распределения. В Н.м.с. используются такие понятия, как «ранжирование» и «ранг». Ранжирование – процедура упорядочения объектов изучения, которая выполняется на основе предпочтения. Ранг – порядковый номер значений признака, расположенных в порядке возрастания или убывания их величин. Если значения признака имеют одинаковую количественную оценку, то ранг всех этих значений принимается равным средней арифметической от соответствующих номеров мест, которые определяют в соответствии со значениями признака; эти ранги называются связными. Напр., дана задача ранжирования пр-тия агропромышленного комплекса одного из регионов РФ по величине балансовой прибыли (см. табл. 1):

Таблица 1

Балансовая прибыль предприятий агропромышленного комплекса одного из регионов РФ в 2008 г. (цифры условные)

№ предприятия	Балансовая прибыль, млн руб.	Ранжирование (ранги)
1	10	6,5
2	20	4
3	10	6,5
4	20	4
5	20	4
6	50	2
7	70	1

Наиболее предпочтительному пр-тию, величина балансовой прибыли которого наибольшая, присваивается ранг «1»; затем в порядке уменьшения величины балансовой прибыли проранжируются все рассматриваемые пр-тия агропромышленного комплекса.

Принцип нумерации значений исследуемых признаков – основа Н.м.с. для изучения взаимосвязи между социально-экономическими явлениями и процессами. Среди непараметрических методов оценки тесноты связи наибольшее значение имеют ранговые коэффициенты корреляции Спирмена и Кендалла. Эти коэффициенты используются для определения тесноты связей как между количественными, так и между качественными признаками при условии, если их значения проранжировать по степени убывания или возрастания признака.

Под ранговой корреляцией понимается статистическая связь, существующую между двумя или несколькими ранжировками одного и того же конечного множества объектов $O_1, O_2, ..., O_n$. Измерить и проанализировать совокупную статистическую связь, существующими между ранжировками одних и тех же объектов $O_1, O_2, ..., O_n$, полученными в соответствии со степенью проявления в них сначала свойства $x^{(k)}$ (первый способ ранжирования), затем – свойства $x^{(j)}$ (второй способ ранжирования), $k, j = 1, 2, ..., p,$ где p – количество рассматриваемых свойств (признаков), составляет основную задачу непараметрических методов статистики.

Коэффициент корреляции рангов (коэффициент К.Спирмена, предложен им в 1904) рассчитывается по формуле (для случая, когда нет связных рангов в обеих исследуемых ранжировках):

$$\rho_{kj} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (x_i^{(k)} - x_i^{(j)})^2 ,$$

$k, j = 1, 2, ..., p,$ используется для измерения степени тесноты связи между ранжировками

$$X^{(k)} = (x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)})^T$$

и

$$X^{(j)} = (x_1^{(j)}, x_2^{(j)}, ..., x_n^{(j)})^T .$$

Коэффициент корреляции рангов Спирмена принимает любые значения на отрезке [-1; 1]. Значимость коэффициента корреляции рангов Спирмена проверяется на основе t- критерия Стьюдента. Расчётное значение критерия определяется по формуле

$$t_p = \rho_{kj} \sqrt{\frac{n-2}{1-\rho_{kj}^2}} .$$

Значение коэффициента корреляции Спирмена считается статистически значимым, т.е. не равным нулю, если $t_p > t_{kp}(\alpha; l = n-2)$, где $t_{kp}(\alpha; l = n-2)$ – критическая точка двусторонней критической области, которую находят по таблице критических точек распределения Стьюдента по уровню значимости α и числу степеней свободы $l = n-2$. Напр., два экспер-

та проранжировали 10 предложенных им проектов реорганизации нефтегазовой компании с точки зрения их эффективности при заданных ресурсных ограничениях. Пронумеровав проекты в порядке ранжировки 1- го эксперта, получаем в качестве исходных данных: $X^{(1)} = (1, 2, 3, 4, 5, 6, 7, 8, 9,10)^T$; $X^{(2)} = (2, 3, 1, 4, 6, 5, 9, 7, 8,10)^T$. Вычисление коэффициента Спирмена для случая, когда нет связных рангов в обеих исследуемых ранжировках, даёт следующий результат:

$$\rho_{12} = 1 - \frac{6}{1000-10}(1+1+4+0+1+1+4+1+1+0) = 1 - \frac{6}{990}\cdot 14 = 0{,}915,$$

что свидетельствует о существенной положительной ранговой связи между исследуемыми переменными.

Если совокупность значений по исследуемому признаку содержит связные ранги, то коэффициент корреляции Спирмена между ранжировками $X^{(k)}$ и $X^{(j)}$ вычисляется по формуле:

$$\rho^*_{kj} = \frac{\frac{1}{6}(n^3-n) - \sum_{i=1}^{n}(x_i^{(k)} - x_i^{(j)})^2 - T^{(k)} - T^{(j)}}{\sqrt{\left[\frac{1}{6}(n^3-n) - 2T^{(k)}\right]\left[\left[\frac{1}{6}(n^3-n) - 2T^{(j)}\right]\right]}},\ T^{(l)} = \frac{1}{2}\sum_{t=1}^{m^{(l)}}\left[(n_t^{(l)})^3 - n_t^{(l)}\right],\ \text{l=k, j;}$$

где $m^{(l)}$ – число групп одинаковых рангов у переменной $x^{(l)}$, а $n_t^{(l)}$ – число элементов (рангов), входящих в t - ю группу одинаковых рангов. В частном случае отсутствия групп одинаковых рангов имеем

$$m^{(l)} = n,\ n_1^{(l)} = n_2^{(l)} = \ldots = n_n^{(l)} = 1$$

и соответственно $T^{(l)} = 0$; кроме того, группы одинаковых рангов, состоящие из единственного элемента, по существу не участвуют в расчете величины $T^{(l)}$.

Другая широко используемая характеристика статистической связи между двумя ранжировками – ранговый коэффициент корреляции Кендалла, определяемый для случая отсутствия связных рангов в обеих ранжировках соотношением:

$$\tau_{kj} = 1 - \frac{4\nu(X^{(k)}, X^{(j)})}{n(n-1)},$$

где $\nu(X^{(k)}, X^{(j)})$ – миним. число обменов соседних элементов последовательности $X^{(j)}$, необходимое для приведения к упорядочению $X^{(k)}$. Коэффициент корреляции рангов Кендалла принимает любые значения на отрезке [-1; 1].

Если совокупность значений по исследуемому признаку содержит связные ранги, то коэффициент корреляции Кендалла между ранжировками $X^{(k)}$ и $X^{(j)}$ вычисляется по формуле:

$$\tau^*_{kj} = \frac{\tau_{kj} - \frac{2(U^{(k)} + U^{(j)})}{n(n-1)}}{\sqrt{\left(1 - \frac{2U^{(k)}}{n(n-1)}\right)\left(1 - \frac{2U^{(j)}}{n(n-1)}\right)}},$$

где $U^{(l)} = \frac{1}{2}\sum_{t=1}^{m^{(l)}} n_t^{(l)}(n_t^{(l)} - 1)$, *l=k, j.*

Смысл величин $m^{(l)}$ и $n_t^{(l)}$ определён выше. Как правило, коэффициент Кендалла меньше коэффициента Спирмена. При достаточно большом объёме совокупности значения данных коэффициентов имеют зависимость: $\tau_{kj} = \frac{2}{3}\rho_{kj}$. Связь между признаками можно признать статистически значимой, если значения коэффициентов ранговой корреляции Спирмена и Кендалла больше 0,5. Кроме указанных коэффициентов в Н.м.с. используются достаточно много коэффициентов: *коэффициенты конкордации*, ассоциации, взаимной сопряжённости Чупрова и т.д.

НЕРАВЕНСТВО ЧЕБЫШЕВА

см. в ст. *Закон больших чисел.*

НЕЧЁТКОЕ МНОЖЕСТВО

(от англ. – fuzzy set) – некоторая совокупность $A = \{(x, \mu_A(x)) | x \in X\}$, где X – фундаментальное множество; $\mu_A(x)$ – функция принадлеж-

ности или характеристическая функция. Функция принадлежности $\mu_A(x)$ принимает значения на некотором упорядоченном множестве принадлежностей M, в качестве которого часто выбирают отрезок [0; 1].

Если множество состоит только из двух элементов {0; 1}, то его рассматривают как обычное множество. Для Н.м. отображение элемента в 0 означает, что элемент не принадлежит данному множеству, 1 – полная принадлежность множеству, значения между 0 и 1 характеризует нечеткие элементы.

Теория Н.м. является расширением классической теории множеств. Впервые предложена Лотфи А. Заде в 1965. Её можно рассматривать как теорию случайных множеств, осн. идея которой сводится к тому, что значение функции принадлежности $\mu_A(x)$ рассматривают как вероятность накрытия элемента *x* некоторым случайным множеством. Однако на практике она выступает как альтернатива теории вероятностей и прикладной статистики.

НОРМАЛИЗУЮЩЕЕ ПРЕОБРАЗОВАНИЕ

функциональное преобразование, обеспечивающее лучшее приближение закона распределения результирующей переменной к нормальному, чем исходное.

Использование преобразования переменных в теории и практике статистики имеет длительную историю. Напр. известны z-преобразование Фишера для коэффициента корреляции, преобразование кубического корня для распределения хи-квадрат, арксинуса для долей или пропорций, логарифмическое, Бокса – Кокса и др.

Чаще всего преобразование переменных используют для устранения асимметрии или островершинности распределения, упрощения статистических выводов, лучшего приближения к предположениям классических линейных моделей.

Иногда выбор функциональной формы вытекает из сущности анализируемых показателей. Так, для процессов, характеризующихся экспоненциальным ростом, естественным является переход к логарифмам исследуемых переменных.

Выбор нормализующего преобразования зависит от конкретной ситуации. Например, если плотность выборочного распределения унимодальна и скошена влево, то для того, чтобы добиться симметрии и лучшего приближения к нормальности, может использоваться логарифмическое преобразование $z=\ln(x)$, или $z=\ln(x+1)$, если диапазон значений исследуемого показателя включает нулевое значение, либо степенное преобразование $z=x^a$, например, $z=\sqrt[3]{x},\sqrt[4]{x},\ldots$. В случае скошенности вправо – преобразования вида $z=x^{3/2}$, x^2, x^3 и т.д.

Если закон распределения анализируемого признака логарифмически нормальный, то логарифмическое преобразование по определению позволит получить переменную, подчиняющуюся нормальному закону. В иных случаях можно надеяться добиться лишь приближенной нормальности.

Для переменной *u*, подчиняющейся χ^2 – распределению с *v* степенями свободы (*v>30*), величина

$$z=\sqrt[3]{\frac{u}{v}}-\frac{1-2/9v}{\sqrt{2/9v}}$$

– приближенно стандартная нормальная.

Иногда добиться приближения к нормальности позволяет извлечение квадратного корня $z=\sqrt{x}$, $z=\sqrt{x+1}$. Если значения изучаемой переменной лежат в интервале (0,1) полезным может быть преобразование $z=\sqrt{x}+\sqrt{x+1}$.

Когда значения случайной переменной лежат в интервале *(a,b)*, то значения преобразованной переменной

$$z=\ln\left(\frac{x-a}{x-b}\right)$$

могут изменяться в интервале от -∞ до +∞. Возможно, что полученная таким способом переменная *z* окажется также приближенно нормальной. Так, подобное преобразование применяется для выборочного коэффициента корреляции. Выборочный коэффициент корреляции *r* принимает значения в интервале (-1,1), выбо-

рочное распределение сильно скошено и зависит от значения коэффициента корреляции ρ в исходной генеральной совокупности. Преобразование Фишера

$$z = \frac{1}{2}\ln(\frac{r+1}{1-r})$$

позволяет получить переменную с приближенно нормальным выборочным распределением с математическим ожиданием

$$\frac{1}{2}\ln(\frac{\rho+1}{1-\rho}) + \frac{\rho}{2(n-1)}$$

и дисперсией

$$\frac{1}{n-3},$$

где n – объём выборки.

Если закон распределения исследуемой непрерывной случайной величины известен, то добиться полной нормализации возможно с помощью преобразования интеграла вероятности $z = \Phi^{-1}(F(x))$, где *F(x)* – функция распределения случайной величины *X* в точке *x*. Нормализующее преобразование зависимой переменной часто используют, чтобы добиться стабилизации дисперсии остатков в линейной модели:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i .$$

Как правило, теория линейных моделей предполагает аддитивность функциональной формы, постоянство дисперсии остатков, нормальный закон распределения случайной составляющей.

Обобщение степенного и логарифмического преобразования зависимой переменной – преобразование Бокса-Кокса:

$$z_i = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y_i), & \lambda = 0 \end{cases},$$

где λ – параметр, рассчитываемый по итерационному алгоритму для конкретного набора данных. Если диапазон данных включает отрицательные величины, в формулу дополнительно включают параметр сдвига *a* – положительную константу. Сравнительно простой интерпретации поддаются значения параметра λ из диапазона (-2, -1, – ½, – 1/3, 0, ¼, 1/3, ½, 2/3, 1, 2).

С вычислительной точки зрения часто легче работать с эквивалентным степенным нормализующим преобразованием:

$$z_i = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda\, \bar{y}^{\lambda-1}}, & \lambda \neq 0 \\ \bar{y}\ln(y_i), & \lambda = 0 \end{cases},$$

где $\bar{y} = \sqrt[n]{y_1 \cdot y_2 \cdot ... \cdot y_n}$ – среднее геометрическое наблюдений $y_1, y_2, ..., y_n$.

Если зависимая переменная является пропорцией или вероятностью в качестве нормализующих могут применяться логит, пробит и арксинус преобразования. Если исследуемая переменная подчиняется равномерному закону на интервале (0,1), то в качестве нормализующего может использоваться пробит – преобразование $z = \Phi^{-1}(y)$, где *Φ(y)* – функция распределения стандартной нормальной переменной. Приближённой нормальности можно добиться с помощью логит – преобразования $z = \ln\frac{y}{1-y}$. Если моделируемые вероятности не слишком близки к 0 или 1, напр., лежат в интервале от 0,2 до 0,8, то регрессии с использованием логит, пробит или arcsin преобразований зависимой переменной дают очень близкие результаты. Однако, для логистичекого преобразования обратную функцию можно записать в явном виде, что существенно упрощает выкладки.

О

ОБРАЩЕНИЕ МАТРИЦЫ

(от англ. – matrix inversion) – операция получения матрицы A^{-1}, обратной к заданной матрице A, т.е. операция, удовлетворяющая условию: $AA^{-1}=A^{-1}A=I$, где I – единичная матрица.

Если задана невырожденная матрица *A*, то обратная ей матрица в общем виде вычисляется:

$$A^{-1} = (a_{ij}^{обр})_{n\times n} = \frac{[(-1)^{i+j} \det A_{ji}]_{n\times n}}{\det A}$$

где $[(-1)^{i+j}\det A_{ji}]_{n\times n}$ – транспонированная матрица *алгебраических дополнений*; $\det A$ – *определитель* (*детерминант*).

Свойства обратных матриц: 1. матрица A^{-1} для любой невырожденной матрицы A является единственной; 2. $(A^{-1})^{-1}=A$; 3. $(AB)^{-1}=B^{-1}A^{-1}$; 4. $(A^{-1})'=(A')^{-1}$; 5. $(A^{-1})^m=(A^m)^{-1}$.

С помощью обратной матрицы в межотраслевом балансе (МОБ) определяются валовые выпуски продукции, необходимые для получения заданных компонентов конечного продукта, т.е. осуществляется решение уравнений МОБ. Обратная матрица также служит важным инструментом вычислительной процедуры *регрессионного анализа*.

ОБРАЩЕНИЕ БЛОЧНОЙ МАТРИЦЫ

если дана *блочно-диагональная матрица*, т.е. внедиагональные элементы равны нулю:

$$A=\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}$$

блочная матрица порядка *n*, где A_{11}, A_{22} – квадратные невырожденные матрицы одинакового порядка. Обратная к ней матрица имеет вид:

$$A^{-1}=\begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{pmatrix}.$$

В общем случае, когда

$$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

и если матрицы A и $B=A_{22}-A_{21}A_{11}^{-1}A_{12}$ невырождены, то:

$$A^{-1}=\begin{pmatrix} A_{11}^{-1}+A_{11}^{-1}A_{12}B^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}B^{-1} \\ -B^{-1}A_{21}A_{11}^{-1} & B^{-1} \end{pmatrix}.$$

Или, если матрица A и $C=A_{11}-A_{12}A_{22}^{-1}A_{21}$ невырождены, то

$$A^{-1}=\begin{pmatrix} C^{-1} & -C^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}C^{-1} & A_{22}^{-1}+A_{22}^{-1}A_{21}C^{-1}A_{12}A_{22}^{-1} \end{pmatrix}.$$

ОБУЧАЮЩАЯ ВЫБОРКА

выборка, по которой производится настройка параметров аналитической модели зависимости; используется для обучения аналитических моделей – искусственных *нейронных сетей*, деревьев решений, *самоорганизующихся карт Кохонена* и др. О.в. включает наблюдения, для которых известны значения признаков, характеризующих как управляемые (факторные) показатели, так и целевой результат.

В задачах *классификации многомерных наблюдений с обучением* О.в. – набор объектов, для каждого из которых известно, к которому из классов он принадлежит.

О.в. должна включать достаточное количество наблюдений, как можно более полно отражающих правила и закономерности исследуемого процесса. Наличие пропусков, аномальных значений или противоречий снижает качество обучения модели.

Оценку качества модели целесообразно производить по контрольной выборке – независимому, не использованному для обучения массиву данных, поскольку показатели качества, рассчитанные по О.в., как правило, оказываются завышенными.

См. также *Дискриминантный анализ.*

ОПРЕДЕЛИТЕЛЬ (ДЕТЕРМИНАНТ) БЛОЧНОЙ МАТРИЦЫ

пусть

$$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

блочная матрица порядка *n*, где A_{11}, A_{22} – квадратные матрицы одинакового порядка О.б. м.:

$$|A|=\det A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}=\begin{cases} \det(A_{11}A_{22}-A_{21}A_{12}), \ \textit{если} \ A_{11}A_{21}=A_{21}A_{11}; \\ \det(A_{22}A_{11}-A_{21}A_{12}), \ \textit{если} \ A_{11}A_{12}=A_{12}A_{11}; \\ \det(A_{11}A_{22}^{T}-A_{12}A_{21}^{T}), \ \textit{если} \ A_{11}A_{21}=A_{21}A_{11}. \end{cases}$$

Или, если существует *обратная матрица* A_{11}^{-1}, то О.б.м. порядка *n*:

$$\det A = \det\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \det A_{11} \cdot \det(A_{22} - A_{21}A_{11}^{-1}A_{12})$$

.

Если существует обратная матрица A_{22}^{-1}, то О.б.м. порядка *n*:

$$\det A = \det\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \det A_{22} \cdot \det(A_{11} - A_{12}A_{22}^{-1}A_{21}) .$$

Частный случай:

$$\det A = \det\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} = \det A_{11} \det A_{22} = \det\begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}.$$

ОПРЕДЕЛИТЕЛЬ (ДЕТЕРМИНАНТ) МАТРИЦЫ

см. в ст. *Детерминант (определитель) матрицы*

П

ПАРАЛЛЕЛЬНЫЕ КЛАСТЕР-ПРОЦЕДУРЫ

процедуры автоматической классификации, использующие на каждом шаге все имеющиеся наблюдения.

Существует значительное число алгоритмов, реализующих идею оптимизации некоторого функционала качества разбиения. При известном числе классов в качестве функционала качества разбиения часто используют характеристики: сумму внутри-классовых дисперсий; сумму внутриклассовых расстояний между элементами; обобщенную внутриклассовую дисперсию. При неизвестном числе классов функционал качества разбиения выбирают в виде алгебраической комбинации двух функционалов, один из которых характеризует внутриклассовый разброс наблюдений, а другой является возрастающей функцией числа классов. Выбор функционала качества разбиения обусловлен, прежде всего, его адекватностью содержательным посылкам решаемой задачи.

Полный перебор всех вариантов разбиений становится практически неосуществи-мым даже при сравнительно небольшом числе объектов. Поэтому большинство алгорит-мов реализуют идею сокращенного перебора. Напр., с помощью какого-либо быстрого алгоритма получают начальное разбиение, а затем, если это приводит к возрастанию значений выбранного функционала качества, последовательно переносят объекты из класса в класс. Работа алгоритма заканчивается, когда перемещение наблюдений перестают приводить к улучшению качества разбиения. Большой класс алгоритмов реализует идею эталонных точек (множеств), понятие центра тяжести и др. см [1-3].

К недостаткам П.к.-п. следует отнести высокие требования к вычислительным ресурсам, что огранивает их использование их при большом числе объектов.

См. также *Кластерный анализ*.

ПАРАМЕТРИЧЕСКИЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

метод классификации многомерных наблюдений при наличии предположений о виде закона распределения вероятностей классов.

Наиболее важный частный случай – нормальный *многомерный закон распределения вероятностей*. Пусть *k* – мерные наблюдения *x*, извлечены из классов π_1, π_2, описываемых многомерным нормальным законом распределения вероятностей со средними μ_1, μ_2 и одинаковой ковариационной матрицей Σ. Правило классификации состоит в следующем: наблюдение *x* относится к классу π_1, если выполняется неравенство:

$$D_T(x) = \left[x - \frac{1}{2}(\mu_1 + \mu_2) \right]^T \Sigma^{-1}(\mu_1 - \mu_2) > \ln \frac{p_2}{p_1} .$$

$D_T(x)$ называют теоретической дискриминантной функцией. Её выборочный аналог получают, подставляя вместо μ_i выборочные средние $\bar{x}_i$ и оценку ковариационной матрицы *S* вместо Σ.

Если ковариационный матрицы классов различны Σ_1, Σ_2 приходят к квадратичной дискриминантной функции:

$$D_T(x) = \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}[x-\mu_1)]^T\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}[x-\mu_2)]^T\Sigma_2^{-1}(x-\mu_2) > \ln\frac{p_2}{p_1}.$$

При большом числе признаков и различных ковариационных матрицах линейное и квадратичное решающие правила дают существенно различные результаты классификации.

В случае *k* классов с общей ковариационной матрицей объект следует отнести к классу π_i, если

$$\ln p_i + (x - \frac{1}{2}\mu_i)^T\Sigma^{-1}\mu_i = \max\left[\ln p_j + (x - \frac{1}{2}\mu_j)^T\Sigma^{-1}\mu_j\right].$$

Важным шагом после построения дискриминантной функции является вопрос о качестве классификации будущих наблюдений. Для нормальных классов с одинаковой ковариационной матрицей теоретическая вероятность ошибочной классификации может быть вычислена как функция априорной вероятности p_1 и расстояния Махалонобисса между классами

$$\delta^2 = (\mu_1 - \mu_2)^T\Sigma_1^{-1}(\mu_1 - \mu_2).$$

Проведённые исследования устойчивости оценок вероятностей неправильной классификации позволяют сформулировать ряд существенных выводов. Параметрические оценки вероятности ошибочной классификации не являются робастными и не могут быть рекомендованы для практического использования. Оценки вероятности ошибочной классификации по обучающей выборке занижены. Лучше себя зарекомендовали процедуры скользящего экзамена. См. также *Дискриминантная функция линейная (Фишера)*.

ПАРНЫЕ СРАВНЕНИЯ

метод исследования отношений между некоторыми совокупностями путём их попарного сопоставления по степени выраженности общего для них свойства.

Метод П.с. широко используется при изучении предпочтений, отношений и построения шкальных значений для неподдающихся непосредственному измерению показателей в социологии, психологии, маркетинге, экспертных оценках, планировании спортивных соревнований и других областях.

Результат сравнения представляется величиной α_{ijk}, принимающей значение 1 или 0, в зависимости от того является ли в соответствии с заданным критерием *i*-я альтернатива предпочтительней *j*-й ($T_i \succ T_j$) в *k*-м сравнении этих объектов. Наиболее известные вероятностные модели П.с. – модели Брэдли-Терри и Терстоуна - Мостеллера. Предполагается, что все сравнения проводятся независимо друг от друга, так что случайные величины α_{ijk} независимы в совокупности.

В модели Брэдли-Терри объектам $T_1, T_2, ..., T_t$ сопоставляются параметры

$$\pi_1, \pi_2, ..., \pi_t, \quad \pi_i \geq 0, \quad i = 1,2,...,t, \quad \sum_{i=1}^{t}\pi_i = 1,$$

трактуемые как вероятности извлечения соответствующих альтернатив. Вероятность предпочтения объекта T_i при сравнении с T_j задается выражением:

$$P(T_i \succ T_j) = \pi_i/(\pi_i + \pi_j) = 1 - P(T_j \succ T_i), \quad i \neq j, \quad i, j = 1,...,t.$$

Оценки параметров π_i могут быть найдены *методом макс. правдоподобия*. Полученные компоненты нормализованного вектора π используются как количественные оценки важности

объектов. Модель Терстоуна-Мостеллера формализует концепцию субъективного континуума, внутренней шкалы восприятия, которая используется человеком для упорядочивания объектов по изучаемому свойству, но не поддается непосредственному измерению.

Каждой альтернативе приписывается некоторое вещественное число μ_i. В опытах фиксируется восприятие объекта X_i – нормально распределенная случайная величина со средним μ_i.

Предполагается, что индивидуум при сравнении объектов T_i и T_j сообщает об упорядочении восприятий X_i и X_j. Так как объекты предъявляются одновременно, допускается корреляция восприятий. При предположении о постоянстве дисперсий X_i и постоянстве корреляций между X_i и X_j вероятность предпочтения объекта T_i при сопоставлении с T_j задается выражением:

$$P(T_i \succ T_j) = P(X_i \succ X_j) = \frac{1}{\sqrt{2\pi}} \int_{-(\mu_i - \mu_j)}^{\infty} \exp\left(-\frac{1}{2} y^2\right) dy .$$

Модели Терстоуна - Мостеллера и Брэдли - Терри – частные случаи линейной модели $P(T_i \succ T_j) = H(V_i - V_j)$, где V_i – «ценность» T_i, H – симметричная функция распределения $H(-x) = 1 - H(x)$, в случаях нормального и логистического распределений. На практике модели Терстоуна – Мостеллера и Брэдли – Терри приводят к сходным результатам.

Линейное упорядочение объектов, адекватно воспроизводящее результаты П.с., не всегда возможно построить. Иногда более соответствующим оказывается представление объектов точками на плоскости или в пространстве большей размерности.

Если результат сравнения представлен не бинарной, а *количественной переменной*, то для анализа результатов может использоваться *дисперсионный анализ*. См. также *Многомерное шкалирование метрическое*, *Многомерное шкалирование неметрическое.*

ПАССИВНЫЙ ЭКСПЕРИМЕНТ

эксперимент, в котором уровни факторов фиксируются исследователем, но не задаются. Информация об исследуемом объекте накапливается путём пассивного наблюдения, т.е. информацию получают в условиях обычного функционирования объекта, в отличие от активного эксперимента, когда наблюдение проводится с применением искусственного воздействия на объект по специальной программе.

ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА

раздел *математической статистики*, изучающий рациональную организацию измерений, подверженных случайным ошибкам. П.э. эффективно применяется при поисковых исследованиях, при неполном знании механизма явления, подтверждении или уточнении параметров уже известных математических описаний этих явлений. Оптимальная организация исследований позволяет уменьшить количество опытов, сократив тем самым расходы на их проведение и время на обработку данных, уменьшить ошибку эксперимента, выработать четкие формализованные правила принятия решения на каждом этапе проведения эксперимента и получить многофакторные математические модели с желаемыми статистическими свойствами.

Цель эксперимента – оценка всех или некоторых параметров вектора неизвестных параметров θ или их функций или проверка некоторых гипотез о параметрах θ. На основе цели формируется критерий оптимальности плана эксперимента. П.э. заключается в получении значений, которые задаются переменным *x* в эксперименте. Переменные *x* по выбору экспериментатора могут принимать значения из некоторого допустимого множества *X*.

Стратегия экспериментирования определяется требованиями, предъявляемыми к модели, и заключается в выборе соответствующего плана размещения экспериментальных точек в заданной области пространства исходя из критериев оптимальности. Критерии оптимальности характеризуют качество плана, точность линей-

ных оценок параметров модели и прогностические свойства модели в целом.

ПОКАЗАТЕЛИ КЛАССИФИКАЦИОННЫЕ (НОМИНАЛЬНЫЕ)

показатели, значения которых измерены в номинальной шкале (шкале наименований), представляющей собой простое перечисление различающихся между собой явлений или объектов. Объекты могут быть произвольным образом пронумерованы, причём цифры, присваиваемые различным градациям, служат лишь для отличий их друг от друга. Единственными отно-шениями между объектами, располагаемыми по номинальной шкале, выступают отношения тождества и отличия. Номера градаций изменяются взаимно однозначным образом.

ПОКАЗАТЕЛИ КАЧЕСТВЕННЫЕ

показатели классификационные и порядковые. Порядковые признаки, значения которых измерены в порядковой шкале позволяют упорядочить объекты друг относительно друга. Ранги, присвоенные изучаемой переменной, должны соответствовать изменению интенсивности изучаемого признака у объектов. Допустимо монотонное преобразование шкалы, т.к. отношения упорядочения между объектами при этом сохраняются.

ПОКАЗАТЕЛИ КОЛИЧЕСТВЕННЫЕ

показатели, измеренные в количественной шкале, которая устанавливает не только отношения порядка, но и величину интервалов между градациями. Теория измерений выделяет следующие типы количественных шкал: интервалов, отношений, разностей, абсолютная. В случае интервальной шкалы можно произвольным образом выбирать начало отсчёта и единицу измерения. В шкале отношений указывают начало отсчёта – нуль. В шкале разностей есть единица измерения, но нет естественного начала отсчета. В абсолютной шкале фиксированы как начало отсчёта, так и единица измерения.

В большинстве статистических процедур не делается различий между свойствами П.к., измеренных в интервальной шкале, шкале отношений, разностей или абсолютной.

ПОСЛЕДОВАТЕЛЬНЫЕ КЛАСТЕР-ПРОЦЕДУРЫ

процедуры *автоматической классификации* многомерных наблюдений, на каждом шаге которых используется небольшая часть исходных наблюдений. П.к.-п. особенно успешно применяются при большом числе наблюдений.

Наиболее популярны и изучены – модификации *метода k-средних*. Обозначим k – заданное число классов. На первом шаге тем или иным способом строится начальное приближение – выбираются эталоны, т.е. центры каждого из k классов. В качестве центров могут быть взяты координаты случайно извлеченных из выборки наблюдений, координаты достаточно удаленных друг от друга объектов, центры случайно сгенери-рованных классов и т.п. Из массива данных извлекаются наблюдение, вычисляются расстояния до текущих центров классов. Наблюдение относится к классу, расстояние до центра которого – наименьшее. Координаты центра этого класса (эталона) пересчитываются. Извлекается следующее наблюдение и т.д.

По исчерпанию наблюдений возможно повторение итераций, т.е. массив данных может просматриваться многократно. Процесс извлечения наблюдений повторяется до стабилизации центров классов, выполнения заданного числа итераций или достижения иного критерия. По завершении работы алгоритма производится окончательное разбиение на классы. Для каждого объекта номер класса определяется по принципу наименьшего расстояния до эталона. Существуют модификации алгоритма *k*-средних при неизвестном числе классов.

К достоинствам метода *k*-средних следует отнести его простоту и скорость, что позволяет анализировать большие совокупности наблюдений. К недостаткам – зависи-мость результатов от выбора начальных центров классов.

Другие популярные методы – метод нечётких *k*-средних (*c*-means), метод *k*-медиан, который используют тогда, когда желательно, чтобы эталон совпадал с реально сущест-вующим объектом.

См. также *Кластерный анализ.*

ПОСЛЕДОВАТЕЛЬНЫЙ АНАЛИЗ

способ *проверки статистических гипотез*, особенность которого состоит в том, что число производимых наблюдений не фиксируется заранее, а выбирается по ходу наблюдений в зависимости от поступающих данных.

Осн. достоинство П.а. по сравнению со способами, в которых число наблюдений заранее фиксировано, заключается в значительно меньшем среднем числе наблюдений.

Исторически стимулом к развитию П.а. явилась необходимость снижения издержек и объёма наблюдений в задачах контроля качества продукции при принятии решения признать большую партию пром. продукции бракованной или годной.

А. Вальд предложил схему последовательного критерия отношения вероятностей, когда наблюдения извлекаются и анализируются поочередно и решение (напр., признать партию годной, признать партию бракованной, продолжить наблюдения) принимается на каждом шаге.

Последовательный метод характеризуется двумя важнейшими особенностями: правилом (процедурой) остановки, которое определяет продолжать или заканчивать сбор данных и решающим правилом, определяющим, какие действия должны быть выполнены на очередном шаге. Заметим, что способы проверки статистических гипотез с фиксированным объёмом выборки можно рассматривать как частный случай последовательных процедур, когда правило остановки заключается в извлечении выборки заранее определенного объёма *m*.

Осн. принципы П.а. Пусть $X_1,\ldots,X_n$ случайные величины соответствующие наблюдениям $x_1,\ldots,x_n$. Предположим, что X_i – независимые одинаково распределенные случайные величины с плотностью распределения $f_\theta(x)$, где θ – неизвестный параметр, принадлежащий некоторому параметрическому множеству Ω. Проблема состоит в проверке гипотезы $H_0:\theta\in\omega_0$ против альтернативной гипотезы $H_1:\theta\in\omega_1$, где ω_0,ω_1 – непересекающиеся подмножества значений параметра θ.

Последовательная процедура проверки гипотезы H_0 против альтернативы H_1 определяется парой (N,D). Правило остановки N (окончательный объем выборки) – множество правил устанавливающих, следует ли на шаге N=n по результатам обработки выборочных данных $(x_1,\ldots,x_n)$ завершить, либо продолжить процесс извлечения наблюдений для каждого n≥1. Решающая функция D – множество правил, применяемых на шаге N=n, определяющих принять гипотезу H_0 (т.е. $D=H_0$), либо отвергнуть гипотезу H_0 (т.е. $D=H_1$), на основании результатов наблюдений $(x_1,\ldots,x_n)$. Если $P_\theta(N<\infty)=1$ для любого $\theta\in\Omega$ говорят, что последовательная процедура закачивается с вероятностью 1. Если $P_\theta(N=m)=1$ для любого $\theta\in\Omega$, то объем выборки заранее фиксирован. Ожидаемый (средний) объём выборки в последовательной процедуре определяется выражением:

$$ASN=M_\theta(N)=\sum_{n=1}^{\infty}P_\theta(N=n).$$

Функции $Q_\theta=P_\theta(D=H_0)$ и $P_\theta=P_\theta(D=H_1)$, описывающие зависимость вероятности принятия гипотез H_0 и H_1 от значений параметра θ называют, соответственно, оперативной характеристикой и функцией мощности последовательного теста. Если процедура заканчивается с вероятностью 1, то $P_\theta+Q_\theta=1$ для любого θ. Вероятности ошибок первого и второго рода определяются как $\alpha_\theta=P_\theta$ для $\theta\in\omega_0$ и $\beta_\theta=Q_\theta$ для $\theta\in\omega_1$.

Цель П.а. – определить оптимальные по некоторому критерию правило остановки и решающее правило. Критерии выбора оптимальных последовательных процедур обычно учитывают совокупность следующих параметров: закон распределения окончательного объёма выборки N, средний объём выборки ASN, вероятности ошибок первого и второго рода, стоимость од-

ного эксперимента. Пример последовательных процедур – последовательный критерий отношения вероятностей (*последовательный критерий Вальда*), который применим к широкому кругу ситуаций, когда проблема формулируется в терминах проверки гипотезы о единственном параметре в отсутствии мешающих параметров.

Существуют многочисленные варианты построения последовательных процедур для проверки сложных гипотез, присутствия мешающих параметров, исследованы асимптотические свойства процедур и т.д.

ПОСЛЕДОВАТЕЛЬНЫЙ КРИТЕРИЙ ВАЛЬДА

критерий отношения вероятностей, определяющий правило остановки и решающее правило.

Рассмотрим последовательную случайную выборку $x_1,\ldots,x_n$ из распределения вероятностей, зависящего от одного неизвестного параметра θ. Допустим подмножества ω_0, ω_1 содержат по единственному элементу θ_0, θ_1 соответственно, т.е. требуется проверить простую гипотезу $H_0 : \theta = \theta_0$ против альтернативной $H_1 : \theta = \theta_1$.

Значения вероятности ошибки первого рода α и вероятности ошибки второго рода β фиксированы. Последовательный критерий отношения вероятностей определяет правило остановки как:

N = *первое* $n \geq 1$, *такое, что* $\lambda_n \leq A$ *или* $\lambda_n \geq B$;

$$N = \infty,\ A < \lambda_n < B\ \ \forall n \geq 1,$$

где $B < A \leq \infty$ и λ_n – отношение правдоподобия

$$\lambda_n = \prod_{i=1}^{n} \frac{f_{\theta_1}(x_i)}{f_{\theta_0}(x_i)}.$$

Решающее правило определяется как

$$D = \begin{cases} H_0,\ \lambda_n \leq B \\ H_1, \lambda_n \leq A \end{cases}.$$

Т.о., процедура на очередном шаге n такова: если $\lambda_n \leq B$ то принимается H_0 и наблюдения заканчиваются; если $\lambda_n \geq A$, то принимается H_1 и наблюдения заканчиваются; если $B < \lambda_n < A$, то делается еще одно наблюдение.

Справедливы неравенства, связывающие пороги с вероятностями ошибок:

$$B \geq \frac{\beta}{1-\alpha},$$
$$A \leq \frac{1-\beta}{\alpha}.$$

Вместо неизвестных значений A и B можно взять их приближенные значения A' и B':

$$B' = \frac{\beta}{1-\alpha},$$
$$A' = \frac{1-\beta}{\alpha}.$$

При таком выборе порогов вероятности ошибок I и II рода будут равны не □ и□□, а некоторым □′ и □′, которые несущественно меньше требуемых □ и □.

Вальд и Вольфовиц показали, что описанное правило является оптимальным в том смысле, что требует минимального (в среднем) объема выборки по сравнению с любым другим решающим правилом, обеспечивающим те же вероятности ошибок первого и второго рода.

ПРОПУЩЕННЫЕ (СТЕРТЫЕ) НАБЛЮДЕНИЯ

наблюдения, у которых отсутствуют значения одной или более переменных.

Данные, содержащие П.н., называют неполными данными или данными с пропусками. Пропуски в массивах данных возникают из-за невозможности сбора информации, несоблюдения правил извлечения выборки, отказа от обследования, неполноты, утраты, искажения или сокрытия информации, цензурирования, исключения резко выделяющихся наблюдений и т.п.

Методы анализа неполных данных основаны на предположениях о распределении пропущенных значений, в особенности от того как связаны законы распределения пропущенных и наблюдаемых значений. В работе Д. Рубина предложена классификация типов пропущенных данных и сформулированы условия, при которых П.н. – игнори-руемы. Пусть задача

состоит в оценке параметров распределения Y, возможно условного от нескольких предикторов X, П.н. содержит лишь переменная Y. Пропуски значений переменной Y называются случайными (англ. – missing at random (MAR)), если распределение пропусков не зависит от пропущенных значений, т.е. если *условная вероятность* P(Y – пропущено | X) не зависит от Y, но может зависеть от переменных X.

Пропуски полностью случайны (англ. – missing completely at random (MCAR)), если условная вероятность P(Y – пропущено | X) одинакова для всех возможных значений полных данных, т.е. не зависит ни от Y, ни от прочих переменных X. При выполнении предположений о случайности или полной случайности П.н. (MCAR или MAR), механизм пропусков несущественен. Если предположения случайности пропусков не выполняются, то П.н. не являются случайными (англ. – missing not a random (MNAR)). В этом случае механизм пропусков является существенным, и для корректного анализа данных этот механизм необходимо знать.

Применяют следующие методы обработки данных, содержащих П.н.: исключение некомплектных объектов, заполнение пропусков перед анализом фактических данных, взвешивание, методы, основанные на модели пропусков. При небольшом числе П.н. часто применяют полное исключение некомплектных объектов из анализа. Достоинства данного подхода состоят в его простоте, сравнимости дескриптивных статистик, т.к. они рассчитаны по одинаковому множеству наблюдений. Однако, при исключении неполных наблюдений происходит значительная потеря информации. Кроме того, если комплектные данные сильно отличаются от выборки в целом, результаты анализа могут быть сильно смещёнными.

При применении метода доступных наблюдений используются все имеющиеся значения. Но, т.к. совокупность наблюдений, по которым производятся расчёты, меняется в зависимости от состава признаков, то возникают трудности при сопоставлении данных. В общем случае данный метод может приводить к неудовлетворительным результатам.

Преимущество предварительного заполнения пропусков – простота представления данных, выполнения вычислений и интерпретации результатов, т.к. после заполнения пропусков используются традиционные методы анализа полных данных.

Существует множество методов заполнения пропущенных значений: замена общим средним, замена аналогом, замена постоянным значением из внешнего источника, замена условным средним, замена с помощью регрессии на присутствующие объекты, случайный выбор подстановки из множества допустимых значений, сплайн-интерполяция, восстановление с помощью алгоритмов *факторного* или *кластерного анализа*, локальные алгоритмы восстановления и др. Недостатки метода восстановления пропущенных значений – зависимость между наблюдениями после заполнения пропусков, т.к. параметры алгоритма восстановления П.н. вычисляются по присутствующим данным, а также отличие распределения данных после заполнения пропусков от истинного.

При многократном заполнении сравнивают результаты, полученные при разных вариантах заполнения П.н. Сопоставление результатов, полученных при многих вариантах подстановок в рамках одной и той же модели пропусков, позволяет строить выводы, отвечающие неопределённости в рамках этой модели. Сравнение результатов анализа подстановок в соответствии с двумя или более моделями позволяет исследовать чувствительность анализа к моделям пропусков, что особенно важно в случае неигнорируемых пропусков.

Процедуры взвешивания. При использовании *выборок расслоенных* среднее рассчиты-вается по формуле средневзешенной арифметической, с весами обратно пропорцио-нальными вероятности попадания наблюдения из *i*-го слоя в выборку:

$$\frac{\sum \pi_i^{-1} y_i}{\sum \pi_i^{-1}}.$$

При наличии П.н. процедура взвешивания может быть модифицирована, что бы учитывать вероятности пропуска:

$$\frac{\sum(\pi_i p_i)^{-1} y_i}{\sum(\pi_i p_i)^{-1}},$$

где p_i – доля полных наблюдений в i-ой подвыборке (оценка вероятности получить наблюдаемое значение наблюдение для единиц в i-ом слое).

Широкий класс процедур составляют методы обработки, основанные на явных или неявных предположениях о механизме порождения П.н. Достаточно общий подход к поиску оценок макс. правдоподобия по неполным данным реализует ЕМ - алгоритм (от англ. expectation maximization), формализующий итеративную идею обработки неполных данных: заполнение пропусков оценками пропущенных значений, оценивание параметров, повторное оценивание пропущенных значений при полученных оценках параметров, повторное оценивание параметров и т.д. до сходимости процесса.

При наличии П.н., особенно, неигнорируемых пропусках, полезным является исследование чувствительности выводов к различным предположениям о механизме их порождения.

ПРОСТРАНСТВЕННАЯ ВЫБОРКА

1. массив данных, содержащий результаты наблюдений для разных единиц совокупности в данный момент времени; 2. вид *выборки*, характеризующий процедуру её формирования: подмножество, состоящее из определён-ного числа единиц, случайно отобранных по двум координатам на местности. Исполь-зуется в археологии, географии, геологии, картографии, экологии.

ПРОЦЕДУРЫ КЛАССИФИКАЦИИ

методы и алгоритмы решения задачи разбиения изучаемой совокупности объектов на отдельные группы, называемые *классами* или кластерами.

Для осуществления П.к., требуется задать способы вычисления *меры близости* или *расстояния между объектами*, *расстояния между классами объектов* и алгоритм классификации. В некоторых случаях использование определённого вида меры близости уже заложено в самом алгоритме. Существование различных П.к. связано с многообрази-ем постановок задач типологизации объектов.

Возможность применения того или иного типа кластер-процедур для решения конкретной задачи зависит от числа наблюдений и наличия априорной информации о числе классов.

Процедуры классификации иерархические используются для классификации относительно небольшого числа наблюдений и не требуют априорной информации о числе классов. Позволяют получить иерархическую структуру разбиений анализируемой совокупности на классы в виде ориентированного дерева (*дендрограммы*), по которому исследователь может судить об отношениях между классами. При необходимости исследователь может, выбрав на основе дендрограммы или априорной информации число классов, получить итоговое разбиение на классы.

Процедуры классификации параллельные применяются для классификации сравнительно небольшого числа наблюдений. Число классов, в зависимости от особенностей алгоритма, задаётся заранее или получается в процессе решения задачи. Эти процедуры в основном нацелены на поиск разбиения, оптимального в смысле выбранного функционала качества, или формируют кластеры по принципу определения мест наибольшего сгущения точек в рассматриваемом факторном пространстве. Как правило, они реализуются с помощью итерационных алгоритмов, которые на каждом шаге используют все имеющиеся наблюдения.

Процедуры классификации последовательные рекомендуются для решения задач *автоматической классификации* при значительном числе наблюдений. Число классов указывается исследователем или, если это предусмотрено алгоритмом, получено в ходе решения задачи. Достоинства *последовательных кластер-процедур* – простота реализации и скорость выполнения, т.к. на каждом шаге используется лишь не-

большое число наблюдений, что позволяет применять последовательные процедуры при больших объёмах выборочной совокупности.

См. также *Иерархические процедуры кластерного анализа*, *Кластерный анализ*, *Параллельные кластер-процедуры*.

ПРОЦЕДУРЫ КЛАССИФИКАЦИИ ИЕРАРХИЧЕСКИЕ

процедуры построения последовательности разбиений исследуемой совокупности объектов на *классы*. Выделяют агломеративные и дивизимные алгоритмы. На первом шаге агломеративных алгоритмов каждое наблюдение рассматривается как отдельный кластер. Дивизимные алгоритмы начинают с разбиения, когда все объекты входят в единственный класс. Далее разбиение получается из предыдущего посредством либо объединения двух и более классов (агломеративные алгоритмы), либо разбиения классов (дивизимные алгоритмы).

Иерархическая структура разбиений на классы изображается в виде ориентированного дерева (*дендрограммы*), по которому можно судить об отношениях между классами. При этом корень дерева представляет заданное множество данных, вершины, отличные от корня – подмножества исходного множества (классы), а ребра, направленные от корня к поддеревьям и от поддеревьев к листьям, связывают включающие классы с включаемыми.

Результат иерархической классификации существенно зависит от способа вычисления *расстояния между классами*. Некоторые популярные агломеративные алгоритмы могут рассматриваться как частные случаи обобщённой агломеративной П.к.и. Обозначим S_a и S_b два подмножества, построенные на шаге *k*, которые на шаге *k+1*объединяются в подмножество S_{ab}, S_c – некоторое третье подмножество, n_a, n_b, n_c – число объектов, входящих в эти подмножества.

Обобщённое расстояние между классами вычисляется по формуле:

$$\rho(S_{ab},S_c)=a_a\rho(S_a,S_c)+a_b\rho(S_b,S_c)+a_a\rho(S_a,S_c)+\beta a_a\rho(S_a,S_b)+\gamma(\rho(S_a,S_c)-\rho(S_b,S_c)),$$

где a_a, a_b, β, γ – константы и $\rho(S_a,S_c), \rho(S_b,S_c), \rho(S_a,S_b)$ – расстояния, вычисленные ранее, на шаге классификации *k*. Важнейшие частные случаи: одиночная связь (*метод «ближайшего соседа»*). Расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах:

$$a_a=\tfrac{1}{2}, a_{b.}=\tfrac{1}{2}, \beta=0, \gamma=-\tfrac{1}{2};$$

полная связь (*метод «дальнего соседа»*). В этом методе расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удалёнными соседями"):

$$a_a=\tfrac{1}{2}, a_{b.}=\tfrac{1}{2}, \beta=0, \gamma=+\tfrac{1}{2};$$

взвешенный *центроидный метод* (медиана). Этот метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них):

$$a_a=\tfrac{1}{2}, a_{b.}=\tfrac{1}{2}, \beta=-\tfrac{1}{4}, \gamma=0;$$

среднее расстояние между всеми парами объектов, взвешенное согласно размеру соответствующих кластеров:

$$a_a=n_a/(n_a+n_b), a_{b.}=n_b/(n_a+n_b), \beta=0, \gamma=0;$$

невзвешенный центроидный метод – расстояние между двумя кластерами определяется как расстояние между их центрами тяжести:

$$a_a=n_a/(n_a+n_b), a_{b.}=n_b/(n_a+n_b), \beta=-a_a a_b, \gamma=0;$$

метод Ворда:

$$a_a = (n_a + n_c) / (n_a + n_b + n_c), a_{b.} = (n_b + n_c) / (n_a + n_b + n_c), \beta = -n_a / (n_a + n_b + n_c), \gamma = 0.$$

Некоторые агломеративные алгоритмы используют способы вычисления расстояний между классами, не сводящиеся к указанной линейной форме.

Практически все алгоритмы иерархической классификации являются эвристическими.

К недостаткам П.к.и. следует отнести высокие вычислительные требования и затраты времени при большом числе объектов. Кроме того, большинство алгоритмов не приводят к разбиению, оптимальному в смысле какого-либо критерия качества. См. также *Автоматическая классификация*, *Кластерный анализ*.

ПРЯМОЕ (КРОНЕКЕРОВО) ПРОИЗВЕДЕНИЕ МАТРИЦ

A, размером $m \times n$, и B, размером $p \times q$ – *блочная матрица* порядка $mp \times nq$:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \dots & \dots & \dots & \dots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix}$$

и обозначается $A \otimes B$.

П.п.м. определено для любой пары матриц.

Свойства кронекерова произведения:

1. $(A \otimes B) \otimes C = A \otimes (B \otimes C)$; 2. $(A+B) \otimes (C+D) = A \otimes C + A \otimes D + B \otimes C + B \otimes D$, если существуют $A+B$ и $C+D$; 3. если существуют AC и BD, то $(A \otimes B)(C \otimes D) = AC \otimes BD$; 4. если α – число, то $\alpha \otimes A = \alpha A = A\alpha = A \otimes \alpha$; 5. если a и b – два вектора-столбца не обязательно одной и той же размерности, то $a^T \otimes b = ba^T = b \otimes a^T$; 6. правило транспонирования кронекерова произведения: $(A \otimes B)^T = A^T \otimes B^T$; 7. Если A и B – квадратные матрицы не обязательно одного и того же размера, то след кронекерова произведения: $tr(A \otimes B) = tr(A) \cdot tr(B)$; 8. если матрицы A и B невырождены, то $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$; 9. ранг кронекерова произведения: $r(A \otimes B) = r(A)r(B)$; 10. для двух матриц: A размерности $n \times n$ и B размерности $m \times m$, определитель: $\det(A \otimes B) = |A \otimes B| = |A|^m |B|^n$; 11. если матрицы A и B положительно (полу)определены, то матрица $A \otimes B$ положительно (полу)определена.

Р

РАНГ МАТРИЦЫ А

наивысший порядок её миноров, отличных от нуля. Р.м. называют также макс. число её линейно независимых строк или столбцов. Матрицей A с элементами a_{ij} (i = 1, 2, …, n; j = 1, 2, …, m), содержащая n строк и m столбцов, т.е. размера n x m называется табл., обозначаемая и изображаемая в виде:

$A=(a_{ij})=$	a_{11}	·	·	·	a_{1j}	·	·	·	a_{1m}
	·				·				·
	·				·				·
	·				·				·
	a_{i1}	·	·	·	a_{ij}	·	·	·	a_{im}
	·				·				·
	·				·				·
	·				·				·
	A_{n1}	·	·	·	a_{mj}	·	·	·	A_{nm}

Подмножество элементов a_{ij}, стоящих на пересечении к любых строк и столбцов матрицы A образуют квадратную подматрицу B размера к x к или квадратную матрицу к-го порядка. Определителем (минором) к-го порядка называется скаляр, обозначаемый как $|M_к|$. Этот минор можно вычислить с помощью рекуррентной формулы:

$$|M_к| = a_{11} \cdot (-1)^{1+1} \cdot M_{11} + \cdots + a_{1j} \cdot (-1)^{1+j} \cdot M_{1j} + \cdots + a_{1n} \cdot (-1)^{1+n} \cdot M_{1n},$$

где M_{1j} – минор (к-1)-го порядка матрицы, получаемой вычеркиванием из матрицы к-го порядка строки и столбца, содержащих элемент a_{1j}.

Для получения Р.м. используется метод последовательного исключения, в результате которого исходная матрица преобразуется в матрицу, содержащую единичную подматрицу r-го порядка, равного Р.м. А.

РАСПОЗНАВАНИЕ ОБРАЗОВ

раздел математической кибернетики, разрабатывающий методы классификации и идентификации объектов и решения на этой основе задач *прогнозирования*. Объекты описываются конечным набором признаков. Описание объекта представляет собой *n*-мерный вектор, где *n* – число признаков, а *i*-я координата вектора равна значению *i*-го признака. Для некоторых объектов может быть определена принадлежность к тому или иному классу, и из таких объектов можно формировать обучающие или контрольные выборки.

Классификация – важнейшая составляющая проблемы Р.о. Она может производиться в ситуации, когда число классов и их смысл известны заранее, или в отсутствие образов классов, когда их число и смысл выявляются в процессе классификации. Использование обучающих выборок в процессе разделения объектов на классы обусловливает решение задачи классификации с обучением, в противном случае классификация производится без обучения.

Задачи Р.о. помимо классификации – минимизация описания исходных объектов и выделение миним. числа наиболее информативных признаков, позволяющих достичь заданного качества классификации объектов. Важную роль в решении задач Р.о. играют методы *математической статистики*. Напр., *кластерный, факторный* и компонентный *анализы* применяют для классификации *ген. совокупностей* на основе выборок из них. Задачи Р.о. – весьма трудны и исследованы лишь в отдельных частных случаях. В Р.о. используются идеи и результаты многих научных направлений. Прикладные задачи Р.о. (прогнозирование, классификация социологических материалов, идентификация тестов и т.д.) часто решаются с помощью т.н. эвристических алгоритмов и моделей распознавания. Среди них известны нейросетевые модели, модели, построенные на принципе потенциалов; модели голосования.

Р.о. используется в исследованиях структуры социально-экономических явлений, напр., при выявлении типологии потребителей и прогнозе структуре потребления.

РАССТОЯНИЕ МАХАЛАНОБИСА

расстояние между объектами, используемое в задачах многомерной классификации. В случае взаимозависимости компонент $x_1, x_2, \ldots, x_k$ вектора наблюдений X и их различной значимости в решении вопроса классификации обычно используют обобщённое (взвешенное) Р.М. между объектами X_i и X_j, где $i, j = 1,2,\ldots,n$, определяемое как:

$$\rho_0(X_i, X_j) = \sqrt{(X_i - X_j)^T \Lambda^T \Sigma^{-1} \Lambda (X_i - X_j)},$$

где Σ – *ковариационная матрица* ген. совокупности X, из которой извлекаются наблюдения; $X_i = (x_{i1}, x_{i2}, x_{ik})^T$ – вектор значений показателей для *i*-го наблюдения;

Λ – некоторая симметрическая неотрицательно-определённая матрица «весовых» коэффициентов, которая чаще всего выбирается диагональной.

Если компоненты $x_1, x_2, \ldots, x_k$ одинаково значимы для классификации, то матрица Λ становится единичной матрицей и имеет место Р.М.

$$\rho_M(X_i, X_j) = \sqrt{(X_i - X_j)^T \quad \Sigma^{-1} \quad (X_i - X_j)}$$

Р.М. является естественным в случае нормальности распределения вектора *X*.

Евклидово расстояние – частный случай Р.М., в случае взаимонезависимости компонент $x_1, x_2, \ldots, x_n$ вектора наблюдений *X*, а также равенства их дисперсий.

В статистической практике Р.М. используют в процедурах классификации и при отклонении

распределения наблюдений внутри классов от нормального, с заменой теоретических характеристик вектора средних значений и *ковариационной матрицы* их оценками, построенными по наблюдениям.

РАССТОЯНИЕ МЕЖДУ ОБЪЕКТАМИ

мера различия объектов, задаваемых количественными или качественными признаками.

Для определения однородности объектов O_i и O_j необходимо задать правила расчёта расстояния между ними $d(O_i,O_j)$, либо степень близости (сходства) $\rho(O_i,O_j)$. Объекты можно считать однородными, если они близки в смысле метрики $d(O_i,O_j)$, и для них производится сравнение $d(O_i,O_j)$ с некоторым пороговым значением в каждом конкретном случае.

При использовании метрики расстояния и степени сходства $\rho(O_i,O_j)$ необходимо выполнение следующих условий: симметрия: $\rho(O_i,O_j)=\rho(O_j,O_i)$; $d(O_i,O_j)=d(O_j,O_i)$; макс. сходство объекта с самим собой: $\rho(O_i,O_j)=\max\rho(O_i,O_j)$, $d(O_i,O_i)=0$; монотонное убывание $\rho(O_i,O_j)$ по $d(O_i,O_j)$: из $d(O_k,O_l)\geq d(O_i,O_j)$ следует $\rho(O_k,O_l)\leq\rho(O_i,O_j)$.

Выбор метрики или меры близости – важный этап анализа, от которого зависит окончательный вариант разбиения совокупности объектов на классы. В каждом конкретном случае это зависит от целей статистического исследования, физического и статистического вида многомерной совокупности, характера вероятностного распределения и т.п. В основном применяются следующие *расстояния* между объектами: Евклида, Каллбэка, *Махаланобиса* и Хемминга. Кроме этого, существуют различные меры близости для дихотомических признаков, меры, задаваемые с помощью потенциальной функции и др.

См. также *Однородность объектов, Евклидово расстояние, Информационное расстояние Каллбэка.*

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ

мера различия между группами (*классами*, кластерами) объектов.

Наиболее распространённые меры расстояния между классами – Р.м.к.о.: *«ближнего соседа», «дальнего соседа», информационное расстояние Каллбэка, обобщённое* (по Колмогорову), *«средний связи», «центров тяжести».*

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ «БЛИЖНЕГО СОСЕДА»

определяется

$$\rho_{\min}(S_l,S_m)=\min_{X_i\in S_l,X_i\in S_m}d(X_i,X_j),$$

где S_l и S_m – соответственно l и m – группа (*класс*, кластер) объектов; вектор X_i и X_j – вектор наблюдений соответственно групп S_l и S_m.

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ «ДАЛЬНЕГО СОСЕДА»

определяется:

$$\rho_{\max}(S_l,S_m)=\max_{X_i\in S_l,X_i\in S_m}d(X_i,X_j),$$

где S_l и S_m – соответственно l и m – группа (*класс*, кластер) объектов; вектор X_i и X_j – вектор наблюдений соответственно групп S_l и S_m.

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ ИНФОРМАЦИОННОЕ

см. в ст. Информационное расстояние Каллбэка.

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ ОБОБЩЁННОЕ (ПО КОЛМОГОРОВУ)

вычисляется по формуле:

$$\rho_\nu(S_l,S_m)=\left(\frac{1}{n_l n_m}\sum_{X_i\in S_l}\sum_{X_i\in S_m} d^\nu(X_i,X_j)\right)^{\frac{1}{\nu}},$$

где S_l и S_m – соответственно l и m – группа (*класс*, кластер) объектов; n_l и n_m – соответственно число объектов групп S_l и S_m, вектор X_i и X_j – вектор наблюдений соответственно групп S_l и S_m; ν – степень среднего.

Если $\nu\to\pm\infty$ имеем $\rho_\infty(S_l,S_m)=\rho_{\max}(S_l,S_m)$ – *расстояние между классами объектов «дальнего соседа»* и $\rho_{-\infty}(S_l,S_m)=\rho_{\min}(S_l,S_m)$ – расстояние между классами объектов «дальнего соседа». Если $\nu=1$, то $\rho_1(S_l,S_m)=\rho_{cp}(S_l,S_m)$ – *расстояние между классами объектов «средней связи».*

Если $\rho_\nu(S_m,S_q)=S_m\cup S_q$ – группа, полученная объединением кластеров S_m и S_q, то расстояние Колмогорова находится:

$$\rho_\nu(S_l,S(m,q))=\left(\frac{n_m(\rho_\nu(S_l,S_m))^\nu+n_q(\rho_\nu(S_l,S_q))^\nu}{n_q+n_m}\right)^{\frac{1}{\nu}}.$$

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ «СРЕДНЕЙ СВЯЗИ»

среднее арифметическое всевозможных попарных расстояний между элементами групп:

$$\rho_{cp}(S_l,S_m)=\frac{1}{n_l n_m}\sum_{X_i\in S_l}\sum_{X_i\in S_m} d(X_i,X_j),$$

где S_l и S_m – соответственно l и m – группа (*класс*, кластер) объектов; n_l и n_m – соответственно число объектов групп S_l и S_m, вектор X_i и X_j – вектор наблюдений соответственно групп S_l и S_m.

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ «ЦЕНТРОВ ТЯЖЕСТИ»

вычисляется по формуле: $\rho(S_l,S_m)=d(\overline{X}(l),\overline{X}(m))$, где S_l – *l-класс* (группа, кластер) объектов; вектора $\overline{X}(l),\overline{X}(m)$ – центры тяжести (средние арифметические) векторных наблюдений соответственно групп S_l и S_m, $\rho(S_l,S_m)$.

РАССТОЯНИЕ ХЕММИНГА

см. в ст. *Хеммингово расстояние*

РАСЩЕПЛЕНИЕ СМЕСИ РАСПРЕДЕЛЕНИЙ

метод автоматической классификации, применяемый для содержательного статистического анализа структуры экономических и социальных объектов и для обеспечения возможности разбиения наблюдаемых объектов на однородные группы. Обычно доля q_i $(j=1,2,\ldots,k)$ объектов каждой из k однородных групп и принадлежность к ним объектов неизвестны. Требуется на основе n наблюдений оценить параметры q_j и θ_j смеси

$$f(x)=\sum_{j=1}^{k} q_j f_j(x;\theta_j),$$

где $f_i(x;\theta_i)$ – плотность вероятности признака x в j-й однородной группе. Для технической реализуемости Р.с.р. требуется одномодальность каждой из функций $f_i(x;\theta_i)$. Вследствие того, что

$$\sum_{j=1}^{k} q_j=1,$$

число оцениваемых независимых параметров снижается на единицу, но объём выборки должен быть как минимум больше этого числа, но желательно, чтобы на каждый оцениваемый параметр приходилось хотя бы три-пять наблюдений. Оценку $\widehat{\theta}$ вектора параметров модели $\theta=(q_1,\ldots,q_k;\theta_1^T,\ldots,\theta_k^T)^T$ целесообразно производить методом максимального правдоподобия путём максимизации логарифма функции правдоподобия

$$\ln l(\widehat{\theta}_{onm})=\max_\theta\sum_{i=1}^{n}\ln f(x_i;\theta),$$

реализуемой, как правило, численными методами. Следует помнить о чувствительности данной параметрической процедуры к наличию искажений и выбросов в исходной совокупности. Поэтому требуется предварительный анализ исходных данных. Кроме того, выбора вида функции плотности вероятности каждой однородной группы объектов $f_i(x;\theta_i)$ должен быть обоснован как теоретически, так и на основе анализа распределения результатов наблюдений. Корректность и качество полученной модели определяются путём проверки соответствия полученного теоретического распределения эмпирическому. После получения оценок встает вопрос классификации объектов. Для однозначного отнесения произвольного объекта к одному из выделенных классов можно использовать байесовский подход, использующий в качестве критерия минимум среднего риска ошибочной классификации. При простой функции стоимости ошибок классификации и отсутствии априорных предпочтений границы страт определяются абсциссами точек пересечения соседних компонентов смеси. В окрестностях этих точек вероятность ошибочной классификации близка к ½. В случае существенного наложения компонент незначительные отличия в значении признака x могут приводить к скачкообразному переходу объекта из одной группы в другую. Для исключения этого эффекта целесообразно использовать алгоритмы мягкой классификации, использующие аппарат *нечётких множеств*.

РЕДУЦИРОВАННАЯ МАТРИЦА

преобразованная матрица с целью сокращения числа переменных в *факторном анализе*. В факторном анализе используется понятие «редуцированная *корреляционная матрица*» – корреляционная матрица, на гл. диагонали которой вместо единиц стоят значения общностей:

$$R^+ = \begin{pmatrix} h_1^2 & r_{12} & \cdots & r_{1n} \\ r_{21} & h_2^2 & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & h_n^2 \end{pmatrix}, \; i,j = \overline{1,n},$$

где r_{ij} – выборочные коэффициенты корреляции, h_i^2 – значения общностей, которые представляют собой сумму квадратов факторных нагрузок.

В модели компонентного анализа ранг корреляционной матрицы равен n, а параметры описаны в терминах не менее чем n общих факторов. В задаче описания n параметров через меньшее, чем n число общих факторов вводится факторное отображение

$$x_i = \sum_{j=1}^{m} q_{ij} f_i + e_i, \quad i = \overline{1,n},$$

(x_i – исходные признаки, q_{ij} – факторные нагрузки; f_i – ненаблюдаемые факторы; e_i – случайная ошибка), с помощью которого может быть вычислена редуцированная корреляционная матрица. Ранг Р.м. m меньше её порядка n. Число общих факторов в факторном отображении равно рангу редуцированной корреляционной матрицы, которое и есть наименьшее число факторов, адекватно описывающие корреляции между параметрами. В геометрической интерпретации это означает, что наименьшим пространством общих факторов, в котором лежат n точек, является m-мерное пространство. Прежде чем определить фактор, нужно построить редуцированную корреляционную матрицу R^+ по значениям общностей. Оценка общностей – проблема общности – первая проблема факторного анализа; вторая проблема – определение фактора – проблема факторов. В методе гл. факторов методика анализа *гл. компонент* используется применительно к редуцированной корреляционной матрице R^+. Для оценивания общностей используют квадрат множественного коэффициента корреляции между соответствующей переменной и совокупностью остальных переменных или наибольший по абсолютной величине коэффициент корреляции в соответствующей переменной строке корреляционной матрицы. После размещения оценок общностей на гл. диагонали корреляционной матрицы проводится факторный анализ, решая характеристическое уравнение: $\det(R^+ - \lambda I) = 0$, где R^+ – редуцированная корреляционная, определяют собственные числа λ_i и матрицу нормированных (характеристических) векто-

ров, а затем находят матрицу факторного отображения. Для получения оценок общностей и факторных нагрузок используется эмпирический итеративный алгоритм, который сходится к истинным оценкам параметров.

РЕЗКО ВЫДЕЛЯЮЩИЕСЯ НАБЛЮДЕНИЯ

нетипичные значения, выбросы, которые существенно отклоняются от центра распределения остальных выборочных данных. Причины выбросов – обычные случайные колебания выборки, обусловленные природой анализируемой ген. совокупности либо искажения стандартных условий сбора статистических данных, ошибки измерения. В случае ошибок измерения «подозрительные» наблюдения исключают из дальнейшего рассмотрения. Один из способов решения вопроса об исключении Р.в.н – формальные (статистическим) методы, основанные на предположении однородности данных. При этом выбросы рассматриваются как наблюдения, нетипично далеко удаляющиеся от центра распределения. Общая логическая схема этих методов: вводится предположение о природе анализируемой совокупности данных в виде функции $\psi(X_{i1}^*, X_{i2}^*, \ldots, X_{ik}^*; X)$ от всех имеющихся наблюдений X, характеризующей степень аномальности (меру удаленности от основной массы наблюдений) выбросов $X_{i1}^*, X_{i2}^*, \ldots, X_{ik}^*$. Путём подстановки в функцию ψ реальных значений наблюдений и сравнивании этих величин с некоторым пороговым значением ψ_0, «подозрительные» наблюдения полностью исключаются или их вклад уменьшается с помощью весовой функции. К методам выделения выбросов также относят аналитические процедуры для идентификации выбросов и оценки значимости их отклонения (напр., метод исключения одного экстремального наблюдения или одновременного исключения нескольких экстремальных наблюдений). Осн. трудность в использовании аналитических методов состоит в том, что реальная доля «засорения» не известна, а оценивается по тем же данным, по которым проверяется значимость отклонения. Наиболее устойчивыми к отклонениям от предположения нормальности осн. части выборки являются графические процедуры.

При исследовании временного ряда выбросы, вместо исключения, можно моделировать с помощью фиктивных переменных, соответствующих фиксированным моментам времени. В случае кратковременного отклонения временного ряда вводят фиктивную переменную $\theta_t^{t^*} = (0,\ldots,0,1,0,\ldots,0)$, которая равна нулю всегда, кроме момента $t = t^*$, соответствующего выбросу. Если же в исследуемом явлении произошел структурный сдвиг, вызвавший скачок в динамике ряда, то фиктивная переменная будет иметь вид: $\delta_t^{t^*} = (0,\ldots,0,1,\ldots,1)$. Эта переменная равна нулю до некоторого фиксированного момента t^*, а после этого момента становится равной единице.

РЕШАЮЩЕЕ ПРАВИЛО

функция δ(x), на основе которой решается общая задача построения *оптимальных (байесовских) процедур классификации*. Р.п. (процедура классификации, дискриминантная функции) δ(x) принимает целые положительные значения 1,2, …, k, причем те X, при которых она принимает значение, равное j, относят к классу j т.е. $S_j = \{X: \delta(x) = j\}$, $j = \overline{1,k}$, где $X_1, X_2, \ldots, X_n$ – классифицируемые наблюдения при наличии обучающих выборок, интерпретируемые как выборка из ген. совокупности, описываемой смесью k классов (одномодальных ген. совокупностей) с плотностью вероятности $f(X) = \sum_{j=1}^{k} \pi_j f_j(X)$, π_j – априорная вероятность появления в выборке элемента из класса (ген. совокупности) j с плотностью $f_j(X)$, S_j – это p-мерные области в пространстве П(X) возможных значений анализируемого многомерного признака X. Р.п.о – функция δ(x) – задается разбиением $S = (S_1, S_2, \ldots, S_k)$ всего пространства П(X) на k непересекающихся областей, таким образом, чтобы сумма S_j: $S_1 + S_2 + \ldots + S_k$ заполняло все пространство П(X) и чтобы S_j попарно не пересекались.

Процедура классификации δ(x) называется оптимальной (байесовской), если она сопровождается миним. потерями

$$C = \sum_{i=1}^{k} \pi_i \sum_{j=1}^{k} c(j \mid i) P(j \mid i)$$

среди всех других процедур классификации, *c(j|i)* – величина потерь при отнесении одного объекта i-го класса к классу j (при i=j c_{ij} = 0), *P(j|i)* – вероятность отнести объект класса i к классу j.

Процедура классификации $S^{(опт)}$ = ($S_1^{(опт)}$, $S_2^{(опт)}$,…., $S_k^{(опт)}$), при которой потери *C* будут оптимальными, определяется как:

$$S_j^{опт} = \left\{ X : \sum_{i=1, i \neq j}^{k} \pi_i f_i(X) c(j \mid i) = \min_{1 \le l \le k} \sum_{i=1, i \neq j}^{k} \pi_i f_i(X)\, c(l \mid i) \right\}.$$

Т.е. наблюдение X_v ($v = \overline{1, n}$) будет отнесено к классу j тогда, когда средние удельные потери от его отнесения в этот класс окажутся миним. по сравнению с аналогичными потерями, связанными с отнесением этого наблюдения в любой другой класс. В случае равных потерь *c(j|i)* (т.е. когда *c(j|i)* = c_0 = const) правило классификации упрощается: наблюдение X_v будет отнесено к классу j, когда

$$\pi_i f_i(X_v) = \max_{1 \le l \le k} \pi_i f_i(X_v).$$

Однако процедура $S_j^{(опт)}$ – это теоретическое оптимальное правило классификации, для его реализации необходимо знание априорных вероятностей π_i и законов распределения $f_j(X)$. В статистическом варианте решения этой задачи априорные вероятности и законы распределения заменяются соответствующими оценками, построенными на базе обучающих выборок.

С

СЖАТИЕ МАССИВОВ ИНФОРМАЦИИ

построение экономной системы вспомогательных признаков, обладающих наивысшей автоинформативностью. С.м.и. – одна из типовых прикладных задач снижения размерности анализируемого признакового пространства наряду с отбором наиболее информативных показателей, включая выявление латентных факторов, наглядным представлением данных, построением условных координатных осей (*многомерным шкалированием*, *латентно-структурным анализом*). При решении задач С.м.и. больших используется сочетание методов классификации и снижения размерности. Методы классификации позволяют перейти от массива, содержащего информацию по всем n статистически обследованным объектам, к соответствующей информации только по k эталонным образцам (k<<n), где в качестве эталонных образцов берутся специальным образом отобранные наиболее типичные представители классов, полученных в результате операции разбиения исходного множества объектов на однородные группы. Методы снижения размерности позволяют заменить исходную систему показателей $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})^T$ набором вспомогательных (наиболее автоинформативных) переменных $Z(X) = (z^{(1)}(X), z^{(2)}(X), \ldots, z^{(p/)}(X))^T$. Т.о., размерность информационного массива понижается от p·n до $p^{/}$·k. При этом новые (вспомогательные) признаки $z^{(1)}(X)$, $z^{(2)}(X)$, …, $z^{(p/)}(X)$ могут выбираться из числа исходных или определяться по какому-либо правилу по совокупности исходных признаков, например, как их линейные комбинации. При формировании новой системы для признаков вводят различные требования, такие, как наибольшая информативность, взаимная некоррелированность, наименьшее искажение геометрической структуры множества исходных данных и т.п. В зависимости от варианта формальной конкретизации этих требований получают различные виды алгоритмов снижения размерности. Выделяют три осн. типа принципиальных предпосылок, обусловливающих возможность перехода от большого числа p исходных показателей состояния (поведения, эффективности функционирования) анализируемой системы к существенно меньшему числу $p^{/}$ наиболее информативных переменных. Во-первых, это дублирование информации, доставляемой сильно взаимосвязанными признаками; во-вторых, неинформативность признаков, мало меняющихся при переходе от одного объекта к другому (малая «вариабельность» признаков); в-третьих, возможность агрегирования, т. е. простого или

«взвешенного» суммирования, по некоторым признакам.

Весь процесс решения задач классификации и снижения размерности разбивается на несколько этапов: предметно-содержательное определение целей исследования; определение типа прикладной задачи в терминах теории классификации и снижения размерности; составление плана сбора исходной информации, его реализация и предварительный анализ исходной информации; выбор базовой математической модели механизма генерации исходных данных; применение методов статистической обработки исходных данных, нацеленных на выявление их вероятностной и геометрической природы; уточнение выбора базовой математической модели; реализация на ЭВМ уточненного плана математико-статистического анализа данных; подведение итогов исследования и интерпретация полученных результатов.

СИММЕТРИЧНАЯ МАТРИЦА

квадратная матрица A = $\{a_{ij}\}, i=\overline{1,n},\ j=\overline{1,n},$ совпадающая со своей транспонированной матрицей $A^T=\{a_{ji}\}$ для «j,i, т.е. $A^T = A$. В симметричной матрице элементы, симметрично расположенные относительно главной диагонали, равны. Примеры С.м.:

$$A=\begin{pmatrix} 3 & 4 & -1 \\ 4 & 2 & 6 \\ -1 & 6 & 1 \end{pmatrix},\quad B=\begin{pmatrix} 6 & 8 \\ 8 & 7 \end{pmatrix}.$$

Частным случаем С.м. является диагональная матрица:

$$D=\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}.$$

Всякая *ковариационная матрица* Σ = { σ_{ij} } и *корреляционная матрица* R = { r_{ij} } являются симметричными.

Свойства С.м.: 1. для любой матрицы A матрица A^TA – симметричная; 2. С.м. A размерностью n x n имеет n собственных чисел (некоторые из них могут совпадать), которым соответствуют n собственных векторов $c_1, c_2, \ldots, c_n$, которые могут быть выбраны попарно ортогональными. Собственные векторы, соответствующие разным собственным значениям С.м., являются ортогональными; 3. С.м. может быть приведена к диагональному виду при помощи ортогонального преобразования X: $X^TAX=D$, где на диагонали матрицы D стоят собственные числа матрицы A.

СЛУЧАЙНАЯ (СТОХАСТИЧЕСКАЯ) МАТРИЦА

матрица размерности *n*x*n*

$$P=\{p_{ij}\}=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix},$$

обладающая следующими свойствами: все элементы матрицы p_{ij} неотрицательны: $p_{ij} \geq 0$ для « $i,\ j=\overline{1,n};$ суммы элементов любой строки матрицы равны единице:

$$\sum_{j=1}^{n} p_{ij} = 1 \qquad \text{для } \forall\ i=\overline{1,n}.$$

Пример С.м.:

$$P=\begin{pmatrix} 0{,}1 & 0{,}3 & 0{,}6 \\ 0{,}7 & 0{,}1 & 0{,}2 \\ 0{,}2 & 0{,}6 & 0{,}2 \end{pmatrix}.$$

Любая стохастическая матрица может служить матрицей переходных вероятностей однородной цепи Маркова, где под переходной вероятностью p_{ij} понимают условную вероятность того, что система S после k-го шага окажется в состоянии s_j, при условии, что непосредственно перед этим она находилась в состоянии s_i.

Свойства С.м.: является частным видом неотрицательной матрицы; для каждой однородной цепи Маркова матрица переходных вероятностей является стохастической и, наоборот, любая С.м. матрица может быть рассмотрена как матрица переходных вероятностей некоторой однородной цепи Маркова; С.м. имеет характеристическое число 1 с положительным собственным вектором z = (1, 1, …, 1). Верно и обратное: всякая матрица P > O, имеющая соб-

ственный вектор (1, 1, …,1) при характеристическом числе 1, является стохастической. При этом единица является максимальным характеристическим числом стохастической матрицы; неотрицательная матрица A≥0, имеющая положительное максимальное характеристическое число r>0 и соответствующий этому числу положительный собственный вектор z = (z_1, z_2,. . ., z_n)>0, всегда подобна произведению числа r на некоторую стохастическую матрицу: A = ZrPZ^{-1}. Здесь Z = {z_1, z_2,. . ., z_n}>0 – диагональная матрица, матрица P определена как: $P = \frac{1}{r} Z^{-1} A Z$. 5. Характеристическому числу 1 С.м. всегда соответствуют только элементарные делители первой степени.

СМЕСЬ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ

закон распределения вероятностей признака, характеризующего совокупность, содержащую элементы из различных однородных групп. Каждая из однородных групп имеет свой закон распределения, а доля группы в общей совокупности соответствует вероятности случайного извлечения объекта данной группы из этой совокупности. Модель в виде С.р.в. адекватно описывает социально-экономические системы, которые включают в себя элементы, принадлежащие, как правило, нескольким классам. Примерами таких совокупностей могут служить смеси распределений вероятностей значений параметров одноимённой продукции, выпускающейся различными предприятиями по единой технологии, но с отличающимися значениями параметров технологических систем, сырья и материалов, распределение работников определённой сферы деятельности по уровню заработной платы, распределение стоимости одного квадратного метра объектов недвижимости. Часто заранее неизвестно, какому классу принадлежит наблюдаемый элемент системы. Отдельный j-й класс объектов можно охарактеризовать некоторой функцией плотности вероятности $f_j(x;\theta_j)$ с набором параметров θ_j, описывающей закон распределения анализируемого признака x, в общем случае векторного. Распределение признака во всей совокупности из k однородных групп представляет собой С.р.в.

$$f(x) = \sum_{j=1}^{k} q_j f_j(x;\theta_j),$$

где q_j – доля объектов j-го класса в совокупности,

$$\sum_{j=1}^{k} q_j = 1.$$

Величину q_j можно рассматривать как дискретную смешивающую функцию при конечном или счётном числе компонентов, или как непрерывную q_w, если множество компонентов $\{w\}$ непрерывно. В последнем случае С.р.в. представляет собой интеграл

$$f(x) = \int_{\{w\}} q_w f_w(x;\theta_w) dw.$$

При конечном числе классов и наличии объектов с заранее известной их принадлежностью к отдельным классам (обучающих выборок) классификация объектов может быть реализована на основе процедур *дискриминантного анализа*. В противном случае для этого требуется решение задачи *расщепления смеси вероятностных распределений*.

СМЕСЬ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ ДВУХ НОРМАЛЬНЫХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ

закон распределения вероятностей признака, характеризующего совокупность, содержащую элементы из двух однородных групп, каждая из которых характеризуется нормальным законом распределения по данному признаку. Модель

$$f(x) = q_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + (1-q_1) \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$

используется для описания социально-экономических систем, состоящих из элементов двух нормальных совокупностей, отличающихся значениями параметров – ген. средней μ и дис-

персии σ^2, а также их долями в общей совокупности соответственно q_1 и $(1-q_1)$. Примером может служить продукция двух пр-тий, выпускаемая под единым брендом, но отличающаяся как по средней величине контролируемого параметра, так и по разбросу его значений относительно этого среднего значения. Если объём выпуска на первом пр-тии в три раза больше, чем на втором, весовые коэффициенты компонентов смеси будут равны соответственно 0,75 и 0,25. Если требуется по значениям оцениваемого параметра определить, к какой из групп относится каждое из наблюдений (какое пр-тие выпустило тот или иной образец продукции), то требуется решить задачу классификации в отсутствие или при наличии обучающих выборок в зависимости от того неизвестна или известна принадлежность хотя бы части наблюдений к тому или иному классу. При наличии обучающих выборок классификация может быть реализована как процедура линейного дискриминантного анализа, а в их отсутствие – на основе *расщепления смеси распределений*, реализация которой состоит в оценивании неизвестных параметров по имеющимся результатам наблюдений.

СМЕСЬ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ МНОГОМЕРНЫХ НОРМАЛЬНЫХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ

многомерный закон распределения вероятностей векторного признака $(x_1, x_2, \ldots, x_m)^T = x$, характеризующего совокупность, содержащую элементы из k различных групп, каждая из которых может быть описана многомерным нормальным законом распределения вероятностей

$$\frac{1}{(2\pi)^{m/2}\sqrt{\det\Sigma}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

с вектором ген. средних $\mu = (\mu_1, \mu_2, \ldots, \mu_m)^T$ и ковариационной матрицей Σ. Доля j-й группы q_j в общей совокупности соответствует вероятности случайного извлечения объекта данной группы из этой совокупности. Модель в виде С.р.в.м.н.з.р.

$$f(x) = \sum_{j=1}^{k} q_j \frac{1}{(2\pi)^{m/2}\sqrt{\det\Sigma_j}} e^{-\frac{1}{2}(x-\mu_j)^T\Sigma^{-1}(x-\mu_j)}$$

позволяет описывать совокупности, включающие в себя элементы, принадлежащие, как правило, нескольким классам по признакам, являющимся результатами аддитивного действия большого числа примерно равноценных факторов. Пример такой совокупности – совокупность домохоз-в со сравнительно небольшим числом типов потребительского поведения, в пределах каждого из которых различия в структуре потребления носят случайный характер и незначительны по сравнению с потребительским поведением домохоз-в, представляющих разные типы. Ни число типов, ни значения параметров каждого из них, заранее не известны. Поэтому требуется оценить эти величины на основе результатов наблюдений и предположений о структуре исследуемой совокупности. Оценка параметров при определённом числе однородных групп представляет собой задачу *расщепления смеси вероятностных распределений*. Решив эту задачу для различного числа классов и сравнив полученные результаты с точки зрения их соответствия теории и эмпирическим данным, можно выбрать лучшую из полученных структур и на её основе произвести классификацию объектов.

СНИЖЕНИЕ РАЗМЕРНОСТИ ПРИЗНАКОВОГО ПРОСТРАНСТВА

переход от исходного набора показателей к небольшому числу вспомогательных переменных, по которым можно достаточно точно воспроизвести осн. свойства анализируемого массива данных. Снижение размерности обычно применяется для построения пространства, более удобного для решения задач классификации и исследования причинных связей, чем исходный набор переменных. Осн. предпосылки С.р.п.п. – возможное дублирование информации, содержащейся в тесно взаимосвязанных показателях, низкая информативность показателей, обладающих малой вариабельностью, т.е. мало изменяющихся при переходе от одного объекта к другому, и возможность агрегирования некото-

рых однотипных показателей. Среди методов, используемых для С.р.п.п. выделяется *метод гл. компонент*, максимизирующий критерий информативности, определяемый суммарной дисперсией заданного небольшого числа нормированных вспомогательных переменных. Для вычисления k-й гл. компоненты $z^{(k)}(X)$ (k = 1,..., p) следует найти собственный вектор $l_k = (l_{k1}, ..., l_{kp})^T$ ковариационной матрицы Σ исходного набора показателей в виде матрицы $X = (x^{(1)}, ..., x^{(p)})$, т.е. решить систему уравнений $(\Sigma - \lambda_k I) l_k^T = \underline{0}$, где $\underline{0}$ – p-мерный вектор-столбец из нулей, λ_k – k-й по величине в порядке убывания корень характеристического уравнения $|\Sigma - \lambda I| = 0$. Компоненты l_{kj} (j = 1,…, p) собственного вектора l_k – искомые весовые коэффициенты, с помощью которых осуществляется переход от исходных показателей $x^{(1)}, ..., x^{(p)}$ к гл. компоненте $z^{(k)}(X)=X^T l_k$. Если ограничиться первыми m гл. компонентами (m<p), то получим пространство меньшей размерности, обладающее, наряду с некоррелированностью его координат, свойствами наименьшего искажения некоторых геометрических характеристик при проецировании в него совокупности исходных многомерных наблюдений. В результате снижения размерности решаются задачи упрощения статистических моделей, наглядного представления многомерных данных путём их визуализации, устранения дублирования информации, содержащихся в признаках за счёт перехода к ортогональной системе координат, сжатия объёмов статистической информации. Гл. недостаток метода гл. компонент – трудность интерпретации новых признаков. Более общий подход к С.р.п.п. связан с использованием общей модели факторного анализа: $X = QF + U$, где Q – прямоугольная матрица коэффициентов линейного преобразования, связывающие исходные признаки с непосредственно не наблюдаемыми общими факторами F, а вектор U определяет часть исследуемого признака, которая не может быть объяснена общими факторами. Линейное преобразование может быть реализовано различными методами, в т. ч. и методом гл. компонент. Однако, если требуется обеспечить интерпретируемость факторов, возможно их независимое вращение в векторном пространстве, приводящее к определенному отходу от ортогональности. Кроме того, для С.р.п.п. могут применяться методы построения интегральных показателей по заданным значениям частных характеристик и различные эвристические методы.

СОБСТВЕННОЕ (ХАРАКТЕРИСТИЧЕСКОЕ) ЗНАЧЕНИЕ (ЧИСЛО) МАТРИЦЫ

некоторое число λ, которое является одним из корней характеристического уравнения: $\det(A - \lambda I_n) = 0$. Здесь A – некоторая матрица размерности *n*x*n*, I_n – единичная матрица размерности *n*x*n*, определитель

$$\det(A - \lambda I_n) = \det \begin{pmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{pmatrix}$$

является алгебраическим полиномом n-й степени относительно λ. Этот полином называется характеристическим многочленом матрицы A и имеет не более n вещественных корней. Характеристическое уравнение $\det(A - \lambda I_n) = 0$ – это алгебраическое уравнение n-й степени относительно λ. Само уравнение и его корни $\lambda_1, \lambda_2, \ldots, \lambda_n$ по построению полностью определяются элементами матрицы A.

Пример нахождения собственных значений. Для матрицы

$$A = \begin{pmatrix} 1 & 3 \\ 12 & 1 \end{pmatrix}$$

составляется характеристическое уравнение:

$$\det(A - \lambda I_n) = \begin{vmatrix} 1 - \lambda & 3 \\ 12 & 1 - \lambda \end{vmatrix} = (1 - \lambda)^2 - 36 = \lambda^2 - 2\lambda - 35 = 0,$$

из которого определяются собственные значения: $\lambda_1 = -5$, $\lambda_2 = 7$.

Некоторые свойства собственных (характеристических) значений матрицы: 1. у матрицы А размерности nxn существует не больше *n* различных собственных чисел; 2. Если $\lambda_1, \lambda_2, \ldots, \lambda_n$ – собственные числа матрицы А, то след матрицы

$$tr(A) = \sum_{i=1}^{n} \lambda_i,$$

а определитель матрицы

$$\det(A) = \prod_{i=1}^{n} \lambda_i.;$$

3. если матрица А идемпотентна, то все её собственные числа равны 0 или 1; 4. все собственные числа вещественной симметричной матрицы вещественны; 5. всякая симметрично положительная (или неотрицательная) определённая матрица размерности *n*x*n* имеет *n* положительных (неотрицательных) действительных характеристических чисел; и наоборот, матрица, все характеристические числа которой положительны (неотрицательны) является симметрично положительной (или неотрицательной) определённой матрицей; 6. всякая симметрично положительная (или неотрицательная) определённая матрица размерности *n*x*n* может быть приведена с помощью ортогонального преобразования X к диагональному виду: $D = X^T AX$, где на диагонали матрицы D стоят характеристические числа матрицы А: $\lambda_1, \lambda_2, \ldots, \lambda_n$.

СОБСТВЕННЫЙ (ХАРАКТЕРИСТИЧЕСКИЙ) ВЕКТОР МАТРИЦЫ

некоторый вектор-столбец $X^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}) \neq 0$, соответствующий характеристическому (собственному) числу λ_i и являющийся решением уравнения: $(A - \lambda I_n)X = 0$. Здесь А – некоторая матрица размерности *n*x*n*, I_n – единичная матрица размерности *n*x*n*.

Всякая квадратная матрица А размерности *n*x*n* имеет *n* собственных чисел (не обязательно различных) и *n* соответствующих им собственных векторов.

Пример нахождения С.в.м.: для матрицы

$$A = \begin{pmatrix} 1 & 3 \\ 12 & 1 \end{pmatrix}$$

характеристическое уравнение имеет вид:

$$\det(A - \lambda I_n) = \begin{vmatrix} 1-\lambda & 3 \\ 12 & 1-\lambda \end{vmatrix} = \lambda^2 - 2\lambda - 35 = 0.$$

Собственные значения: $\lambda_1 = -5$, $\lambda_2 = 7$. Собственный вектор $X^{(1)}$, соответствующий собственному значению $\lambda_1 = -5$, определяется из матричного уравнения: $(A - \lambda_1 I_n) X^{(1)} = 0$ или

$$\begin{pmatrix} 6 & 3 \\ 12 & 6 \end{pmatrix} \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

откуда $x_1^{(1)} = -0{,}5\, x_2^{(1)}$. Положив $x_2^{(1)} = c$, получаем, что векторы $x^{(1)} = (-0{,}5c; c)$ при « $c \neq 0$ являются собственным векторами матрицы А с собственным значением $\lambda_1 = -5$. Аналогично определяются собственные векторы $x^{(2)} = (0{,}5c_1; c_1)$ при « $c_1 \neq 0$ для собственного значения $\lambda_2 = 7$.

Некоторые свойства С.в.м.: 1. если $X^{(i)}$ – С.в.м. А, соответствующий собственному числу λ_i, то для любого скаляра $\alpha \neq 0$, $\alpha X^{(i)}$ – тоже собственный вектор, соответствующий собственному числу λ_i; 2. собственные векторы $X^{(i)}$ и $X^{(j)}$, соответствующие разным собственным числам, всегда взаимно ортогональны, т.е. $(X^{(i)})^T X^{(j)} = 0$ при $\lambda_i \neq \lambda_j$. Нормируя собственные векторы $X^{(i)}$, $i = \overline{1,n}$, в силу того, что собственные вектор определяется с точностью до коэффициента пропорциональности, получим ортонормированную систему:

$$(X^{(i)})^T X^{(j)} = \begin{cases} 1 & при\ i = j \\ 0 & при\ i \neq j \end{cases} \quad i, j = \overline{1,n}.$$

Матрица А, составленная из столбцов $X^{(i)}$: $X = (X^{(1)} X^{(2)} \ldots X^{(n)})$ будет ортогональной, т.е. $X^T X = I_n$.; 3. если матрица А размерности *n*x*n* является вещественной и симметричной, то существуют матрицы X и D, где X – ортогональная матрица размерности *n*x*n* ($X^T = X^{-1}$), столбцы которой – С.в.м. А, а D – диагональная матрица размерности *n*x*n*, состоящая из соответствующих собственных чисел матрицы А, такие что: $A = X D X^T$.

СРЕДНЯЯ МЕРА ВНУТРИКЛАССОВОГО РАССЕЯНИЯ

характеристика качества разбиения исходных многомерных наблюдений X_1, X_2, …, X_n, используемая в задаче кластерного анализа при неизвестном числе классов k и определяемая как:

$$I_\tau^{(K)}(S)=\left\{\frac{1}{n}\sum_{i=1}^{k} n_i\left[Q_\tau^{(K)}(S_i)\right]^\tau\right\}^{\frac{1}{\tau}},$$

где

$$Q_\tau^{(K)}(S_i)=\left[\frac{1}{n_i^2}\sum_{X_j\in S_i}\sum_{X_l\in S_i} d^\tau\left(X_j,X_l\right)\right]^{\frac{1}{\tau}}$$

– функционалы качества разбиения, $S=(S_1, S_2, ..., S_k)$ – некоторое фиксированное разбиение наблюдений X_1, X_2, …, X_n на классы S_i; d – заданная метрика в пространстве $\Pi^p(X)$; числовой параметр τ выбирается исследователем в зависимости от целей разбиения.

При конструировании и сравнении различных кластер-процедур, необходимо учитывать, что объединение двух кластеров S_l и S_m в один даёт прирост величины

$$n\left[I_\tau^{(K)}(S)\right]^\tau,$$

непосредственно характеризующей среднюю меру внутриклассового рассеяния, равный

$$\Delta\left[n\left(I_\tau^{(K)}\right)^\tau\right]=\frac{n_l n_m}{n_l+n_m}\left\{2\left[\rho_\tau^{(K)}(S_l,S_m)\right]^\tau-\left[Q_\tau^{(K)}(S_l)\right]^\tau-\left[Q_\tau^{(K)}(S_m)\right]^\tau\right\}.$$

Для сокращения числа кластеров при наименьших потерях в отношении внутриклассового рассеивания, без учета меры концентрации, целесообразно объединять два кластера, для которых минимальна величина

$$\Delta\left[n\left(I_\tau^{(K)}\right)^\tau\right].$$

В случае, если одновременно ориентироваться и на рост взвешенной концентрации $Z_1(S)$, то объединение кластеров следует подчинить требованию минимизации величины

$$\frac{\Delta\left[n\left(I_\tau^{(K)}\right)^\tau\right]}{\Delta Z_1(S)}=\frac{\left\{2\left[\rho_\tau^{(K)}(S_l,S_m)\right]^\tau-\left[Q_\tau^{(K)}(S_l)\right]^\tau-\left[Q_\tau^{(K)}(S_m)\right]^\tau\right\}n^2}{n_l+n_m}.$$

Понятие С.м.в.р. наряду с понятием меры концентрации было введено А.Н. Колмогоровым в ситуациях, когда не известно, на какое число классов подразделяются исходные многомерные наблюдения. Данные понятия использовались для описания подхода, реализующим идею одновременного учёта двух функционалов качества разбиения, один из которых характеризует внутриклассовый разброс наблюдений, другой – меру взаимной удалённости (близости) классов или меру потерь, которые появляются при излишней детализации рассматриваемого массива исходных наблюдений.

СТАТИСТИЧЕСКИ ЗНАЧИМАЯ СВЯЗЬ

взаимосвязь нескольких признаков или величин, для которой построен измеритель степени тесноты статистической связи (*коэффициент корреляции, корреляционное отношение*, какая-либо информационная характеристика связи, *коэффициент ранговой корреляции* и т.п.) и оценена мера уверенности в «истинности» этого измерителя (т.е. проверена *гипотеза*, показавшая, что полученное числовое значение анализируемого измерителя связи действительно свидетельствует о наличии статистической связи или исследуемая корреляционная характеристика статистически значимо отличается от нуля).

Один из методов анализа С.з.с. нескольких признаков – *корреляционный анализ*, который используется для того, чтобы выбрать (с учётом специфики и природы анализируемых переменных) подходящий измеритель статистической связи; оценить с помощью *оценок точечной* и *интервальной* его числовое значение по имеющимся выборочным данным, проверить гипотезу о статистической значимости анализируемого

измерителя связи, определить структуру связей между компонентами исследуемого многомерного признака, сделав вывод о наличии или отсутствии связи. Степень линейной зависимости между *количественными переменными* характеризуется с помощью *парных, частных* и *множественных коэффициентов корреляции* и *детерминации*. При этом значимость частных и парных коэффициетов корреляции проверяется по *t – критерию Стьюдента* или с помощью табл. Фишера-Иейтса, значимость множественного коэффициента корреляции и коэффициента детерминации – по F-критерию. В случае нелинейной зависимости между количественными переменными для оценки статистической связи используется корреляционное отношение, а статистическая значимость корреляционного отношения проверяется по F-критерию. Для оценки *степени тесноты статистической связи* между порядковыми (ординальными) переменными используют ранговый коэффициент корреляции Спирмена и ранговый коэффициент корреляции Кендалла в случае парной связи, коэффициент конкордации (или согласованности) – в случае нескольких переменных, значимость которого проверяется с помощью χ^2-критерия. Для измерения степени тесноты статистической связи между фиктивными (категоризованными) переменными используют характеристику X^2 квадратичной сопряжённости признаков, коэффициент Крамера или информационную характеристику Y^2 признаков, значимость данных характеристик проверяется с помощью χ^2-критерия.

СТАТИСТИЧЕСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ

методы группировки анализируемой совокупности объектов на сравнительно небольшое число (заранее известное или нет) однородных групп или классов по одному или нескольким информативным с точки зрения исследования признакам. Выделяют методы *дискриминантного анализа*, методы расщепления смесей вероятностных распределений и процедуры кластерного анализа в зависимости от наличия и характера априорных сведений о природе искомых классов и от конечных прикладных целей исследования. Параметрические и непараметрические методы дискриминантного анализа, используемые в случае наличия обучающих выборок, выявляют различия между группами по некоторым переменным и дают возможность классифицировать объекты по принципу максимального сходства. Методы расщепления смесей вероятностных распределений применимы, когда отсутствует предварительная выборочная информация, а исследуемая ген. совокупность задана в виде параметрического семейства законов распределения вероятностей. Методы *кластерного анализа*, наряду с таксономией, распознаванием образов «без учителя», иерархической классификацией, используют, когда отсутствует априорная информация о распределении ген. совокупности и заданы только некоторые самые общие предположения о законе распределения.

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

исследование объектов, явлений или систем посредством построения, изучения, экспериментальной проверки статистических моделей, основанное на одновременном использовании информации о природе и содержательной сущности явлений, представленной в виде теоретических закономерностей, и исходных статистических данных, характеризующих процесс и результаты функционирования изучаемого явления. Цель С.м. – объяснение исследуемых явлений и предсказание. При С.м. используется широкий спектр инструментария *теории вероятностей*, *математической статистики* и прикладной статистики. Теория вероятностей предоставляет исследователю набор *математических моделей*, имитирующих механизмы функционирования гипотетических реальных явлений или систем стохастической природы, а прикладная статистика – обоснованный выбор среди множества возможных моделей той, которая наилучшим образом соответствует имеющимся в распоряжении исследователя статистическим данным, характеризующим реальное поведение исследуемой системы. В рамках статистического моделирования выделяют два

подхода. Один из них представлен методами статистического анализа, предусматривающими возможность вероятностной интерпретации обрабатываемых данных и полученных в результате обработки статистических выводов. В рамках второго подхода используется широкий класс методов статистической переработки исходной информации, которые не опираются на вероятностную природу обрабатываемых данных (например, разнообразные методы кластерного анализа, многомерного шкалирования, теории измерений и др.). Построение и экспериментальная проверка модели, математическое описание интересующих исследователя связей и отношений между реальными элементами анализируемой системы, обычно основаны на одновременном использовании информации двух типов. Первый тип – априорная информация о природе и характере исследуемых соотношений; второй тип – исходные статистические данные, характеризующие процесс и результат функционирования анализируемой системы. Если исследователь располагает информацией обоих типов, то используется прием содержательного (реалистического) математического моделирования, когда из априорной информации о природе искомых соотношений (математически формализованной в виде некоторых исходных предпосылок или исходных допущений) выводится общий вид аналитических уравнений, описывающих эти соотношения. Далее с помощью статистического анализа информации второго типа оцениваются численные значения параметров, входящих в построенные аналитические уравнения (этап подгонки). Если же в распоряжении исследователя только априорная информация первого типа или информация обоих типов, и имитируется поведение анализируемой реальной системы при варьировании численных значений параметров, входящих в аналитическую модель, или искусственно (опираясь на модельные соотношения) генерируются статистические данные второго типа с целью их пополнения. В данном случае наряду с элементами математического моделирования используются средства ЭВМ. Этот тип моделирования называют моделированием типа «Монте-Карло».

С.м., опирающееся на вероятностную природу изучаемого явления, включает в себя несколько осн. этапов. На первом этапе проводится предварительный анализ исследуемой реальной системы, определяются цели моделирования; набор входных и выходных факторов и показателей, взаимосвязи между которыми изучаются. В случае, если исходная статистическая информация отсутствует, то задача сбора статистических данных тоже включается в содержание первого этапа. Второй этап – этап формирования априорной информации, состоящий в формализации ряда гипотез и исходных допущений об исследуемом явлении. Третий этап включает в себя вывод общего вида модельных соотношений, связывающих между собой входные и выходные показатели, определение структуры модели. Четвертый этап моделирования посвящён статистическому оцениванию неизвестных параметров, входящих в модель, и исследованию свойств полученных оценок, их точности. На пятом этапе (этапе верификации модели) используются различные процедуры анализа адекватности и точности модели. Присутствие шестого этапа зависит от результатов предыдущего этапа. Он заключается в планировании и проведении исследований, направленных на уточнение модели и на дальнейшее развитие и углубление второго этапа.

В моделировании типа «Монте-Карло» этапы исследования несколько отличаются. Первый этап – исходный анализ изучаемого явления, второй этап – составление детального плана сбора исходной статистической информации с учетом схемы дальнейшего статистического анализа. На третьем этапе осуществляется сбор исходных статистических данных и их ввод в ЭВМ. Четвертый этап – первичная статистическая обработка данных: статистическое описание исходных совокупностей с определением пределов варьирования переменных; анализ резко выделяющихся наблюдений; восстановление пропущенных наблюдений; проверка статистической независимости последовательности наблюдений, составляющих массив исходных данных; экспериментальный анализ закона распределения исследуемой ген. совокупности, параметризация сведений о природе

изучаемых распределений и т.д. Пятый этап – составление детального плана вычислительного анализа материала, описывается блок-схема анализа с указанием привлекаемых методов, формулируется оптимизационный критерий, на основании которого выбирается один из альтернативных методов основной статистической обработки исходных данных. Шестой этап – вычислительная реализация статистической обработки данных, эффективное управление вычислительным процессом путем формулировки задачи обработки и описания данных на входном языке пакета, либо составление программы на алгоритмическом языке для реализации ЭВМ. Седьмой этап – подведение итогов исследования.

С.м. нашло широкое применение для работы с вероятностными моделями на всех этапах исследования. Так, в *теории массового обслуживания* это осн. метод решения сложных систем. В классической статистике – один из способов изучения устойчивости оценок к отклонениям от базовых предположений, используемый как самостоятельно, так и как дополнительный прием к асимптотическим аналитическим методам. При планировании исследований, в случае сложной модели изучаемого явления, статистическое моделирование может помочь найти объемы основной и контрольной выборок и т.д. Одним из примеров техники статистического моделирования наблюдений, подчиняющихся заданному распределению, является получение равномерно распределённых на отрезке [0, 1] случайных чисел, моделирование дискретных и непрерывных случайных величин и т.д. Так, напр., для моделирования случайных чисел с помощью ЭВМ используют два метода: «физический», когда с ЭВМ соединяется тот или иной «физический» датчик случайных чисел (напр., счётчик числа α-частиц, вылетающих из некоторого радиоактивного источника за фиксированный промежуток времени), или математический, когда в ЭВМ с помощью стандартных машинных команд генерируется регулярная последовательность случайных чисел, удовлетворяющая основным неравенствам. Эту последовательность называют последовательностью псевдослучайных чисел. Чаще всего используется математический метод, т.к. в С.м. важно иметь возможность воспроизвести последовательности случайных чисел, чтобы проанализировать, как на тех же данных будет работать другой метод статистической обработки.

СТЕПЕНЬ СОГЛАСОВАННОСТИ МНЕНИЙ ЭКСПЕРТОВ

характеристика степени согласованности мнений группы экспертов, в том случае, когда речь идёт о согласованности мнений двух экспертов, используют ранговые коэффициенты корреляции (Спирмена и Кендалла). Это мнение заключается в установлении ранга, места, занимаемого в последовательности статистически обследованных объектов, ранжированных по степени проявления в них анализируемого свойства. В результате получается т.н. ранжировка (перестановка) n заданных объектов. В случае, если объекты невозможно упорядочить по степени проявления данного свойства, но возможно отнести исследуемый объект в определённый класс или категорию, то применяются коэффициенты согласованности каппа. Т.о., мы имеем дело со *случайными величинами* – ординальными (порядковыми) с одной стороны, с другой – с *номинальными* (классификационными).

При анализе С.с. мнений двух и более экспертов и сравнении их компетентности выделяется группа экспертов более компетентных, чем остальные. Для исследования различных комбинаций соответствующих порядковых переменных применяется *коэффициент конкордации*, измеряющий С.с. В случае увеличения числа экспертов и числа категорий классификационного признака вводится обобщённый коэффициент согласованности каппа.

СТЕПЕНЬ ТЕСНОТЫ СТАТИСТИЧЕСКОЙ СВЯЗИ

см. в ст. Теснота статистической связи

Т

ТАБЛИЦА (МАТРИЦА) «ОБЪЕКТ-СВОЙСТВО»

представляет исходные данные в наиболее общем и распространенном виде. Эта матрица задаёт результаты статистического наблюдения и сопоставления объект-признак в виде табл.:

$X=(x_{ij})=$	x_{11}	·	·	·	x_{1j}	·	·	·	x_{1m}	,
	·	·	·	·	·	·	·	·	·	
	x_{i1}	·	·	·	x_{ij}	·	·	·	x_{im}	
	·	·	·	·	·	·	·	·	·	
	x_{m1}	·	·	·	x_{mj}	·	·	·	X_{nmn}	

где x_{ij} значение j-го признака у i-го объекта. Т.о., X^{T}_{i}, = (x_{i1}, …, x_{ij}, …, x_{im}) есть вектор-столбец значений координат случайного вектора для i-го наблюдения. Столбец x_{1j}, …, x_{ij}, …, x_{nj} – множество n значений j-го свойства, измеренного на совокупности n объектов. Объекты представляются строкой, а вместо термина свойство употребляют термины одномерная случайная величина, признак, показатель. Обработка матриц данных предваряется *оцифровкой* номинальных и порядковых свойств.

ТАБЛИЦА СОПРЯЖЕННОСТИ

табл. сопоставления частот (количеств объектов выборки) двух неколичественных признаков; чаще всего встречается при обработке исходной информации и сравнении с теоретическими, гипотетическими моделями распределения ген. совокупности.

Т.с. имеет r строк, отвечающих количеству категорий первого признака X и s столбцов соответствующих категориям признака Y, следовательно размер таблицы r x s. В первом внешнем столбце Т.с. обозначены номерами неколичественные значения, категории признака X: 1, 2, …, i, …, r. Аналогично в первой внешней строке расположены оцифрованные категории номинальных или ординальных категорий признака Y: 1, 2, …, j, …, s. Далее, в последней внешней итоговой строке помещаются суммарные значения наблюдаемых частот n_{ij}, находящихся в столбце j; эта сумма обозначается символом n_{*j}. Правый, крайний внешний столбец соответствует итоговым частотам категорий признака X. Внутри Т.с. располагаются клетки, обозначаемые через (i, j), в которых помещаются наблюдаемые частоты n_{ij} – число элементов (объектов, индивидов), принадлежащих i-й категории признака X и одновременно j-й категории признака Y. На пересечении итоговой строки и итогового столбца находится итоговое балансовое число $n_{**} = n$ – объем выборки. Заметим, что вместо n_{ij}, i = 1,2, …, r и j = 1, 2, …, s могут стоять вероятности p_{ij} или так называемые теоретические частоты n^{*}_{ij}, оцениваемые по модели распределения вероятностей. Кроме того используются логарифмы частот для исследования логарифмически линейных моделей. Если вместо оцифрованных количественных признаков исследуются числовые, количественные признаки, то Т.с. называют корреляционной табл. Т.с. может быть одномерной или трёхмерной и т.д.

ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ

отдельный раздел *теории вероятностей*, изучающий системы массового обслуживания, реализующие многократное выполнение однотипных задач (требований, заявок) случайного характера. Теоретической основой теории являются *случайные процессы*, числовые характеристики которых имеют непосредственное отношение к формулированию и решению задач Т.м.о. Типичный пример таких систем массового обслуживания – автоматическая телефонная станция, где случайным образом поступают заявки – вызовы абонентов, а обслуживание состоит в соединении их с другими абонентами. К системам массового обслуживания относится деятельность многих отраслей производ-

ственной и непроизводственной сфер, таких как ремонт и наладка оборудования, автотранспорта, сервисное, медицинское, социальное обслуживание, финансовые учреждения (банки, кредитные, страховые орг-ции, налоговые инспекции, аудиторские службы) и т.д.

Обслуживающие структуры, входящие в системы массового обслуживания называются каналами обслуживания, поэтому системы массового обслуживания бывают одноканальными или многоканальными. Для проведения расчётов на основе Т.м.о. необходимо иметь информацию о потоке заявок, поступающих в случайные моменты времени, длительности их исполнения, времени между поступлениями заявок, числе обслуживающих каналов, данные о потоке обслуженных и необслуженных заявок и т.д.

Различают два осн. вида массового обслуживания: система массового обслуживания (СМО) с потерями (или отказами). В этом случае заявка, пришедшая в момент, когда все каналы обслуживания заняты, получает отказ и покидает систему. Для того, чтобы заявка была обслужена, надо её вновь подать на вход СМО, как поступившую впервые. Одной из осн. характеристик этого вида СМО является вероятность отказа, т.е. вероятность того, что в момент заявки все каналы окажутся занятыми; СМО с ожиданием (или очередью). В этом случае заявка не покидает систему, а становится в очередь и ждет, пока не освободится какой-нибудь канал. Время ожидания и число мест в очереди могут быть как неограниченными, так и ограниченными. Одна из осн. характеристик данного вида массового обслуживания – среднее время ожидания.

Кроме этих двух видов массового обслуживания в Т.м.о. рассматривается третий – модель обслуживания с ограничениями. К этому виду модели относятся смешанные, промежуточные, усложненные системы, в которых на заявки могут накладываться отдельные ограничения. Напр., система с ограниченным временем ожидания, после истечения которого заявка аннулируется; система с ожиданием и ограничением на длину очереди – заявка аннулируется, если очередь достигает предельную длину (число объектов обслуживания); система с числом ожидающих, равным предельному числу; система с приоритетной группой требований на обслуживание – заявки от участников ВОВ, инвалидов; система с объединением нескольких систем массового обслуживания в единую систему высшего порядка – к работе привлекаются диспетчеры. О процессах массового обслуживания требуется следующая информация: входящий поток требований, для задания которого необходимо знать распределение моментов поступления их, или, в более частном случае, распределение интервалов между моментами поступления; закон распределения времени обслуживания требования; дисциплина обслуживания очереди, число обслуживающих каналов, организация очереди и процесса обслуживания.

Наиболее распространённой дисциплиной является дисциплина «первый пришёл – первый обслужен», т.е. непосредственный порядок обслуживания, когда обслуживание требования начинается, как только оно достигает начала очереди.

Предполагается, что входящий поток требований не зависит от размера очереди и что интервалы между последовательными моментами поступления требований являются независимыми одинаково распределенными положительными случайными величинами. Такие потоки иногда называют рекуррентными, или потоками восстановления. Понятие «простейший поток» применяется в случае, когда поступления заявок образуют пуассоновский поток, а интервалы между моментами поступления распределены по показательному закону. Предполагается, что время обслуживания отдельных требований – независимые одинаково распределённые случайные величины, не связанные с входящим потоком. При решении многих задач Т.м.о. используются такие случайные потоки, как стационарные, эрланговские, марковские и др.

ТЕСНОТА СТАТИСТИЧЕСКОЙ СВЯЗИ

количественная характеристика, измеряемая числовой переменной, которая изменяется в пределах от нуля до единицы, когда подразумевается абсолютная величина тесноты связи, или от минус единицы до плюс единицы, когда учитывается направление (знак) связи или взаимозависимости между двумя случайными величинами. Измерители связи называются *коэффициентами корреляции*, связи, взаимозависимости, сопряжённости и т.п. Обычно от такого коэффициента требуется, чтобы были известны границы изменения этого коэффициента. Коэффициент должен принимать среднее или нижнее значение интервала изменения в случае, когда случайные величины не связаны (независимы). Если коэффициент меняется от (– 1) до (+ 1) , то случай независимости должен соответствовать нулевому значению коэффициента. Такое требование удобно тем, что соответствует широко употребляемому обычному коэффициенту корреляции. В случаях, когда используется квадрат коэффициента связи двух величин или коэффициент, измеряющий тесноту зависимости между результативным признаком и совокупностью, содержащей более одного признака – аргумента, тогда изменению в интервале (0, 1) соответствует нулевое значение в случае отсутствия зависимости или взаимосвязи. Измерение Т.с. – одна из задач *корреляционного анализа*. Показателями Т.с. являются *коэффициенты детерминации* (квадратов соответствующих *коэффициентов корреляции*), *парный*, *частный* и *множественный* коэффициенты корреляции и, ранговые коэффициенты. Для изучения Т.с. между номинальными признаками, а также ординальными признаками используются коэффициенты сопряжённости, построенные на основе статистики хи-квадрат Пирсона. Применяются меры связи Гудмена-Краскала, основанные на качестве прогноза категории одного признака при известной и неизвестной информации о другом признаке. Кроме того, используются меры связи, возникающие в логлинейных моделях. Для четырёхклеточных *табл. сопряжённости* применяются коэффициенты Юла, основанные на статистике отношения перекрестных произведений частот.

Ф

ФАКТОРНАЯ ДИСПЕРСИЯ

дисперсия, характеризующая влияние некоторого качественного фактора F, имеющего p уровней $F_1, F_2, \ldots, F_p$ на изучаемую величину X.

Ф.д. применяется в *дисперсионном анализе*, заключающемся в сравнении Ф.д., порождаемой воздействием фактора, и остаточной дисперсии, обусловленной случайными причинами.

Пусть на количественный нормально распределённый признак X воздействует фактор F, имеющий p постоянных уровней (см. табл.):

Таблица

Номер испытания	Уровни фактора F_j			
	F_1	F_2	…	F_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…
q	x_{q1}	x_{q2}	…	x_{pp}
Групповая средняя	$\bar{x}_{\text{гр.}1}$	$\bar{x}_{\text{гр.}2}$	…	$\bar{x}_{\text{гр.}p}$

Принимается, что число наблюдений (испытаний) на каждом уровне одинаково и равно q. Наблюдается n = pq значений x_{ij} признака X, где i – номер испытания (I = 1, 2, …, q), j – номер уровня фактора (j = 1, 2, …, p).

Ф.д. определяется отношением:

$$s^2_{\text{факт.}} = \frac{S_{\text{факт.}}}{p-1},$$

где

$$S_{\text{факт.}} = q\sum_{j=1}^{p}(\bar{x}_{\text{гр.}j} - \bar{x})^2$$

– факторная сумма квадратов отклонений групповых средних от общей средней, характеризующая рассеяние между группами; p – число уровней фактора; (p-1) – число степеней свободы.

ФОРМУЛА БАЙЕСА

см. в ст. Теорема (формула) Байеса

Х

ХЕММИНГОВО РАССТОЯНИЕ

расстояние, используемое как мера различия объектов, задаваемых дихотомическими (номинальными) признаками. Это расстояние определяется по формуле:

$$\rho_H(x_i, x_j) = \sum_{e=1}^{k}\left|x_{ie} - x_{je}\right|$$

и равно числу несовпадений значений соответствующих признаков, в рассматриваемых i-м и j-м объектах.

Ц

ЦЕЛЕВАЯ ФУНКЦИЯ

функция обобщённого свойства объекта**,** характеризуемого значениями его частных критериальных признаков; средство формализации зависимости между частными измеряемыми факторами и латентным агрегированным показателем на основе статистической и экспертной информаци. Решается задача построения единого обобщающего латентного показателя качества некоторой системы *Y*, характеризующейся вектором измеряемых частных критериев $X = (X_1, X_2, ..., X_p)$. В качестве исходных данных в этой задаче используются значения критериев, определенные по результатам наблюдения за объектами $O_1, O_2, ..., O_n$, а также набор экспертных оценок искомого латентного показателя для каждого объекта от *m* независимых экспертов. Под Ц.ф. понимают некоторое преобразование $\varphi(X)$, которое сохраняет соотношение порядка между этими объектами, т. е. из системы предпочтений $O_{i_1} \succ O_{i_2} \succ ... \succ O_{i_n}$ следует система неравенств $\varphi(O_{i_1}) \succ \varphi(O_{i_2}) \succ ... \succ \varphi(O_{i_n})$. Эта функция определяется с точностью до произвольного монотонного преобразования и отражает установки экспертов при выборе оценок. В основе отыскания Ц.ф. лежит принцип минимизации отклонений экспертных оценок искомого показателя и его оценок, полученных с использованием этой функции. Алгоритм построения Ц.ф. зависит от вида экспертных данных. Если эти данные представлены в виде матрицы $\|y_{ij}\|_{n\times m}$ бальных оценок *Y* для каждого *i*-го объекта каждым *j*-м экспертом, то задача сводится к обычной схеме регрессионного анализа:

$$\begin{cases} y_{ij} = f(X_i, \theta) + \varepsilon_j(X_i) \\ \mathrm{E}\varepsilon_j(X_i) = 0, \quad D\varepsilon_j(X_i) = \sigma_{ij}^2 \end{cases}$$

Оценка параметров Θ находится путём решения оптимизационной задачи вида:

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{1}{\sigma_{ij}^2}(y_{ij} - f(X_i, \theta))^2 \to \min_{\theta}$$

Если информация дана в виде матриц парных сравнений

$$\gamma_k = \left\|\gamma_{ij}^{(k)}\right\|_{n\times n}, \quad k = 1,2,...,m,$$

задающих экспертные разбиения множества $\{O_1, O_2, ..., O_n\}$ на классы по искомому показателю, то оценки параметров Ц.ф. определяются из минимизации близости таких разбиений и разбиения $\gamma(\delta;\theta)$, полученного с помощью Ц.ф.:

$$\sum_{j=1}^{m} d(\gamma_j; \gamma(\delta;\theta)) \to \min_{\theta},$$

где

$$d(\gamma_j; \gamma(\delta;\theta)) = \frac{1}{2}\sum_{i,\,j=1}^{n}\left|\gamma_j - \gamma(\delta;\theta)\right|.$$

Разбиение $\gamma(\delta;\theta)$ строится на основе линейной аппроксимации Ц.ф.

$$\hat{f}(X;\theta)=\sum_{l=1}^{p}X_l\theta_l\ ,$$

так, что по заданному $\delta>0$ относят в один класс объекты, для которых $0\le\hat{f}(X;\theta)\le\delta$, в другой – те, у которых $\delta\le\hat{f}(X;\theta)\le 2\delta$ и т.д.

В случае парных сравнений в виде отношения предпочтения вектор оценок $\hat{\boldsymbol{\theta}}$ определяется из условия минимизации невязок.

$$b_{i_q k_q}(\theta)=\begin{cases}0, & f(X_{i_q};\theta)-f(X_{k_q};\theta)\ge 0\\ -(f(X_{i_q};\theta)-f(X_{k_q};\theta)), & f(X_{i_q};\theta)-f(X_{k_q};\theta)<0\end{cases},$$

введённых в каждое неравенство системы:

$$f(X_{i_q};\theta)-f(X_{k_q};\theta)\ge 0,\quad q=1,2,\ldots,N,\quad N\le C_n^2\ ,$$

при некоторых ограничениях (типа нормировки) на θ:

$$\sum_{q=1}^{N}b_{i_q k_q}(\theta)\to\min_{\theta}\ .$$

Если экспертные оценки заданы в виде матрицы $\|R_{ij}\|_{n\times m}$ рангов, присвоенных каждому *i*-му объекту каждым *j*-м экспертом, то при отыскании Ц.ф. эту информацию предварительно преобразуют в соответствующий набор матриц парных сравнений.

См. также *Экспертно-статистический метод.*

ЦЕЛЕНАПРАВЛЕННОЕ ПРОЕЦИРОВАНИЕ

совокупность методов снижения размерности многомерных данных путём их линейного проецирования в пространство меньшей размерности (в зарубежной литературе – «projection pursuit»). Эти методы являются обобщением классических методов *многомерного статистического анализа*, таких, как *факторный анализ*, анализ главных компонент, линейный *дискриминантный анализ*. Задача Ц.п. заключается в определении такого линейного отображения (способа проецирования) U исходного множества X в пространство меньшей размерности, которое бы оптимизировало заданный функционал (критерий) качества Q(U,X), который при этом называется проекционным индексом и подбирается так, чтобы в спроецированных данных сохранялась вся информация о структуре исходных многомерных данных. На первом этапе Ц.п. в зависимости от целей исследования определяется проекционный индекс, напр., используются проекционные индексы, минимизирующие расстояние от исходных точек до их проекций, а также проекционные индексы, описывающие меру искажения взаимных расстояний между точками в исходном и результирующем пространстве. На втором этапе отыскивается отображение U, оптимизирующее Q(U,X). При этом используются пошаговые процедуры условной оптимизации (напр., при условиях линейной независимости или ортогональности искомых векторов), а также безусловной оптимизации на всем многообразии операторов проецирования заданной размерности. Они основаны на градиентных методах и требуют больших вычислительных затрат, поэтому при практическом применении рекомендуется предварительно сокращать размерность исходного массива классическими методами, напр., *методом главных компонент*. Кроме того, рекомендуется также предварительно подавлять влияние аномальных наблюдений, существенно искажающих результаты Ц.п. Методы Ц.п. находят своё применение в моделях разведочного анализа данных, в частности, *визуализации данных,* и определения аномальных наблюдений. Ц.п. служит инструментом для выделения нелинейных структур в исходном массиве, используется для построения обобщенной линейной модели множественной регрессии, а также широко применяется в различных задачах классификации, например, технической и медицинской диагностике. Кроме того, процедуры Ц.п., позволяющие восстанавливать плотность распределения вероятностей *случайной величины многомерной* по её проек-

циям, лежат в основе решения задач вычислительной томографии.

ЦЕНТРОИДНЫЙ МЕТОД

метод статистического оценивания факторных нагрузок и остаточных дисперсий в модели

$$X = QF + U,$$

$$X = (X_1, X_2, ..., X_p)^T,\ F = (F_1, F_2, ..., F_{p'})^T,\ Q = \|q_{ij}\|_{p\times p'},\ U = (U_1, U_2, ..., U_p)^T,$$

$$\mathrm{E}X = 0,\ \Sigma = \|\sigma_{ij}\|_{p\times p'},\ \sigma_{ij} = cov(X_i, X_j),\ \mathrm{E}U = 0,\ V = \|v_{ij}\|_{p\times p'},\ v_{ij} = cov(U_i, U_j).$$

Ц.м. – способ определения Q и V при дополнительном условии отсутствия связи между исходными факторами и некоторыми общими факторами имеет простую геометрическую интерпретацию. Если сопоставить признакам $X_1, X_2, ..., X_p$ набор векторов, выходящих из начала координат, так, что $r_{ij} = cos(X_i, X_j)$ $|X_i| = \sqrt{\sigma_{ii}}$, а затем скорректировать их знаки для формирования тенденции к группировке в одном направления в «пучок», то нормированная сумма этих векторов определит первый общий фактор F_1. Переход к остаточным переменным $X_i^{(1)} = X_i - q_{i1}F_1$, $\Sigma^{(1)} = \Sigma - q_1 q_1^T$, и повторение относительно них той же процедуры даёт второй общий фактор F_1 и т. д. Поскольку $F_1, F_2, ..., F_{p'}$ проходят через центры «пучков», то их называют «центроидами». Для реализации Ц.м. используют итерационную схему вычислений. На нулевом шаге выбирают начальные приближения $V^{(0)}$ для V (напр., $v_{ij}^0 = \hat{\sigma}_{ii}[1 - max|r_{ij}|]$), а также $b_1^{(0)} = (1,1,...,1)^T$ для первого столбца b_1 матрицы B, содержащей веса, с которыми векторы «пучков» образуют «центроиды». Последовательно определяют начальное приближение

$$Q^{(0)} = (q_1^{(0)}, q_2^{(0)}, ..., q_{p'}^{(0)})$$

матрицы Q:

$$q_1^{(0)} = \frac{\psi^{(0)} b_1^{(0)}}{\sqrt{b_1^{(0)T} \psi^{(0)} b_1^{(0)}}},$$

где

$$\psi^{(0)} = \hat{\Sigma} - V^{(0)},$$

факторного анализа. Рассматривается линейная модель *факторного анализа* в виде:

$$q_2^{(0)} = \frac{\psi_1^{(0)} b_2^{(0)}}{\sqrt{b_2^{(0)T} \psi_1^{(0)} b_2^{(0)}}},$$

где

$$\psi_1^{(0)} = \psi^{(0)} - q_1^{(0)} q_1^{(0)T},$$

и вектор $b_2^{(0)}$ состоит из +1 или -1 так, чтобы максимизировать знаменатель и т.д. На первом шаге вычисляют $V^{(1)} = \Sigma - Q^{(0)} Q^{(0)T}$ и переходят к следующей итерации и т.д. Оценки Ц.м. близки к оценкам макс. правдоподобия, однако, как и всякий непараметрический метод, Ц.м. является более «устойчивым» по отношению к отклонениям от нормальности признаков и требует меньшего объёма вычислений. Недостаток Ц.м. – наличие определённого произвола в его процедуре и невозможность его статистической оценки. Недостатком Ц.м. является также зависимость факторных нагрузок от шкалы, в которой измерены исходные признаки. Поэтому исходные признаки обычно нормируют с помощью среднеквадратических отклонений, так что выборочная *ковариационная* матрица заменяется во всех рассуждениях выборочной *корреляционной* матрицей.

ЦЕПОЧЕЧНЫЙ ЭФФЕКТ

эффект, который возникает при образовании *классов* с помощью агломеративного иерархического *метода «ближайшего соседа»* и проявляется в том, что элементы одного класса могут оказаться более далекими (в смысле *меры близости*), чем элементы разных классов. Реализация алгоритма выделения классов указанным методом может приводить к получению

кластеров достаточно сложной формы. В частности, они не обязаны быть выпуклыми, поскольку два элемента попадают в один кластер, если существует соединяющая их цепочка близких между собой элементов. В качестве примера приводятся *дендрограммы* для процедур метода «ближайшего соседа» (рис. 1б) и *метода «дальнего соседа»* (рис. 1в) на основе данных, представленных на рис. 1а.

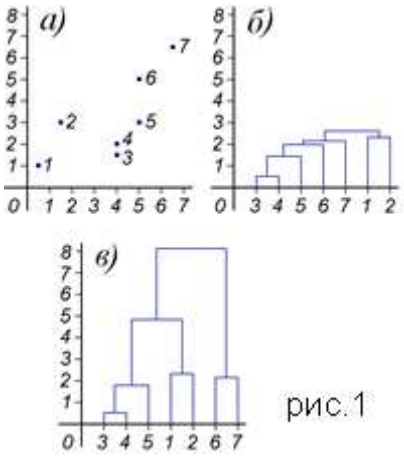


рис.1

На рис. 1б хорошо заметно проявление цепочечного эффекта: происходит последовательное «удлинение» кластера путём присоединения ближайших к границе объектов, в результате чего расстояние между крайними элементами (3 и 7) в этом классе превосходит расстояние между элементами в разных классах (2 и 7). Метод «дальнего соседа» не приводит к подобному эффекту. Существуют различные способы устранения Ц.э. Наиболее простым и естественным из них можно признать, напр., введение ограничения сверху на максимальное расстояние между элементами одного класса: если при формировании классов для некоторых элементов получаемого кластера взаимное расстояние превысит некоторый заданный порог, то эти элементы следует разнести по какому-либо дополнительному правилу в различные классы. См. также *Кластерный анализ.*

Ч

ЧАСТНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

коэффициент, измеряющий тесноту линейной связи между двумя переменными $x^{(j)}$ и $x^{(k)}$ в ситуации, когда значения остальных переменных, входящих в модель, зафиксированы на их средних уровнях, т.е. в ситуации, когда исключено опосредованное влияние этих переменных на взаимосвязь между $x^{(j)}$ и $x^{(k)}$. Если исследуемые переменные $x^{(1)}$, $x^{(2)}$,…, $x^{(p)}$ подчиняются p-мерному нормальному закону, то для подсчёта значения Ч.к.к. используется формула:

$$\rho_{jk/N(j,k)} = -\frac{R_{jk}}{\sqrt{R_{jj}\cdot R_{kk}}},$$

где N(j, k) – набор номеров всех анализируемых переменных за исключением номеров j и k; $R_{q,s}$ – алгебраическое дополнение элемента $\rho_{q,s}$ в *корреляционной матрице* R, соответствующей ряду исследуемых переменных. Напр., Ч.к.к. $\rho_{12/3}$ между $x^{(1)}$ и $x^{(2)}$ при исключённом опосредованном влиянии переменной $x^{(3)}$ будет рассчитываться:

$$\rho_{12/3} = -\frac{R_{12}}{\sqrt{R_{11}\cdot R_{22}}} = \frac{\rho_{12}-\rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}.$$

Ч.к.к. обладает всеми свойствами парного, т.е. изменяется в пределах от -1 до +1. Если Ч.к.к. равен ±1, то связь между исследуемыми признаками функциональная, а равенство его нулю свидетельствует о линейной независимости этих величин. Выборочные значения Ч.к.к. $r_{jk/N(j,k)}$, $r_{12/3}$, вычисляются по вышеприведенным формулам с заменой теоретических значений парных коэффициентов корреляции их выборочными аналогами.

При проверке гипотез H_0: $\rho_{jk/N(j,k)} = 0$ и при построении *оценок интервальных* для $\rho_{jk/N(j,k)}$ следует пользоваться теми же правилами, что и для обычных *парных коэффициентов корреляции* $\rho_{jk/l1,\,l2,\,\ldots,\,lm}$ с одной поправкой: объём выборки n надо заменить на n – m, где m – порядок Ч.к.к.

См. также *Корреляция*, *Многомерный статический анализ*.

Ш

ШКАЛИРОВАНИЕ МНОГОМЕРНОЕ

см. в ст. *Многомерное шкалирование*

Э

ЭВРИСТИЧЕСКИЕ МЕТОДЫ СНИЖЕНИЯ РАЗМЕРНОСТИ

методы снижения размерности, в основе которых лежит опыт и интуиция исследователя. В основе Э.м.с.р. отсутствует строгая вероятностная модель, хотя могут быть использованы отдельные понятия *теории вероятностей* и *математической статистики*. Эти методы подчинены некоторым частным целевым установкам (наименьшее искажение геометрической структуры исходных «выборочных точек», наименьшее искажение их эталонного разбиения на классы и т.д.), но не формулируемых в терминах вероятностно-статистической теории. Процедура выбора целевой установки, подходящей именно для данной конкретной задачи носит эвристический характер.

Применение Э.м.с.р. обусловлено отсутствием в анализе предварительной информации, напр., обучающих и квазиобучающих выборок, и не столь эффективно по сравнению с применением наиболее обоснованных *математико-статистических методов*. К Э.м.с.р. относится метод экстремальной группировки признаков, который используется в целях лаконичного объяснения природы многомерных данных путём определения наиболее информативных признаков – детерминант, обуславливающих изменения в соответству-ющих им группах. В отличие от классических моделей *факторного анализа*, применяя-емых в аналогичных задачах, при эвристически оптимизационном подходе группировка признаков и выделение общих факторов производятся на основе экстремизации некоторых эвристически введённых функционалов, в частности, в качестве критерия оптимальности используются функционалы, представляющие собой сумму квадратов или сумму модулей *парных коэффициентов корреляции*, определённых для каждой группы между всеми признаками из этой группы и соответствующим этой группе признаком-детерминантом. К Э.м.с.р. относится также *метод корреляционных плеяд*, который используется в целях систематизации изучаемых признаков. Метод корреляционных плеяд, так же как и метод экстремальной группировки, предназначен для нахождения таких групп признаков – «плеяд», когда корреляционная связь между параметрами одной группы (внутриплеядная связь) достаточно велика, а связь между параметрами из разных групп (межплеядная) – мала. По определённому правилу по *корреляционной матрице* признаков образуют граф, который затем с помощью различных приёмов разбивают на подграфы. Элементы, соответствующие каждому из подграфов, и образуют плеяду. См. также *Метод главных компонент*, *Снижение размерности исследуемого пространства*, *Экстремальная группировка признаков*, *Фактор общий*.

ЭКСПЕРТНОЕ УПОРЯДОЧЕНИЕ

обследованных объектов по степени проявления в них анализируемого свойства – форма представления результатов экспертного оценивания изучаемых объектов $O_1, O_2, ..., O_n$ в виде матрицы $R = \|R_{ij}\|_{n\times m}$ значений рангов, присвоенных экспертами этим объектам, (ранжировки).

Статистический анализ Э.у. осуществляется с помощью методов *ранговой корреляции*. При этом решаются осн. задачи: задача А: – анализ структуры Э.у. путём её интерпретации в виде множества точек $R_j = (R_{1j}, R_{2j}, ..., R_{nj})$ в n-мерном пространстве.; задача В – анализ интегральной согласованности мнений экспертов и согласованности мнений групп экспертов, а также условная ранжировка последних по их компетентности. В основе анализа лежит расчёт *коэффициента конкордации* (согласованности) $W(s)$ $(s \le m)$ используемого в качестве меры связи между несколькими порядковыми переменными. Его выборочное значение $\hat{W}(s)$ определяется по формулам для случаев соответственно отсутствия и наличия неразличимых рангов:

$$\hat{W}(s) = \frac{12}{s^2(n^3 - n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{s} R_{ik_j} - \frac{s(n+1)}{2} \right)^2 ,$$

$$\hat{W}(s)=\frac{\sum_{i=1}^{n}(\sum_{j=1}^{s}R_{ik_j}-\frac{s(n+1)}{2})^2}{\frac{1}{12}s^2(n^3-n)-s^2\sum_{j=1}^{s}T_{k_j}},\quad T_{k_j}=\sum_{i=1}^{l_{k_j}}(t_{ik_j}^3-t_{ik_j}).$$

l_{k_j} – число случаев неразличимых рангов;

t_{ik_j} – число повторений для *i*-го случая неразличимых рангов.

Проверка значимости коэффициента *W*(*s*) основана на том, что в условиях отсутствия исследуемой ранговой связи в генеральной совокупности распределение случайной величины

$$\frac{1}{2}ln\frac{(s-1)\hat{W}(s)}{1-\hat{W}(s)}$$

приближенно описывается Z-распределением Фишера с числом степеней свободы числителя

$$\nu_1=n-1-\frac{2}{m}$$

и знаменателя $\nu_2=(s-1)\nu_1$; Задача С – построение единого упорядочения объектов на основе совокупности *m* согласованных упорядочений. В качестве решения этой задачи часто принимают ранжировку, составленную из средних арифметических или медиан, определенных для каждого объекта на исходном множестве рангов. Обоснованием этого подхода является минимизация близости (в смысле определённой меры) между искомым упорядочением объектов и исходными ранжировками, относящимися к *m* экспертам.

Э.у. используется для решения задачи построения единого обобщающего латентного показателя качества некоторой системы, характеризующейся вектором измеряемых частных критериев. В качестве исходных данных в этой задаче выступают значения критериев, определенные по результатам наблюдений за объектами $O_1,O_2,...,O_n$, а также набор экспертных оценок искомого латентного показателя, полученных для каждого объекта от *m* экспертов. Э.у. представляет собой одну из трёх форм выражения такой экспертной информации.

См. также *Значимость коэффициента корреляции*, *Экспертно-статистический метод*.

ЭКОНОМИКО-МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

см. в ст. *Модель экономико-математическая*

ЭКСТРЕМАЛЬНАЯ ГРУППИРОВКА ПРИЗНАКОВ

оптимальное в смысле определённого, эвристически введённого критерия, разбиение исходного множества признаков на группы с выделением наиболее информативных признаков – детерминант (общих факторов), обуславливающих изменения в соответствующих им группах. Метод Э.г.п. используется в целях лаконичного объяснения природы многомерных данных при отсутствии в анализе предварительной информации, напр., обучающих и квазиобучающих выборок. В общем случае под задачей экстремальной группировки случайных величин

$X_1,X_2,...,X_p$ на заранее заданное число групп p' $(p'<p)$ понимают отыскание такого набора групп

$$A_k=\{X_i, i\in S_k\},\quad k=1,2,...,p,$$

$$S_k\subset\{1,2,...,p\},\ \bigcup_{k=1}^{p'}S_k=\{1,2,...,p\},$$

$$S_k\cap S_l=0,\ k\neq l,$$

и таких нормированных (с единичной дисперсией) общих факторов f_k, $k=1,2,...,p'$, которые максимизируют какой-либо критерий оптимальности. Для отыскания Э.г.п. используются следующие два эвристически определённые функционала качества разбиения:

$$I_1=\sum_{k=1}^{p'}\sum_{i\in S_k}[corr(X_i,f_k)]^2,$$

$$I_2=\sum_{k=1}^{p'}\sum_{i\in S_k}|corr(X_i,f_k)|,$$

где $corr(X_i,f_k)$ – *парный коэффициент корреляции* между X_i и f_k.

Максимизация функционалов как по разбиению признаков на группы, так и по выбору общих факторов отвечает требованию такого разбиения параметров, когда в одной группе оказываются наиболее «близкие» между собой признаки. Для случая использования I_1 при заданных классах $S_1, S_2, ..., S_{p'}$ оптимальный набор факторов $f_1, f_2, ..., f_{p'}$ определяется из условия независимой максимизации каждого слагаемого

$$\sum_{i \in S_k} [corr(X_i, f_k)]^2, \quad k = 1,2,..., p'$$

по формулам:

$$f_k = \frac{\sum_{i \in S_k} \alpha_{ik} X_i}{\sqrt{\sum_{i \in S_k} \alpha_{ik} \alpha_{jk} corr(X_i, X_j)}}, \quad k = 1,2,..., p'$$

где $\alpha_k = (\alpha_{1k}, \alpha_{2k}, ..., \alpha_{pk})$ – собственный вектор, соответствующий макс. собственному значению λ_k матрицы, составленной из коэффициентов корреляции переменных, входящих в оптимальное разбиение, максимизирующее I_1, при фиксированных переменных задается формулой:

$$S_k = \{i : corr^2(X_i, f_k) \geq corr^2(X_i, f_q), \quad \forall q = 1,2,..., p'\}$$

и используются итерационные алгоритмы с применение указанных соотношений. Для функционала доказано, что необходимыми и достаточными условиями его максимума являются: 1.разбиение параметров на группы $A_1, A_2, ..., A_{p'}$ таково, что функционал достигает максимума как по разбиению на группы, так и по значениям некоторых коэффициентов, равных либо +1, либо -1 (D(Z) – дисперсия Z);

2. факторы определяются соотношениями:

$$I_3 = \sum_{k=1}^{p'} \sqrt{D(\sum_{i \in S_k} g_i X_i)}$$

$$f_k = \frac{\sum_{i \in S_k} g_i X_i}{\sqrt{\sum_{i \in S_k} g_i g_j corr(X_i, X_j)}}, \quad k = 1,2,..., p'$$

.

При фиксированном разбиении на группы функционал I_1 достигает максимума тогда, когда для каждого k соответствующие коэффициенты g_i максимизируют величину

$$D(\sum_{i \in S_k} g_i X_i).$$

Поэтому при расчётах используют рекуррентную процедуру максимизации.

Идеи метода Э.г.п. близки идеям методов *факторного анализа*. Так, при максимизации функционала I_1 для каждой группы признаков $A_1, A_2, ..., A_p$, строится фактор, имеющий смысл первой гл. компоненты для признаков этой группы. А максимизация функционала I_2 приводит к построению для каждой группы признаков фактора, отличающегося на некоторый множитель от первого общего фактора, который был бы построен для этой группы *центроидным методом*.

См. также *Эвристические методы снижения размерности*, *Экстремальная постановка задач классификации*.

ЭКСПЕРТНО-СТАТИСТИЧЕСКИЙ МЕТОД

построения обобщающего показателя – метод снижения размерности исследуемого признакового пространства до единицы путём формализации зависимости между частными измеряемыми факторами и *латентным* агрегированным *показателем* на основе статистической и экспертной информации. Этот метод используется для решения задач оценки качества анализируемой системы, ранжирования изучаемых объектов по некоторому не поддающемуся

непосредственному измерению свойству, а также в задачах оптимального управления данной системы. Речь может идти, напр., о сравнении стран по прогрессивности их макроструктуры потребления, оценке эффективности деятельности предприятий отрасли и её оптимизации. Пусть характеристика анализируемого интегрального свойства объекта задается латентным показателем и определяется набором частных критериев, задаваемых вектором измеряемых переменных $X = (X_1, X_2, ..., X_p)$.

Базовая идея Э.-с.м. состоит в построении такой целевой функции $f(X)$, которая устанавливает соответствие между статистической информацией об обследованных объектах по переменной *X* и экспертной информацией, относящейся к сравнению этих объектах $O_1, O_2, ..., O_n$ по по переменной *X* и экспертной информацией, относящейся к сравнению этих объектов по анализируемому интегральному свойству. Алгоритм отыскания целевой функции определяется одним из трёх вариантов представления исходной экспертной информации. В первом варианте используется матрица $\|y_{ij}\|_{n\times m}$, содержащая значения бальных оценок искомого показателя для каждого *i*-го объекта каждым *j*-м экспертом. Во втором варианте – т. н. *экспертные упорядочения* объектов по степени проявления в них анализируемого свойства в виде матрицы $\|y_{ij}\|_{n\times m}$, содержащей значения рангов, присвоенных каждому *i*-му объекту каждым *j*-м экспертом. В третьем варианте – набор из *m* булевых матриц парных сравнений $\|\gamma_{ij}^{(k)}\|_{n\times n}$, $k = 1,2,...,m$, содержащих либо отношения предпочтения в виде:

$$\gamma_{ij}^{(k)} = \begin{cases} 1, & O_i \succ = O_i \\ 0, & O_i \prec O_i \end{cases},$$

либо отношения принадлежности объектов к однородному классу в виде:

$$\gamma_{ij}^{(k)} = \begin{cases} 1, & O_i, O_i - \text{однородны}, \\ 0, & O_i, O_i - \text{неоднородны}. \end{cases}$$

В любом случае представления исходных данных производится исследование качества экспертной информации. В первом варианте это сводится к анализу резко выделяющихся наблюдений, во втором – к проверке гипотезы об отсутствии согласованности в упорядочениях различных экспертов, в третьем – к исследованию структуры попарных расстояний между экспертными разбиениями на классы. Исходная статистическая информация однозначно задается в виде матрицы «объект-свойство» $\|y_{ij}\|_{n\times m}$, содержащей значения *p* показателей *n* объектов. Вычислительная реализация Э.-с.м. сводится к обычной схеме регрессионного анализа и к использованию *метода наименьших квадратов* лишь в первом варианте и существенно усложняется во втором и третьем вариантах.

См. также *Целевая функция.*

ЭКСТРЕМАЛЬНЫЕ (ОПТИМАЛЬНЫЕ) СВОЙСТВА ГЛАВНЫХ КОМПОНЕНТ

свойства, позволяющие оптимизировать решения определённых задач *многомерного статистического анализа* и *эконометрики*. К Э.с.г.к. относятся: свойство наименьшей ошибки автопрогноза или наилучшей воспроизводимости, которое используется в *регрессионном анализе*, свойства наименьшего искажения геометрической структуры множества исходных наблюдений при их проектировании в пространство первых главных компонент, используемые в методах группировки данных.

Свойство наименьшей ошибки автопрогноза или наилучшей воспроизводимости состоит в том, что с помощью первых *гл. компонент* $Z_1^*, Z_2^*, ..., Z_{p'}^*$ исходных признаков $X_1, X_2, ..., X_p$ достигается наилучший прогноз этих признаков среди всех прогнозов, построенных с помощью линейных комбинаций $Z_1, Z_2, ..., Z_{p'}$ произвольных признаков. Доказано, что меры ошибки прогноза

$$\|\Delta\| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}\Delta_{ij}}$$

и

$$tr(\Delta) = \sum_{i=1}^{p}\Delta_{ii},$$

где

$$\Delta_{ij} = \mathrm{E}\{(X_i - \sum_{l=1}^{p'} b_{il} Z_l)(X_j - \sum_{l=1}^{p'} b_{jl} Z_l)\},$$

b_{il}, b_{jl}

– МНК-оценки параметров прогноза, достигают одновременно минимума только на векторе ($Z_1^*, Z_2^*, ..., Z_{p'}^*$). При этом

$$\|\Delta\| \approx \sqrt{\sum_{i=p'+1}^{p} \lambda_i^2}\ ,\ tr(\Delta) \approx \sum_{i=p'+1}^{p} \lambda_i\ ,$$

где λ_i – собственное число выборочной *ковариационной матрицы*, построенной по исходным наблюдениям.

К свойствам наименьшего искажения геометрической структуры множества исходных наблюдений при их проектировании в пространство первых гл. компонент относятся следующие три свойства: 1. сумма квадратов расстояний от исходных точек-наблюдений до пространства, натянутого на первые гл. компоненты, наименьшая относительно других подпространств той же размерности, полученных с помощью произвольного линейного преобразования исходных координат; 2. сумма квадратов расстояний между парами точек–наблюдений наименее искажается в подпространстве, натянутом на первые гл. компоненты относительно других подпространств той же размерности, полученных с помощью произвольного линейного преобразования исходных координат; 3. расстояния от исходных точек–наблюдений до их общего «центра тяжести», а также углы между отрезками, соединяющими пары точек–наблюдений с их общим «центром тяжести», наименее искажаются в пространстве, натянутом на первые гл. компоненты относительно других подпространств той же размерности, полученных с помощью произвольного линейного преобразования исходных координат.

См. также *Метод гл. компонент.*

ЭТАПЫ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКОГО МОДЕЛИРОВАНИЯ

последовательные стадии построения и исследования *вероятностно-статистической модели.*

Вероятностно-статистический способ исследования различных явлений и процессов позволяет, с одной стороны, соблюдать математическую строгость при описании их свойств и закономерностей функционирования под воздействием влияния случайных факторов, а, с другой стороны, настраивать получаемые теоретические модели на реально существующие системы, характеризуемые имеющимися эмпирическими данными. Можно выделить осн. Э.в.-с.м. 1-й этап (постановочный) включает в себя определение: конечных прикладных целей моделирования; набора факторов и показателей (переменных), описание взаимосвязей между которыми нас интересует; наконец, роли этих факторов и показателей – какие из них, в рамках поставленной конкретной задачи, можно считать входными (т.е. полностью или частично регулируемыми или хотя бы легко поддающимися регистрации и прогнозу; подобные факторы несут смысловую нагрузку объясняющих в модели), а какие – выходными (эти факторы обычно трудно поддаются непосредственному прогнозу; их значения формируются как бы в процессе функционирования моделируемой системы, а сами факторы несут смысловую нагрузку объясняемых). 2-й этап (априорный, предмодельный) состоит в предмодельном анализе содержательной сущности моделируемого явления, формировании и формализации имеющейся априорной информации об этом явлении в виде ряда *гипотез* и исходных допущений (последние должны быть подкреплены теоретическими рассуждениями о механизме изучаемого явления или, если возможно, экспериментальной проверкой).

3-й этап (информационно-статистический) посвящён сбору необходимой статистической информации, т.е. регистрации значений участвующих в описании модели факторов и показателей на различных временных и (или) пространственных тактах функционирования моделируемой системы. 4-й этап (спецификация

модели) включает в себя непосредственный вывод (опирающийся на гипотезы, принятые на 2-м этапе и исходные допущения) общего вида модельных соотношений, связывающих между собой интересующие нас входные и выходные переменные. Говоря об общем виде модельных соотношений, мы имеем в виду то обстоятельство, что на данном этапе будет определена лишь структура модели, её символическая аналитическая запись, в которой наряду с известными числовыми значениями (представленными в основном исходными статистическими данными) будут присутствовать величины, содержательный смысл которых определен, а числовые значения – нет (их обычно называют параметрами модели, неизвестные значения которых подлежат статистическому оцениванию). 5-й этап (*идентифицируемость* и идентификация модели) предназначен для проведения статистического анализа модели с целью «настройки» значений её неизвестных параметров на исходные статистические данные, которыми мы располагаем. При реализации этого этапа «модельер» должен сначала ответить на вопрос, возможно ли в принципе однозначно восстановить значения неизвестных параметров модели по имеющимся исходным статистическим данным при принятой на 4-м этапе структуре (способе спецификации) модели. Это составляет т.н. проблему идентифицируемости модели. А затем, после положительного ответа на этот вопрос, необходимо решить уже проблему идентификации модели, т.е. предложить и реализовать математически корректную процедуру оценивания неизвестных значений параметров модели по имеющимся исходным статистическим данным. Если проблема идентифицируемости решается отрицательно, то возвращаются к 4-му этапу и вносят необходимые коррективы в решение задачи спецификации модели. 6-й этап (*верификация* модели) заключается в использовании различных процедур сопоставления модельных заключений, оценок, следствий и выводов реально наблюдаемой действительностью. Этот этап называют также этапом статистического анализа точности и адекватности модели. При пессимистическом характере результатов этого этапа необходимо возвратиться к этапу 4, а иногда и к этапу 1.

Построение и экспериментальная проверка (верификация) вероятностно-статистической модели обычно основаны на одновременном использовании информации двух типов: а) априорной информации о природе и содержательной сущности анализируемого явления, представленной, как правило, в виде тех или иных теоретических законномерностей, ограничений, гипотез; б) *исходных статистических данных*, характеризующих процесс и результаты функционирования анализируемого явления или системы.Основой для построения и анализа модели является только априорная информация, что не предусматривает проведения этапов 3 и 5. Тогда модель не является вероятностно-статистической (эконометрической при моделировании экономических закономерностей).

А

АВТОКОРРЕЛЯЦИЯ

корреляционная зависимость между последовательными уровнями временного ряда, обусловленная наличием во временном ряде тенденции и циклических колебаний. Количественно А. можно измерить с помощью линейного коэффициента корреляции между уровнями *временного ряда* отстоящими друг от друга на несколько тактов времени. Одна из рабочих формул для расчёта *коэффициента корреляции* имеет вид:

$$r_{xy} = \frac{\sum_{i=1}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Возьмём в качестве переменной x ряд $y_2, y_3, \ldots, y_n$, а в качестве переменной y – ряд $y_1, y_2, \ldots, y_{n-1}$. Тогда формула примет вид:

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \bar{y}_1)\cdot(y_{t-1} - \bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \bar{y}_1)^2 \cdot \sum_{t=2}^{n}(y_{t-1} - \bar{y}_2)^2}},$$

где

$$\bar{y}_1 = \frac{\sum_{t=2}^{n} y_t}{n-1}; \quad \bar{y}_2 = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1}.$$

Величину r_1 называют коэффициентом автокорреляции уровней ряда первого порядка, т.к. он измеряет зависимость между соседними уровнями ряда t и $t-1$, т.е. при лаге = 1. Аналогично определяются коэффициенты автокорреляции второго и более высоких порядков. Так, коэффициент автокорреляции второго порядка характеризует тесноту связи между уровнями y_t и y_{t-2} и определяется по формуле:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \bar{y}_3)\cdot(y_{t-2} - \bar{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \bar{y}_3)^2 \cdot \sum_{t=3}^{n}(y_{t-2} - \bar{y}_4)^2}},$$

где

$$\bar{y}_3 = \frac{\sum_{t=3} y_t}{n-2}; \quad \bar{y}_4 = \frac{\sum_{t=3} y_{t-2}}{n-2}.$$

Число периодов, по которым рассчитывается коэффициент автокорреляции, называют *лагом*. С увеличением лага число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается. Существует два важных свойства коэффициента автокорреляции: 1. он строится по аналогии с линейным коэффициентом корреляции и таким образом характеризует тесноту только линейной связи текущего и предыдущего уровней ряда. Поэтому по коэффициенту автокорреляции можно судить о наличии линейной (или близкой к линейной) тенденции. Для некоторых временных рядов, имеющих сильную нелинейную тенденцию (напр., параболу второго порядка или экспоненту), коэффициент автокорреляции уровней исходного ряда может приближаться к нулю; 2) по знаку коэффициента автокорреляции нельзя делать вывод о возрастающей или убывающей тенденции в уровнях ряда. Большинство временных рядов экономических данных содержит положительную автокорреляцию уровней, однако при этом могут иметь убывающую тенденцию. Последовательность коэффициентов автокорреляции уровней первого, второго, и т.д. порядков называют *автокорреляционной функцией* временного ряда. График зависимости её значений от величины лага (порядка коэффициента автокорреляции) называется коррелограммой.

АВТОКОРРЕЛЯЦИОННАЯ ФУНКЦИЯ

характеризует корреляционную связь между значениями случайного процесса X(t), наблюдаемыми в различные моменты времени. А.ф. является нормализованным вариантом автоковариационной функции:

$$\rho[X(t), X(t+\tau)] = \operatorname{cov}[X(t), X(t+\tau)] / \sqrt{DX(t)\cdot DX(t+\tau)},$$

где $\operatorname{cov}[X(t), X(t+\tau)] = M[(X(t) - MX(t))\cdot(X(t+\tau) - MX(t+\tau))]$.

При $\tau = 0$ значение автоковариационной функции равно $DX(t)$ – дисперсии случайного процесса X в момент t. Благодаря нормализации А.ф., в отличие от автоковариационной функ-

ции, безразмерна, а её значения, взятые по модулю, не превышают 1. А.ф. стационарного случайного процесса не зависит от значения времени t, а зависит только от сдвига по времени (лага) τ .

А.ф. широко используется при изучении *временных рядов*. Для построения выборочной А.ф. рассчитываются коэффициенты автокорреляции r(τ), совокупность которых и служит оценкой А.ф.:

$$r(\tau) = \frac{\frac{1}{n-\tau}\sum_{t=1}^{n-\tau}(x_t-\bar{x})(x_{t+\tau}-\bar{x})}{\frac{1}{n}\sum_{t=1}^{n}(x_t-\bar{x})^2}, \qquad (1)$$

где n – длина временного ряда $x_1, x_2, \ldots, x_t, \ldots, x_n$; τ – временной сдвиг (τ = 1, 2, …, m); $\bar{x}$ – оценка среднего значения, найденная по формуле:

$$\bar{x} = \frac{\sum_{t=1}^{n} x_t}{n}$$

Числитель выражения (1) представляет выборочную оценку коэффициента автоковариации. Иногда график А.ф., отражающий изменение величины r(τ) в зависимости от значений сдвига τ, называют коррелограммой (correlogram). Очевидно, что с увеличением значения лага τ число пар наблюдений (n–τ), используемых для расчёта в (1), уменьшается. Поэтому в практических руководствах рекомендуется поддерживать соотношение m ≤ n/4.

АВТОКОРРЕЛИРОВАННОСТЬ ОСТАТКОВ

корреляция ошибок регрессии, относящихся к разным моментам времени. Одно из предположений теоремы Гаусса-Маркова, формулирующей условия, при которых оценивание регрессии методом наименьших квадратов (МНК) дает наилучшие линейные несмещённые оценки, заключается в независимости (и, следовательно, в отсутствии автокоррелированности) ошибок регрессии. Отметим, что об А.о. можно говорить только тогда, когда выборочные данные, на которых оценивается регрессия, упорядочены во времени, и, следовательно, упорядочены и регрессионные ошибки. А.о. приводит к тому, что формальное применение МНК дает заниженные оценки дисперсий оценок коэффициентов регрессионного уравнения. Как следствие, t-статистики, вычисляемые как отношение оценок коэффициентов к их стандартным ошибкам, оказываются завышенными, и, т.о., неприменимыми для проверки значимости коэффициентов, т.е. для проверки гипотезы об их равенстве нулю. В связи с этой проблемой разработан ряд критериев для обнаружения А.о. См. также *Критерий Дарбина-Уотсона*.

АВТОРЕГРЕССИЯ

регрессия, определяющая зависимость значений Y_n некоторой случайной последовательности от предыдущих значений $Y_{n-1}, Y_{n-2, \ldots}, Y_{n-p}$. Линейная А. порядка p (autoregression of order p) определяется уравнением:

$$Y_n = \beta_1 Y_{n-1} + \ldots + \beta_p Y_{n-p} + \varepsilon_n \qquad (1),$$

где $\beta_1, \ldots, \beta_p$ – коэффициенты модели; ε_n – одинаково распределённые, некоррелированные случайные величины с нулевым математическим ожиданием и дисперсией σ^2. Уравнение (1) является основой стохастических моделей А., используемых для описания некоторых встречающихся на практике *временных рядов*. Модель А. широко используется для описания стационарных временных рядов, причём рекомендации по идентификации этой модели опираются на анализ *автокорреляционной и частной автокорреляционной функции* временного ряда. На практике при решении социально-экономических задач, как правило, используются авторегрессионные модели невысоких порядков.

См. также *Авторегрессии модель*.

АВТОРЕГРЕССИИ МОДЕЛЬ

модель, которая в математической форме выражает связь текущего значения x_t стационарного временного ряда с его прошлыми значениями $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$,и записываемая в виде

$$x_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots + a_p x_{t-p} + \varepsilon_t \qquad (1),$$

где $a_0, a_1, ..., a_p$ – параметры, подлежащие оцениванию на основе имеющихся статистических данных (наблюдений); ε_t – случайная компонента или ошибка уравнения в момент *t*, *t=1,2,...,T*; *p* – порядок авторегрессии.

Основа для построения А.м. временного ряда – изучение его автокорреляционных свойств. В наиболее общем виде автокорреляционные свойства временного ряда учитываются в смешанной модели авторегрессии-скользящего среднего (АРСС или, англ. ARMA – autoregressive-moving average). В простейшем случае, когда выборочных данных очень мало, буквально несколько наблюдений, на практике часто используют так называемые «наивные модели». Наивная модель 1 предполагает, что «завтра будет то же, что сегодня». Тогда уравнение (1) превращается в $x_t = x_{t-1} + \varepsilon_t$ (2). В этом уравнении лишь один параметр – коэффициент при лаговом значении x_{t-1}, который не оценивается, а постулируется равным 1. В наивной модели 2 предполагается, что «прирост завтра будет таким, как сегодня». Это означает, что с учётом ошибки ε_t: $x_t - x_{t-1} = x_{t-1} - x_{t-2} + \varepsilon_t$ (3), откуда следует: $x_t = 2x_{t-1} - x_{t-2} + \varepsilon_t$ (4). Как видим, опять получено авторегрессионное уравнение, но параметры в нём не оцениваются, а постулируются.

В более сложных случаях, когда требуется одновременно изучать движение многих экономических показателей и исследовать существующие между этими показателями статистические связи, в эконометрических исследованиях используются многомерные или векторные авторегрессионные модели. Уравнение имеет такой же вид, как и (1), но записывается в векторной форме. В этом случае в модели часто оказывается слишком много параметров при небольшом объёме выборки, возникают трудности с оцениванием коэффициентов уравнения. Выход обычно находят в том, что часть параметров задается (как правило, приравнивается нулю) на основе теории или содержательных соображений.

См. также *Бокса-Дженкинса подход*, *Модель авторегрессии со скользящим средним в остатках*.

АДАПТИВНЫЕ МЕТОДЫ ПРОГНОЗИРОВАНИЯ

методы прогнозирования временных рядов на основе построения статистических моделей с автоматической корректировкой параметров и структуры модели во времени – адаптивных моделей. Отличие адаптивных моделей от других прогностических моделей состоит в том, что они отражают текущие свойства ряда и способны непрерывно учитывать эволюцию динамических характеристик изучаемых процессов. Цель А.м.п. заключается в построении самокорректирующихся (самонастраивающихся) математических моделей, которые способны отражать изменяющиеся во времени условия, учитывать информационную ценность различных членов временной последовательности и давать достаточно точные оценки будущих членов данного ряда. Такие модели предназначены прежде всего для краткосрочного прогнозирования, а также для анализа на выборочном периоде долгосрочных тенденций. Отметим различие понятия краткосрочного прогноза в экономике и статистике. В экономике под краткосрочным прогнозом обычно понимают прогноз с периодом упреждения до одного года. В статистике информацию о процессе обычно получают в виде записей значений, наблюденных через равные промежутки времени. Соответственно под краткосрочным прогнозом, как правило, понимается прогноз на один интервал времени (в крайнем случае, на несколько интервалов); сам же интервал может быть любым.

По-видимому, трудно провести чёткую грань, отделяющую А.м.п. от неадаптивных. Уже прогнозирование методом экстраполяции обычных регрессионных кривых содержит некоторый элемент адаптации, когда с каждым новым получением фактических данных параметры регрессионных кривых пересчитываются, уточняются. Однако здесь степень адаптации весьма незначительна; к тому же с течением времени она падает вместе с увеличением общего числа наблюденных точек и соответственно с уменьшением в выборке удельного веса каждой новой точки.

В узком смысле под адаптивной моделью в анализе временных рядов принято понимать модель, процедура корректировки параметров которой основывается на использовании рекуррентных соотношений, и, в частности, формулы экспоненциально-взвешенной скользящей средней (EWMA – Exponentially Weighted Moving Average). Разработано большое число моделей адаптивного типа, способных отражать полиномиальные и экспоненциальные тренды, сезонные колебания аддитивного и мультипликативного типа, эволюционирующие законы распределения вероятностей, линейные и нелинейные регрессионные зависимости, нестационарные корреляционные связи.

Адаптация в данных моделях слагается из небольших дискретных сдвигов. В основе адаптации лежит метод проб и ошибок. Последовательность процесса адаптации в основном выглядит следующим образом. Пусть модель находится в некотором исходном состоянии (т.е. определены текущие значения её коэффициентов) и по ней делается прогноз. Выжидаем, пока истечёт одна единица времени (шаг моделирования), и анализируем, насколько далек прогноз, полученный по модели, от фактического значения ряда. Ошибка прогнозирования через обратную связь поступает на вход системы и используется моделью в соответствии с её логикой (заранее заложенным алгоритмом) для перехода из одного состояния в другое с целью большего согласования своего поведения с динамикой ряда. На изменения динамики ряда модель должна отвечать адекватными изменениями своих параметров. Затем делается прогноз на следующий момент времени, и весь процесс повторяется. Быстроту реакции модели на изменения в движении ряда характеризует т.н. параметр адаптации. Процесс обучения модели состоит в выборе наилучшего параметра адаптации на основе проб на ретроспективном материале. Отметим, что в более сложных моделях сам параметр адаптации может быть переменным, параметров адаптации может быть несколько, когда каждый из них отвечает за адаптацию своей части модели. Помимо корректировки параметров в т.н. комбинированных моделях может предусматриваться и смена структуры прогнозного уравнения путём переключения с одного варианта на другой.

См. также *Анализ временных рядов, Модель Брауна, Модель Брауна обобщённая, Модель адаптивная, Модель авторегрессии-проинтегрированого скользящего среднего в остатках, Модель авторегрессии со скользящим средним в остатках, Параметр адаптации.*

Б

БОКСА-ДЖЕНКИНСА ПОДХОД

систематизированный подход к построению моделей авторегрессии-проинтегрированного скользящего среднего (АРПСС) или в англоязычном варианте Autoregressive Integrated Moving Average (ARIMA) models. В специальной литературе модель ARIMA также известна как модель Бокса-Дженкинса, по имени авторов, разработавших этот подход (Box-Jenkins approach).

Модель ARIMA используется для описания нестационарных временных рядов, обладающих следующими свойствами: 1. ряд включает (аддитивно) составляющую $f(t)$, имеющую вид алгебраического полинома (от времени t) степени k-1 ($k \geq 1$), коэффициенты которого могут носить как стохастический, так и нестохастический характер; 2. ряд, получившийся после применения к нему процедур последовательных разностей, может быть описан моделью ARMA(p, q). В модели ARIMA(p, d, q) параметры p и q определяют, соответственно, порядок авторегрессионной составляющей и порядок скользящего среднего (аналогично модели ARMA(p, q)), а параметр *d* – порядок разности (дискретной производной). Модель Бокса-Дженкинса может быть записана в виде: $\Delta^k y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}$, где $\Delta^k y_t$ – k-я последовательная разность исходного ряда с уровнями y_t; $\alpha_1, \alpha_2, \ldots, \alpha_p, \theta_1, \theta_2, \ldots, \theta_q$ – параметры модели. Такая форма записи предполагает, что временной ряд после взятия последовательных разностей (после дифференцирования k-ого порядка) стал стационарным, удовлетворяющим модели ARMA(p,q) Очевидно, что из модели для $\Delta^k y_t$ легко получить модель для исходного ряда. Б.-

Д.п. к построению ARIMA-модели для исследуемого временного ряда носит итеративный характер и включает следующие основные этапы: идентификация пробной модели; оценивание параметров модели; диагностическая проверка адекватности модели.

Под идентификацией подразумевается использование данных и любой информации для определения подкласса экономичных моделей (с точки зрения количества параметров) для дальнейшего оценивания. На этом этапе определяются значения p, d, q модели ARIMA и получаются предварительные оценки ее параметров. Процесс идентификации содержит две стадии: определение порядка разности для исходного ряда с целью обеспечения стационарности; последующая идентификация модели ARMA для временного ряда, полученного после взятия соответствующих разностей.

Ведущая роль на этом этапе отводится анализу автокорреляционной функции (АКФ) и частной автокорреляционной функции (ЧАКФ).

Под оцениванием понимается эффективное использование данных для получения числовых значений параметров пробной модели, выявленной на предыдущем этапе.

После идентификации модели и оценивания её параметров проводится диагностическая проверка, направленная на анализ согласованности полученной модели с исходными данными (см. рис.1).

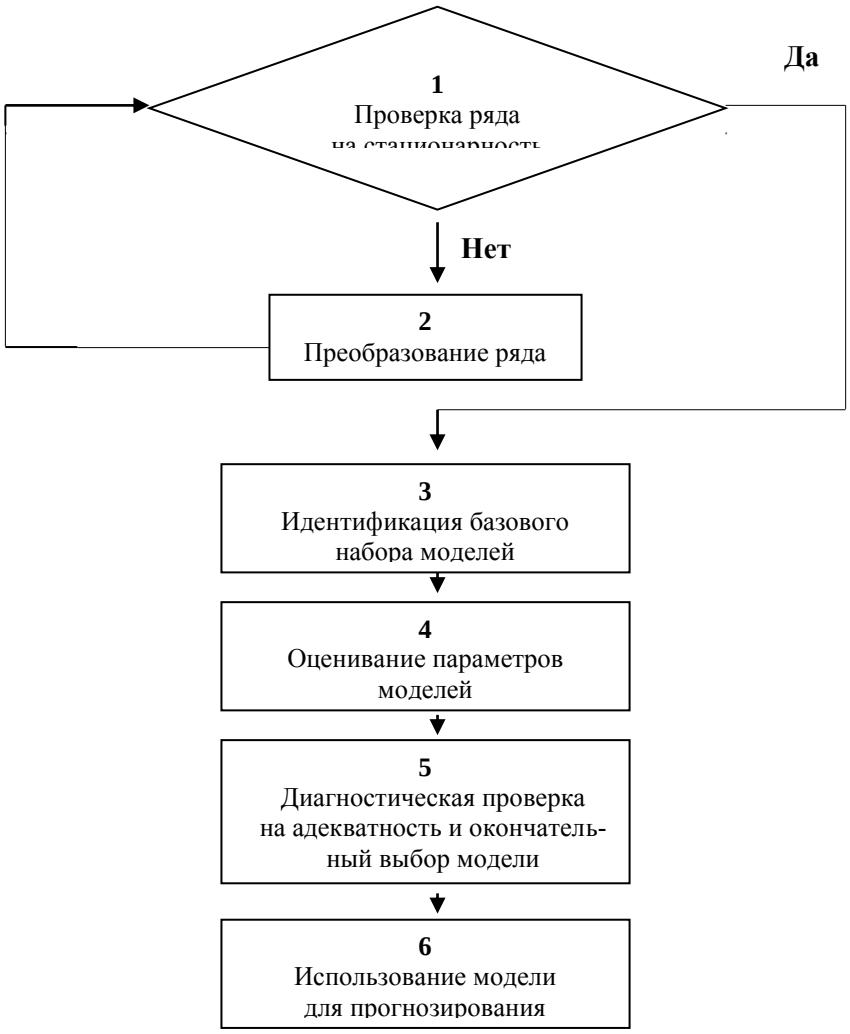


Рис. 1. **Укрупненная схема подбора модели ARIMA**

При «серьёзной» неадекватности модели возникает необходимость её изменения на следующем итеративном цикле. Поясним содержание осн. этапов, представленных на рис. 1

Сначала необходимо получить стационарный ряд. Для перехода к стационарному ряду в Б.-Д.п. рекомендуется применять оператор взятия последовательных разностей (процедуру дискретного дифференцирования). О том, что необходимая для стационарности ряда степень разности достигнута, будет свидетельствовать быстрое затухание АКФ. В практических исследованиях d, как правило, не превышает 2. После получения стационарного ряда исследуется характер поведения выборочных АКФ и ЧАКФ и выдвигаются гипотезы о значениях параметров p (порядок авторегрессии) и q (порядок скользящего среднего).

При этом следует иметь в виду, что выборочные корреляционные функции могут не демонстрировать детального сходства с теоретическими. Напр., умеренно большие значения выборочной АКФ могут наблюдаться после затухания теоретической функции, а также могут наблюдаться всплески, не имеющиеся в теоретической функции. Поэтому для идентификации модели могут использоваться главные черты АКФ, при расхождении более тонких деталей. В результате может быть сформирован базовый набор, включающий 1–2 или даже большее число моделей (см. табл. 1).

Таблица 1

Свойства автокорреляционных и частных автокорреляционных функций

Функция	ARMA(1,0)	ARMA(2,0)	ARMA(0,1)	ARMA(0,2)	ARMA(1,1)
АКФ	Экспоненциально затухает (монотонно или знакопеременно)	Экспоненциально затухает или имеет форму синусоидаль-ной волны	Выброс (пик) на лаге 1	Выбросы (пики) на лагах 1,2	Экспоненциально затухает от значения при лаге 1 (монотонно или знакоперемен-но)
ЧАКФ	Выброс (пик) на лаге 1	Выбросы (пики) на лагах 1,2	Экспоненциально затухает (монотонно или знакопере-менно)	Экспоненциально затухает или имеет форму синусоидаль-ной волны	Экспоненциально убывает от значения при лаге 1 (монотонно или знакоперемен-но)

Исследования показывают, что при использовании в экономических задачах модели ARMA(p,q), потребностям практики, как правило, удовлетворяют пять видов этой модели со свойствами автокорреляционных функций (АКФ и частной АКФ) для этих моделей, представленные в табл.1. После осуществления идентификации моделей необходимо оценить их параметры. В совр. эконометрических пакетах используются разные подходы (*метод наименьших квадратов* (МНК), нелинейный МНК, *метод макс. правдоподобия* (ММП)). Все эти оценки при больших объёмах выборок асимптотически эквивалентны.

Для проверки каждой пробной модели на адекватность анализируется ряд остатков. У адекватной модели остатки должны быть похожими на *белый шум*, т.е. их выборочные автокорреляции не должны существенно отличаться от нуля. При проверке значимости коэффициентов АКФ используются два подхода: проверка значимости каждого коэффициента автокорреляции отдельно; проверка значимости множества коэффициентов автокорреляции как группы. Если модель адекватна исходным данным и ошибки являются *белым шумом*, то распределение коэффициентов автокорреляции приближается к нормальному с нулевым математическим ожиданием и дисперсией $1/n$. Поэтому

если выборочный коэффициент автокорреляции r_k выходит за интервал

$$\pm\frac{t_\gamma}{\sqrt{n}},$$

то нулевая гипотеза о равенстве нулю коэффициента ρ_k отвергается.

Другой подход опирается на Q-статистику Бокса-Пирса, позволяющую проверить равенство нулю сразу τ первых значений АКФ остатков. Отметим, что в некоторых совр. эконометрических пакетах включена модификация этого подхода, опирающаяся на статистику Льюнга-Бокса.

Кроме того, при построении модели ARIMA необходимо проверить значимость коэффициентов (по t-критерию). При этом модель не должна содержать лишних параметров, т.е. уменьшение числа параметров будет способствовать появлению значимой автокорреляции остатков. Если в результате проверки несколько моделей оказываются адекватны исходным данным, то при окончательном выборе следует учесть два требования: повышение точности (качество подгонки модели); уменьшение числа параметров модели.

Совр. подходы построения модели arima используют для этого *информационный критерий Акайке* (Akaike information criterion, aic) и байесовский *информационный критерий Шварца*. С помощью окончательно выбранной модели строится точечный и интервальный прогноз на L шагов вперед. Сезонная модель Бокса-Дженкинса представлена в виде: Arima(p, d, q) (p_s,d_s,q_s), где к параметрам модели p, d, q добавлены сезонные параметры: p_s –сезонный параметр авторегрессии; q_s – сезонный параметр скользящего среднего; d_s – порядок сезонной разности (сезонной производной). При наличии ярко выраженной сезонной компоненты целесообразно включение в модель сезонного дифференцирования. Определение значений параметров сезонной авторегрессии sar(p_s) и сезонного скользящего среднего sma(d_s) также опирается на исследование АКФ и ЧАКФ. Только теперь все типичные проявления, всплески будут удалены друг от друга на величину лага s, где s – период сезонности. Во многих современных эконометрических пакетах реализованы процедуры автоматического подбора вида модели Бокса-Дженкинса. Успех применения мощного, гибкого, но в то же время сложного аппарата модели arima во многом зависит от практического опыта и квалификации исследователя, а процедуры автоматического выбора вида модели призваны лишь облегчить его аналитическую деятельность. См. также *Модель авторегрессии со скользящим средним в остатках, Статистика Бокса-Пирса*.

БЕЛЫЙ ШУМ

стационарный случайный процесс X(t), ординаты $X(t_1)$, $X(t_2)$,… которого независимы для любого множества t_1, t_2, … допустимых значений аргумента t. Название «Б.ш.» объясняется аналогией с белым светом, который представляет собой сумму всех спектральных составляющих, имеющих одинаковую интенсивность. Б.ш. можно представить в виде суммы колебаний всех частот спектра, амплитуды которых являются случайными величинами с одинаковыми дисперсиями. Вследствие бесконечного числа спектральных составляющих в непрерывном временном представлении дисперсия белого шума бесконечна, и в качестве характеристики его интенсивности используется спектральная плотность.

Непрерывный случайный процесс является Б.ш., если его *математическое ожидание* равно нулю, а ковариация его произвольных отсчётов $\operatorname{cov}\big(X(t_1),X(t_2)\big)=\xi\delta\big(t_1,t_2\big)$

с точностью до постоянной величины ξ, называемой его спектральной плотностью, представляет собой дельта-функцию Дирака

$$\delta\big(t_1,t_2\big):\ \delta\big(t_1,t_2\big)=\begin{cases}\infty, & t_1=t_2,\\ 0, & t_1\neq t_2\end{cases}.$$

В дискретном случае Б.ш. может считаться случайный процесс с равномерным спектром, ширина которого не меньше величины, обратной наименьшему временному интервалу между соседними отсчётами, что обеспечивает независимость любых значений этого процесса.

Ковариация временных отсчётов Б.ш. $\operatorname{cov}\big(X(t_1),X(t_2)\big)=\sigma^2\delta\big(t_1,t_2\big)$, где

$$\delta(t_1,t_2)=\begin{cases}1, & t_1=t_2,\\ 0, & t_1\neq t_2\end{cases}$$

– символ Кронекера, σ^2 – дисперсия, которая является конечной величиной при конечной ширине спектра шума.

Б.ш. используется в статистических исследованиях в качестве модели идеальных остатков классической линейной регрессионной модели и других видах статистического анализа для описания компонентов, не поддающихся описанию в виде детерминированной функции.

БИНАРНЫЕ ПЕРЕМЕННЫЕ (БУЛЕВЫЕ, ДИХОТОМИЧЕСКИЕ)

дискретные переменные, принимающие только два значения, напр., (0,1), (да, нет) и т.п.

Как правило, независимые переменные в регрессионных моделях имеют «непрерывные» области изменения. Однако теория не накладывает никаких ограничений на характер регрессоров, в частотности, некоторые переменные могут принимать всего два значения. Необходимость рассматривать такие переменные возникает довольно часто, если приходится принимать во внимание какой-либо качественный признак.

Б.п. имеют большое значение в анализе данных, поскольку многие задачи могут быть сведены к двоичной классификации. Кроме того, такие переменные используются в качестве индикаторных в регрессии, с их помощью представляются признаки исследуемых объектов и процессов, имеющие только два возможных значения.

Во множественной линейной регрессии такие переменные могут быть разного рода атрибутивные признаки. Б., фиктивные п. будучи экзогенными, не создают каких-либо трудностей при применении *метода наименьших квадратов*. Они являются эффективным инструментом построения регрессионных моделей и проверки гипотез. Для того чтобы ввести такие переменные в регрессионную модель, им должны быть присвоены те или иные цифровые метки, т.е. качественные переменные необходимо преобразовать в количественные. Напр., пол клиента: в результате опроса группы людей 0 может означать, что опрашиваемый – мужчина, а 1 – женщина. В принципе можно оценивать соответствующие уравнения внутри каждой категории, а затем изучать различия между ними, но введение Б.п. позволяет оценить уравнение сразу по обеим категориям

Однако, когда Б.п. являются эндогенными переменными, возникают определённые проблемы в оценке параметров модели. Данные модели не могут быть оценены с помощью метода наименьших квадратов, т.к. нарушаются практически все предпосылки классической модели линейной регрессии, поэтому требуются специальные методы оценивания. Существует развитый комплекс аналитических моделей, работающих с бинарной выходной переменной, напр., *пробит-* и *логит-модели*, логистическая регрессия.

В

ВОССТАНОВЛЕНИЕ ПРОПУЩЕННЫХ НАБЛЮДЕНИЙ

механизм заполнения пропусков в неполных статистических данных. На сегодняшний день в *математической статистике* существует несколько путей решения проблемы неполных данных: метод исключения некомплектных объектов. При отсутствии у ряда объектов значений каких-либо переменных некомплектные объекты удаляются из анализа. Подход легко реализуется и может быть удовлетворительным при малом числе пропусков. Однако иногда он приводит к серьезным смещениям и обычно не очень эффективен. Гл. недостаток такого подхода обусловлен потерей информации при исключении неполных наблюдений, так как неполные данные несут в себе новую информацию, необходимую для исследования, и поэтому их важно включать в анализ; методы с заполнением. При данном подходе пропущенные значения исходной выборки заполняются и полученные «полные» данные обрабатываются обычными методами. Наиболее часто используются процедуры заполнения пропусков: заполнение средними, заполнение с пристрастным подбором, подстановка с подбором внутри групп и подбор ближайшего соседа, заполнение

с помощью регрессии, заполнение без подбора, многократное заполнения, составные и другие методы.

Методы взвешивания изменяют веса, чтобы учесть отсутствие значений. Взвешивание связано с заполнением средними.

Методы, основанные на моделировании: широкий класс методов основывается на построении модели порождения пропусков. Выводы получают с помощью функции правдоподобия, построенной при условии справедливости этой модели, с оцениванием параметров методами типа максимального правдоподобия. В методах, использующих функцию правдоподобия, реализована относительно старая идея обработки неполных данных: заполнение пропусков оценками пропущенных значений; оценивание параметров; повторное оценивание пропущенных значений (оценки параметров считаются точными); повторное оценивание параметров и так далее до сходимости процесса. Преимущества такого подхода состоят в том, что он гибок; позволяет отказаться от методов, разработанных для частных случаев; позволяет оценивать в приближении большой выборки дисперсии оценок с помощью матрицы вторых производных функций правдоподобия для неполных данных; обеспечивает надежную сходимость, т.е. в определенных нестрогих условиях каждая итерация увеличивает логарифм правдоподобия и последовательность сходится к некоторому стационарному значению. Недостаток алгоритма заключается в том, что скорость сходимости может быть очень низкой, если пропущено много данных.

Г

ГЕТЕРОСКЕДАСТИЧНОСТЬ

(от лат. – heteroskedastic – неоднородность) – свойство дисперсии остатков ε, когда для каждого значения факторов x_i остатки ε имеют различную дисперсию. При этом оценки коэффициентов регрессии, полученные *методом наименьших квадратов* (МНК–оценки) перестают быть эффективными. С целью обнаружения Г. в остатках строиться график зависимости остатков ε_i от теоретических значений результативного признака $\hat{y}_x$. Если на графике получена горизонтальная полоса, то остатки ε_i представляют собой случайные величины и применение метода наименьших квадратов для оценки параметров уравнения регрессии оправдано, теоретические значения $\hat{y}_x$ хорошо аппроксимируют фактические значения y. Если же ε_i зависит от x_i, то возможны следующие случаи. Остатки ε_i: не случайны; не имеют постоянной дисперсии; носят систематический характер. В этих случаях необходимо применять другие методы оценивания и анализа.

ГОМОСКЕДАСТИЧНОСТЬ

(от лат. – homoscedastic– однородность) – свойство данных, на основе которых строится регрессионная модель; заключается в том, что разброс точек наблюдений относительно плоскости регрессии является равномерным на всем диапазоне изменения независимых переменных: $\sigma_i^2 = \sigma^2 = const$, *для всех* $i = 1,2,...,n$.

Г. данных – одно из требований применимости регрессионной модели, построенной на их основе. Одно из предположений классической регрессионной модели состоит в том, что случайные ошибки некоррелированы между собой и имеют постоянную дисперсию. Т.е. не должно быть априорной причины, вызывающей большую ошибку (отклонение) при одних наблюдениях и меньшую – при других. В тех случаях, когда наблюдаемые объекты однородны, не сильно отличаются друг от друга, такое допущение оправдано и для оценки параметров модели применим метод наименьших квадратов. Однако во многих случаях такое предположение нереалистично. Это означает, что дисперсия зависимых величин (а следовательно и случайных ошибок) не постоянны. Это явление в эконометрике называется *гетероскедастичностью*. На практике гетероскедастичность не так уж и редка. Зачастую есть основание считать, что вероятностное распределение случайных отклонений ε_i при различных наблюдениях будут различны. Это не означает, что случайные отклонения обязательно будут большими при определенных наблюдениях и малыми –

при других, но это означает, что вероятность этого велика.

Д

ДВУХШАГОВЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (2МНК)

метод оценки параметров отдельного уравнения *системы одновременных уравнений* (СОУ). Сущность метода состоит в том, что для оценивания параметров структурного уравнения, *метод наименьших квадратов* (МНК) применяют в два этапа. Он даёт состоятельные, но в общем случае смещённые оценки коэффициентов уравнения, является достаточно простым с теоретической точки зрения и удобным для вычисления. Запишем исходное i -е структурное уравнение системы в виде: $y_i = Y_i\beta_i + X_i\gamma_i + \varepsilon_i$, где y_i – вектор n наблюдений над i -й *эндогенной переменной*; Y_i – матрица порядка ($n \times q_i$) значений эндогенных переменных, входящих в i-е уравнение (кроме y_i -й); β_i – вектор размерности ($q_i \times 1$) значений структурных коэффициентов эндогенных переменных из матрицы Y_i; X_i – матрица порядка $(n \times k_i)$ значений *экзогенных переменных*, входящих в уравнение; γ_i – вектор размерности $(k_i \times 1)$ коэффициентов, относящихся к переменным X_i; ε_i – вектор случайных возмущений, имеющий размерность $(n \times 1)$, причём $M\varepsilon_i = 0$; $\Sigma_{(\varepsilon)} = \sigma_i^2 E_n$.

Непосредственно применить в данном случае МНК нельзя, т.к. эндогенные переменные, содержащиеся в матрице Y_i коррелированы со случайными составляющими ε_i.

Поэтому представляются эндогенные переменные Y_i, входящие в уравнение, как функция всех содержащихся в модели экзогенных переменных (X). Находится оценка $\hat{Y}_i$ матрицы Y_i, которая согласно МНК определяется из выражения: $\hat{Y}_i = X_i(X_i^T X_i)^{-1} X_i^T Y_i$. Тогда $Y_i = \hat{Y}_i + \hat{U}$, где $\hat{U}$ – матрица оценок остаточных величин преобразованной системы. Исходное структурное уравнение может быть преобразовано к виду: $y_i = \hat{Y}_i\beta_i + X_i\gamma + \nu_i$, где $\nu_i = \varepsilon_i + \hat{U}\beta_i$.

При применении МНК, для нахождения оценок параметров вновь полученного уравнения регрессии имеется:

$$d = \begin{pmatrix} \hat{\beta}_i \\ \hat{\gamma}_i \end{pmatrix} = \begin{pmatrix} \hat{Y}_i^T\hat{Y}_i & \hat{Y}_i^T X_i \\ X_i^T\hat{Y}_i & X_i^T X_i \end{pmatrix}^{-1} \begin{pmatrix} \hat{Y}_i^T y_i \\ X_i^T y_i \end{pmatrix},$$

где, d – вектор оценок коэффициентов размерности $((q_i + k_i) \times 1)$. Переходя к исходным переменным, получается:

$$d = \begin{pmatrix} Y_i^T X_i \\ X_i^T X_i \end{pmatrix} \begin{pmatrix} Y_i^T X_i (X_i^T X)^{-1} X_i^T y_i \\ X_i^T y_i \end{pmatrix} .$$

И

ИДЕНТИФИЦИРУЕМОСТЬ

возможность вычисления структурных коэффициентов *систем одновременных уравнений* (СОУ) по коэффициентам приведённой формы. Проблема И. является центральной при работе с СОУ. Уравнение структурной формы эконометрической модели точно идентифицируемо, если все участвующие в нем неизвестные коэффициенты однозначно восстанавливаются по коэффициентам приведенной формы без каких-либо ограничений на значения последних. Уравнение структурной формы (УСФ) называется неидентифицируемым, если хотя бы один из участвующих в нем неизвестных коэффици-

ДИХОТОМИЧЕСИКИЕ (БИНАРНЫЕ) ПЕРЕМЕННЫЕ

см. в ст. *Бинарные переменные (булевые, дихотомические)*

ентов не может быть восстановлен по коэффициентам приведённой формы. УСФ называется сверхидентифицируемым, если все участвующие в нем неизвестные коэффициенты восстанавливаются по коэффициентам приведенной формы, причем некоторые из исходных параметров могут принимать одновременно несколько числовых значений, соответствующих одной и той же приведенной форме.

Если сверхидентифицируемость – это проблема количества наблюдений: с увеличением объёма выборки все различные состоятельные оценки параметра стремятся к одному и тому же истинному значению, то *неидентифицируемость* – это проблема структуры модели.

См. также *Идентифицируемость модели*.

ИДЕНТИФИЦИРУЕМОСТЬ МОДЕЛИ

При анализе *модели эконометрической*, представленной системой уравнений вида

$$\begin{cases} \mathbf{B}_1 Y_t^{(1)} + \mathbf{B}_2 Y_t^{(2)} + \mathbf{C}_1 X_t = \Delta_t \\ \mathbf{B}_3 Y_t^{(1)} + \mathbf{B}_4 Y_t^{(2)} + \mathbf{C}_2 X_t = 0, \quad t = 1,2,\ldots,n, \end{cases} \qquad (1)$$

или $\mathbf{B}Y_t + \mathbf{C}X_t = \bar{\Delta}_t, \quad t = 1,2,\ldots,n,$ (1'), исследователя в конечном счёте интересует прежде всего поведение *эндогенных переменных* Y_t. Из соответствующей приведённой формы модели:

$$Y_t = -\mathbf{B}^{-1}\mathbf{C}X_t + \mathbf{B}^{-1}\bar{\Delta}_t, \quad t = 1,2,\ldots,n, \qquad (2)$$

видно, что эндогенные переменные Y_t – по своей природе случайные величины, поведение которых определяется внутренней структурой модели, а именно элементами матриц B и C и природой случайных остатков Δ_t. Возникает вопрос возможности во время следования в «обратном направлении», восстановить структурную форму (1'), т.е. всех элементов матриц B и C, располагая знанием значений коэффициентов приведённой формы (2), т.е. знанием числовых значений всех элементов матрицы $\mathbf{\Pi}$ и природы случайных остатков ε_t. Именно этот вопрос и отражает сущность проблемы И.м. эконометрической (нельзя смешивать с проблемой идентификации модели, заключающейся в выборе и реализации методов статистического оценивания её неизвестных параметров).

Ответ на поставленный вопрос в общем случае, очевидно, отрицательный: без дополнительных ограничений на внутреннюю структуру модели (т.е. без соблюдения некоторых условий *идентифицируемости*) по $m_1 \times (p+1)$ элементам матрицы $\mathbf{\Pi}$ невозможно восстановить гораздо большее число элементов матриц **B** и **C** (нетрудно подсчитать, что общее число коэффициентов β_{ij} и c_{lk} в структурной форме равно $m_1 \times (m_1 + m_2 + p + 1)$ хотя, конечно, общее число коэффициентов, подлежащих статистическому оцениванию, оказывается меньшим).

В эконометрической теории приняты следующие определения, связанные с проблемой идентифицируемости *системы одновременных уравнений* (СОУ): 1. уравнение структурной формы эконометрической модели называется точно идентифицируемым, если все участвующие в нём неизвестные (т.е. априори не заданные) коэффициенты однозначно восстанавливаются по коэффициентам приведённой формы без каких-либо ограничений на значения последних; 2. эконометрическая модель называется точно идентифицируемой, если все уравнения её структурной формы – точно идентифицируемы; 3. уравнение структурной формы называется сверхидентифицируемым, если все участвующие в нём неизвестные коэффициенты восстанавливаются по коэффициентам приведённой формы, причём некоторые из его коэффициентов могут принимать одновременно несколько (более одного) числовых значений, соответствующих одной и той же приведённой форме; 4. уравнение структурной формы называется неидентифицируемым, если хотя бы один из участвующих в нём неизвестных коэффициентов не может быть восстановлен по коэффициентам приведённой формы. Соответственно модель называется неидентифицируемой, если хотя бы один из коэффициентов структурной формы является неидентифицируемым.

В проблеме И.м. исследователя, в конечном счёте, интересует поведение эндогенных переменных, и с этой точки зрения может показаться несущественной, более того, надуманной проблема «однозначного возврата» от приведённой формы к структурной. Однако в дей-

ствительности исследователя могут интересовать оценочные значения коэффициентов именно структурной формы как имеющие прозрачную экономическую интерпретацию (различные эластичности, мультипликаторы и т. п.). Именно поэтому проблема идентифицируемости крайне важна с позиций выработки предложений по решению проблемы идентификации эконометрической модели, т.е. проблемы выбора и реализации методов статистического оценивания участвующих в ней неизвестных параметров.

Решение проблемы идентификация предусматривает «настройку» модели, записанной в общей структурной форме (1'), на реальные статистические данные:

$$Y = \begin{pmatrix} Y_i^{\mathrm{T}} \\ \vdots \\ Y_n^{\mathrm{T}} \end{pmatrix} \quad \text{и} \quad X = \begin{pmatrix} Y_i^{\mathrm{T}} \\ \vdots \\ Y_n^{\mathrm{T}} \end{pmatrix} \qquad (3).$$

Другими словами, речь идёт о выборе и реализации методов статистического оценивания неизвестных параметров модели (1) (т.е. той части элементов матриц B и **C**, значения которых не являются априори известными) по исходным статистическим данным (3).

Проблема верификация модели, так же, как и проблема идентификации, является специфичной, связанной с построением именно эконометрической модели. Собственно построение эконометрической модели завершается её идентификацией, т.е. статистическим оцениванием участвующих в ней неизвестных коэффициентов (параметров) b_{ij} и c_{lk}. После этого, однако, возникают вопросы: насколько удачно удалось решить проблемы спецификации, идентифицируемости и идентификации модели, т.е. можно ли рассчитывать на то, что использование построенной модели в целях прогноза эндогенных переменных и имитационных расчётов, определяющих варианты социально-экономического развития анализируемой системы, даст результаты, достаточно адекватные реальной действительности; какова точность (абсолютная, относительная) прогнозных и имитационных расчётов, основанных на построенной модели? Получение ответов на эти вопросы с помощью тех или иных математико-статистических методов и составляет содержание проблемы верификации эконометрической модели.

ИНСТРУМЕНТАЛЬНЫХ ПЕРЕМЕННЫХ МЕТОД

см. в ст. Метод инструментальных переменных

ИНТЕРПРЕТАЦИЯ ФАКТОРОВ

анализ величин и знаков факторных нагрузок на исследуемые признаки с целью определения смысловой сущности выделенных общих факторов в *факторном анализе.* Геометрическая интерпретация помогает выбрать вращение системы общих факторов, наиболее подходящее в отношении возможности их содержательной интерпретации. Параметры модели факторного анализа, в том числе и сами общие факторы $f^{(1)}, f^{(2)}, \ldots, f^{(p')}$, определяются не однозначно, а лишь с точностью до некоторого ортогонального преобразования, т.е. с точностью до вращения осей $f^{(1)}, f^{(2)}, \ldots, f^{(p')}$ в пространстве. При этом выбор окончательного решения, т.е. закрепление системы $f^{(1)}, f^{(2)}, \ldots, f^{(p')}$ в определенном положении, находится в распоряжении исследователя. Другими словами, исследователь должен решить вопрос: как выбрать такое преобразование, такой поворот осей $f^{(1)}, f^{(2)}, \ldots, f^{(p')}$, при котором получаемые новые общие факторы $\tilde{f}^{(1)}, \tilde{f}^{(2)}, \ldots, \tilde{f}^{(p')}$ допускают наиболее естественную и убедительную интерпретацию. Типичные стратегии вращения – варимакс, биквартимакс, квартимакс и эквимакс, являются ортогональными преобразованиями, обеспечивающими максимально возможную концентрацию дисперсии исходных данных на координатных осях выделенных факторов. Также существуют косоугольные (не ортогональные) преобразования, в результате которых факторы располагаются не вполне перпендикулярно друг другу, но достигается более простая структура факторных нагрузок – проекций факторных осей на исходные оси координат. Однако не всегда удается легко интерпретировать и косоугольные факторы из-за

перекрестных нагрузок, возникающих в силу их коррелированности. Факторный анализ и *метод главных компонент* являются сугубо количественными методами и формально позволяют получить результат (выделить факторы) практически при любом составе исходных показателей независимо от того, существует объективно этот фактор или нет. Однако выделенные факторы не носят очевидный содержательный характер, то нередко ставится под сомнение справедливость полученных выводов и возможность использования в практической деятельности результатов факторного анализа.

ИНФОРМАЦИОННЫЙ ЭТАП ИССЛЕДОВАНИЯ

сбор необходимой статистической информации для процесса эконометрического моделирования, т.е. регистрация значений участвующих в модели факторов и показателей на различных временных или пространственных тактах функционирования изучаемого явления. При составлении плана сбора исходной статистической информации необходимо по возможности учитывать полную схему дальнейшего статистического анализа. Априорное представление о том, как и для чего данные будут использоваться, может оказать существенное влияние на их сбор.

При планировании особого внимания заслуживают случаи, когда: а) используется аппарат теории выборочных обследований, т.е. определяется, какой должна быть выборка – случайной, пропорциональной, расслоенной и т.п.; б) хотя бы для части исходных переменных эксперимент носит активный характер, т.е. переменные допускают фиксацию в каждом конкретном наблюдении на определенном уровне, и выбор плана обследования осуществляется с привлечением метода планирования экспериментов. В некоторых руководствах по статистике этот этап называют этапом «организационно-методической подготовки». При вводе *исходных статистических данных* в вычислительное устройство одновременно вносятся полные и краткие (для автоматизированного воспроиз-ва в табл.) определения используемых терминов. Не зависимо от того, производится ли исследователем выбор метода и плана статистического обследования или он уже располагает результатами т.н. пассивного эксперимента, к моменту определения основного статистического инструментария исследователь в общем случае располагает в качестве массива исходных статистических данных матрицами наблюдений одного из видов: 1. пространственно-временная выборка

$$(\text{и.с.д.})_1 = \begin{pmatrix} x_1^{(1)}(t) & x_1^{(2)}(t) & \dots & x_1^{(\text{p})}(t) \\ x_2^{(1)}(t) & x_2^{(2)}(t) & \dots & x_2^{(\text{p})}(t) \\ \dots & \dots & \dots & \dots \\ x_{\text{n}}^{(1)}(t) & x_{\text{n}}^{(2)}(t) & \dots & x_{\text{n}}^{(\text{p})}(t) \end{pmatrix}, \quad t = t_1, t_2, \dots, t_N,$$

где $x_i^{(\text{j})}(t_k)$ – значение j-го анализируемого признака, характеризующего состояние i-го объекта в момент времени t_k. На каждом из n объектов регистрируются значения p характеризующих его признаков в N последовательные моменты времени $t_1, t_2, \dots, t_N$. 2. пространственная статическая выборка

$$(u.c.\partial.)_1 = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \dots & \dots & \dots & \dots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{pmatrix},$$

$$(\text{и.с.д.})_2 = \begin{pmatrix} \gamma_{21}(t), & \gamma_{22}(t), & \ldots, & \gamma_{2m}(t), \\ \ldots & \ldots & \ldots & \ldots \\ \gamma_{m1}(t), & \gamma_{m2}(t), & \ldots, & \gamma_{mm}(t), \end{pmatrix}, \quad \begin{pmatrix} m = n \text{ или } p, \\ t = t_1, \ldots, t_N \end{pmatrix}.$$

где $N = 1$, так называемые одномоментные наблюдения, тогда для сокращения обозначений индекс времени t можно опускать. 3. временная последовательность матриц парных сравнений размера $n \times n$ (если рассматриваются характеристики парных сравнений объектов) или $p \times p$ (если рассматриваются характеристики парных сравнений признаков): В статическом варианте, т.е. при N = 1, исследователь располагает лишь одной матрицей парных сравнений (γ_{ij}), описывающей ситуацию в один какой-то фиксированный момент времени.

К

КОЛИЧЕСТВЕННЫЕ ПЕРЕМЕННЫЕ

переменные, числовые значения которых получены в процессе измерения по интервальной шкале или шкале отношений. К К.п. относятся интервальная переменная и переменная отношений. Обычно в статистике различают три типа значений переменных: К.п., номинальные и ранговые.

Количественные данные получают при измерениях (напр., данные о весе, размерах, температуре, времени, результатах тестирования и т. п.). Их можно распределить по шкале с равными интервалами. Значения К.п. являются числовыми, могут быть упорядочены и для них имеют смысл различные вычисления (например, среднее значение). На обработку К.п. ориентировано подавляющее большинство статистических методов. Но даже к количественным данным такие методы можно применить лишь в том случае, если число этих данных достаточно велико.

КОЭФФИЦИЕНТ ЭЛАСТИЧНОСТИ

логарифмическая производная результирующего признака y по фактору x:

$$k_Э = \frac{\partial \ln y}{\partial \ln x}.$$

К.э. приближенно характеризует относительное изменение одного признака при единичном относительном изменении другого:

$$k_Э \approx \frac{\Delta y / y}{\Delta x / x},$$

где Δy – прирост показателя y, Δx – прирост показателя x. К.э. показывает, на сколько процентов изменится y при изменении x на один процент.

Вследствие наглядной интерпретации К.э. – важный инструмент экономического анализа. В линейной регрессионной модели

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k$$

средний К.э. $\bar{k}_{Эi}$ пропорционален коэффициенту b_i с коэффициентом пропорциональности, равным отношению средних значений регрессора $\bar{x}_i$ и зависимой переменной $\bar{y}$. Параметры показательных регрессионных моделей представляют собой К.э. по соответствующим факторам. В *производственной функции Кобба-Дугласа,* широко используемой в практике статистических исследований параметры модели равны К.э. по факторам произ-ва.

Л

ЛАГ

(от англ. lag – запаздывание) – время (временной интервал), за которое изменение аргумента приведёт к изменению результативного показателя.

В экономике много примеров, когда влияние изменений одной переменной на другую сказывается через какой-то промежуток времени. Напр., затраты компании на рекламу, проведение маркетинговых исследований в предыдущие моменты времени отражаются на объёме реализации текущего периода. Также существует определённая задержка во времени между капитальными вложениями и вводом в строй осн. фондов. Объём сбережений зависит от

уровня доходов не только в текущий период времени, но и в предыдущие периоды. В эконометрических моделях эндогенная переменная в текущий момент времени может зависеть не только от текущих, но и от лаговых значений экзогенных переменных, также правая часть модели может содержать *лагированную* зависимую *переменную*.

Л. выступает в качестве аргумента *автокорреляционной функции*, значение которой для Л. τ равно коэффициенту корреляции между величинами X(t) и X(t-τ), разделёнными временным интервалом τ, где X(t) – стационарный случайный процесс.

ЛАГИРОВАННАЯ ПЕРЕМЕННАЯ (ЛАГОВАЯ ПЕРЕМЕННАЯ, ЗАПАЗДЫВАЮЩАЯ ПЕРЕМЕННАЯ)

переменная, рассматриваемая в модели в предыдущие, прошлые моменты времени, т.е. с некоторым запаздыванием, *лагом*. При анализе и моделировании экономических процессов часто приходится рассматривать ситуации, когда значение результирующего признака в текущий момент времени t зависит от значений переменных в предыдущие моменты времени t-1, t-2, … , t-τ. При этом возможны различные типы моделей. Напр., значение эндогенной переменной в текущий момент времени может зависеть не только от текущих, но и от лаговых значений экзогенных переменных (*модель распределённых лагов*):

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_k x_{t-k} + \varepsilon_t,$$

где ε_t – последовательность гомоскедастичных и взаимно не коррелированных регрессионных остатков, $\alpha, \beta_0, \beta_1, \ldots, \beta_k$ – неизвестные коэффициенты модели. При этом предполагается, что две переменные (x и y) связаны так, что наблюдается распределенный во времени эффект воздействия изменения одной из них (x) на другую (y).

Другой тип модели предполагает, что в правой части может стоять Л.п. зависимая, т.е. значение эндогенной переменной в текущий момент времени может зависеть от значений той же переменной в предыдущие моменты времени, напр.,

$$y_t = \alpha + \beta_0 x_t + \beta_1 y_{t-1} + \ldots + \beta_k y_{t-k} + \varepsilon_t.$$

Возможно сочетание в модели двух ранее рассмотренных случаев:

$$y_t = \alpha_0 + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_k x_{t-k} + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + \varepsilon_t.$$

Один из центральных моментов при построении моделей с Л.п. – выбор величины лага.

ЛАГОВАЯ СТРУКТУРА

последовательность коэффициентов $\beta_0, \beta_1, \ldots$ ……в регрессионных *моделях с распределёнными лагами*:

$$y_t = \alpha + \sum_{j=0}^{q} \beta_j x_{t-j} + \varepsilon_t,$$

где ε_t – последовательность гомоскедастичных и взаимно не коррелированных регрессионных остатков, $\alpha, \beta_0, \beta_1, \ldots$ – неизвестные коэффициенты модели. Эта последовательность коэффициентов может быть конечной или бесконечной (в зависимости от q).

Если все $\beta_j \geq 0$ ($j = 0, 1, 2, \ldots$), то последовательность коэффициентов $w_0, w_1, w_2, \ldots,$ полученная с помощью соотношения

$$w_j = \beta_j / \sum_{j=0}^{q} \beta_j,$$

называется нормированной структурой лага модели.

Различные типовые модели распределенных лагов отличаются способом параметризации Л.с.

Наибольшее распространение в практике эконометрического моделирования получили *лаговая структура геометрическая* (Койка) (для модели с бесконечным числом лагов), *лаговая структура полиномиальная* (Алмон) (для модели с конечным числом лагов). Практический интерес также представляет *лаговая структура вероятностная*.

ЛАГОВАЯ СТРУКТУРА ВЕРОЯТНОСТНАЯ

лаговая структура модели распределённых лагов, опирающаяся на сходство нормированной структуры *лага* и закона *распределения вероятностей* дискретной случайной величины. Пусть модель с распределёнными лагами имеет вид:

$$y_t = \alpha + \sum_{j=0}^{q} \beta_j x_{t-j} + \varepsilon_t \qquad (1),$$

где ε_t – последовательность гомоскедастичных и взаимно не коррелированных регрессионных остатков, $\alpha, \beta_0, \beta_1, \ldots$ – неизвестные коэффициенты модели. В модели (1) q может быть как конечным, так и бесконечным. Вводится случайная величина τ («время задержки») с законом распределения вероятностей:

$$P(\tau = j) = w_j, \; j = 0, 1, 2, \ldots q,$$

где $w_j = \beta_j / \sum_{j=0}^{q} \beta_j$, $\beta_j \geq 0$.

Такой подход открывает возможность параметризации нормированной структуры лага (последовательности $w_0, w_1, w_2, \ldots$) с помощью моделей законов распределения дискретных случайных величин.

ЛАГОВАЯ СТРУКТУРА ГЕОМЕТРИЧЕСКАЯ (КОЙКА)

бесконечная *лаговая структура*, получающаяся в предположении, что параметры β_j в *модели распределённых лагов*:

$$y_t = \alpha + \sum_{j=0}^{\infty} \beta_j x_{t-j} + \varepsilon_t \qquad (1)$$

убывают в геометрической прогрессии $\beta_j = \beta_0 \lambda^j$, $j = 0, 1, 2, \ldots$, $0 < \lambda < 1$.

Т.о., модель (1) может быть представлена в виде

$$y_t = \alpha + \beta_0 x_t + \beta_0 \lambda x_{t-1} + \beta_0 \lambda^2 x_{t-2} + \ldots + \varepsilon_t \;, \qquad (2)$$

где ε_t – последовательность гомоскедастичных и взаимно не коррелированных регрессионных остатков, $\alpha, \beta_0, \lambda.$ – неизвестные параметры модели. Параметр λ характеризует скорость убывания коэффициентов модели при увеличении лага. Нормированная Л.с.г. модели Койка определяется положительными весами $w_j = (1-\lambda)\lambda^j$.

Модель (2) с бесконечным числом лагов можно привести к модели с конечным числом членов. Для этого из уравнения (2) надо вычесть то же самое уравнение для момента времени t -1, умноженное на λ. В результате модель будет иметь вид:

$$y_t = (1-\lambda)\alpha + \beta_0 x_t + \lambda y_{t-1} + u_t \; (3),$$

где $u_t = \varepsilon_t - \lambda \varepsilon_{t-1}$.

Модель содержит три параметра (α, λ, β_0), однако включает лагированную эндогенную переменную и ошибки, не удовлетворяющие условиям классической модели линейной регрессии. В связи с этим применяются нестандартные методы оценивания коэффициентов, так как *метод наименьших квадратов* (МНК)-оценки являются несостоятельными.

ЛАГОВАЯ СТРУКТУРА ПОЛИНОМИАЛЬНАЯ (АЛМОН)

лаговая структура, предполагающая полиномиальную форму зависимости весовых коэффициентов β_j от j в *модели распределённых лагов*:

$$y_t = \alpha + \sum_{j=0}^{q} \beta_j x_{t-j} + \varepsilon_t \; (1),$$

где ε_t – последовательность гомоскедастичных и взаимно не коррелированных регресси-

онных остатков, $\alpha, \beta_0, \beta_1, \ldots, \beta_q$ – неизвестные коэффициенты модели.

Т.о., зависимость весовых коэффициентов β_j от j аппроксимируется полиномом некоторой степени m:

$\beta_j = \gamma_0 + \gamma_1 j + \ldots + \gamma_m j^m$, j =0,1,…,q. (2)

Подход опирается на использование полиномов невысокой степени (как правило, $m \leq 3$). После подстановки (2) в (1) получается модель с m+2 неизвестными коэффициентами:

$y_t = \alpha + \gamma_0 \tilde{x}_{t0} + \ldots + \gamma_m \tilde{x}_{tm} + \varepsilon_t$ (3),

где переменные $x_{t0}, \ldots, x_{tm}$ являются линейными комбинациями $x_t, \ldots, x_{t-q}$.

Значения коэффициентов $\alpha, \gamma_0, \ldots, \gamma_m$ в модели (3) определяются с помощью *метода наименьших квадратов*, параметры β_j – из соотношения (2)

ЛИНЕАРИЗАЦИЯ

преобразование, позволяющее перейти от нелинейных относительно исходных переменных моделей (уравнений регрессии) к линейным. Многие важные связи в экономике являются нелинейными. К ним относятся производственные функции, функции спроса и предложения и другие. Для таких моделей пытаются подобрать такие преобразования, которые позволили бы представить искомую зависимость в виде линейного соотношения между преобразованными переменными. Некоторые виды нелинейных зависимостей поддаются непосредственной Л. Напр., логарифмические преобразования позволяют перейти от степенной модели к линейной. При неизвестном виде функции, описывающей нелинейную зависимость, можно воспользоваться формализованной процедурой универсального линеаризующего преобразования, предложенного английскими статистиками Г. Боксом и Д. Коксом.

Недостаток Л. с помощью преобразования исследуемых переменных состоит в том, что оценки параметров, полученные после Л. с помощью *метода наименьших квадратов*, минимизируют сумму квадратов отклонений для преобразованных, а не для исходных переменных, поэтому полученные с помощью Л. зависимостей оценки нуждаются в уточнении.

М

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

совокупность математических соотношений, описывающих изучаемую систему.

Один из осн. инструментов изучения некоторой системы – метод моделирования, заключающийся в разработке и исследовании модели – некоторой вспомогательной системы, которая отражает существенные для исследователя свойства исходной системы. В зависимости от целей исследования используют как материальные модели, воспроизводящие материальные характеристики системы, так и нематериальные модели, отражающие её свойства в виде некоторой абстракции. М.м. – идеальный мыслимый аналог изучаемой системы в форме определённой математической структуры (напр., системы уравнений). Процесс математического моделирования включает построение такой структуры (спецификацию), её исследование (решение М.м.) с применением математических методов (идентификацию и верификацию), а также использование на практике (эксплуатацию). М.м. появились вместе с математикой много веков назад, но решающую роль в исследованиях стали играть с появлением ЭВМ, которое позволяет решать задачи, не поддающиеся строгому аналитическому исследованию и требующие проведения большого объема вычислений. М.м. используются для анализа, прогнозирования и оптимизации показателей самых различных систем (технических, социально-экономических, финансовых и др.). Преимущество М.м. состоит в её компактности, обусловленной применением математической символики, и широкой вариативности, позволяющей одни и те же хорошо изученные математические объекты применять к различным исследуемым системам. Ограниченность М.м. заключается в определённой схематичности отображения реальной действительности, а также проблеме интерпретации решения при исследовании сложных систем. Искусство по-

строения М.м. состоит в том, чтобы совместить как можно большую лаконичность в её математическом описании с достаточной точностью модельного воспроиз-ва именно тех сторон анализируемой реальности, которые интересуют исследователя.

Принято выделять условные этапы математического моделирования. Первый этап – выбор типа модели в зависимости от целей и задач исследования. Этот этап базируется на содержательном анализе исходной проблемы, предполагает сбор и осмысление всех уже имеющихся данных, относящихся к задаче. Второй этап: определение состава, структуры, вида входных и выходных переменных модели, а также её параметров. Вектор входных переменных $X = (X_1, X_2, ..., X_n)$ задается вне изучаемой системы и может разделяться на подвектор управлений $U = (U_1, U_2, ..., U_{p_1})$ и подвектор возмущений $V = (V_1, V_2, ..., V_{p_2})$ $(p_1 + p_2 = p)$. Вектор выходных переменных $Y = (Y_1, Y_2, ..., Y_m)$ определяется в рамках модели. Переменные V также называют экзогенными, а переменные Y – эндогенными. Переменные могут быть как детерминированными, так и *случайными величинами*. Выбор подлежащих учёту в М.м. существенных входных и выходных переменных и абстрагирование от прочих, предположительно незначимых существенно влияет на её качество и эффективность. Здесь необходимо глубокое понимание существа решаемой задачи, тщательное изучение воспроизводимой в модели исходной реальной системы, необходим опыт и эвристические способности. Вектор параметров $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_k)$ характеризует внутреннюю структуру модели, задается экспертно или определяется на основе статистических данных. Процесс отыскания параметров Θ состоит в минимизации отклонения результатов измерения величины Y от результатов её прогноза по модели и называется идентификацией модели. Третий этап: формулировка М.м. в конкретном математическом представлении и математический анализ модели, включающий определение области допустимых значений переменных, выяснение существования, единственности и устойчивости решения. На этом этапе формируется функционал (критерий) качества модели в виде некоторой зависимости $W = W(U, V, Y, \Theta)$. Математическая формулировка задачи может быть сформулирована, напр., т.о.: найти во множестве допустимых управлений $\tilde{U} \subseteq U^{(p_1)}$ такое оптимальное управление $U^* \in \tilde{U}$, которое при заданных значениях параметров Θ^0 и с учётом V и Y доставляет экстремум функции W, т.е.:

$$W^* = W(U^*, V, Y, \Theta^0) = \min_{U \in \tilde{U}} (\max)\, W(U, V, Y, \Theta^0).$$

Четвертый этап: решение М.м. аналитически и (или) с применением численных методов, а также компьютерного программного обеспечения. При этом широко используются т.н. проблемно-ориентированные интерактивные системы, позволяющие в диалоге с ЭВМ манипулировать параметрами модели, визуализировать и обрабатывать различным образом результаты расчетов. Пятый этап: проверка адекватности модели реальной изучаемой системе. В случае несоответствия модели реальному процессу возвращаются к одному из предыдущих этапов. Наилучшее в практическом отношении качество или эффективность М.м. измеряется с помощью величины W и достигается как разумный субъективно определяемый компромисс между близостью модели к оригиналу и простотой модели. Шестой этап: практическое использование модели.

Все многообразие М.м. можно систематизировать, руководствуясь различными признаками. В зависимости от задач исследования выделяют дескриптивные (описательные) модели, оптимизационные модели. Дескриптивные модели используются для осознания сущности изучаемой системы, а также прогнозирования её показателей. Оптимизационные модели, однокритериальные и многокритериальные, направлены на поиск значений переменных, улучшающих в смысле некоторого критерия основные показатели системы. По форме выражения исследуемых свойств системы различают аналитические (теоретические) модели, содержащие явные зависимости между переменными, а также

идентифицируемые модели, построение которых базируется на результатах измерений соответствующих величин переменных. Аналитические модели, как правило, представляют теоретические балансовые соотношения между входными и выходными переменными и выявляют сущность протекающих в системе процессов в терминах соответствующих теорий. Идентифицируемые модели формируются преимущественно на экспериментальной основе по наблюдаемым данным о входах и выходах системы. В зависимости от того, учитываются или нет при моделировании случайные воздействия на систему, различают детерминированные модели и недетерминируемые модели. Последние различаются на вероятностные модели, в которых случайные воздействия подчинены некоторым закономерностям, и модели с элементами неопределенности, в которых информации о распределении случайной составляющей нет. В зависимости от того, учитывается или нет показатель времени в задаче, различают динамические и статические модели. По типу переменных (непрерывные или дискретные), входящих в М.м., выделяют соответственно аналоговые (непрерывные) и дискретные модели. По типу используемого при моделировании математического инструментария, выделяют модели математического программирования, корреляционно-регрессионные модели, модели теории массового обслуживания, модели сетевого планирования и управления, модели теории игр и т.д. По форме выражения зависимостей между переменными различают линейные и нелинейные модели (относительно *X* и *Y*). Линейные модели наиболее полно изучены и дают более простые решения, поэтому на практике обычно стремятся свести к ним нелинейные модели, используя специальные процедуры *линеаризации*. По способу формирования М.м. различают феноменологическую модель как результат прямого изучения системы, асимптотическую модель как результат процесса дедукции из определенной общей модели, а также модель ансамблей как результат процесса индукции на базе некоторых элементарных моделей. Классифицируют также модели по отраслям наук (М.м. в физике, биологии, социологии и т.д.).

МЕТОД ИНСТРУМЕНТАЛЬНЫХ ПЕРЕМЕННЫХ

метод оценивания параметров регрессии при наличии *корреляции* между объясняющими переменными и регрессионными остатками. Рассматривается множественная линейная регрессия со стохастическими объясняющими переменными (регрессорами):

$$Y = X\beta + \varepsilon,$$
$$Y = (y_1, y_2, ..., y_n)^T, \quad \beta = (\beta_0, \beta_1, ..., \beta_p)^T,$$
$$X = \|x_{ij}\|_{n\times(p+1)}, \; x_{i0} = 1, \; i = 1,2,...,n,$$
$$P\{rankX = p+1\} = 1, \quad \varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T,$$
$$E\varepsilon = 0, \; \Sigma(\varepsilon) = \|\mathrm{cov}(\varepsilon_i, \varepsilon_j)\|_{n\times n} = \Sigma_0\sigma^2.$$

Ковариация $\mathrm{cov}(\varepsilon_i, x_{jk})$ не равна нулю хотя бы для одного из X_k, $k = 1,2,..., p$.

Корреляция между регрессорами и ошибками может быть вызвана либо существованием неучтённых в модели переменных, коррелированных с учтёнными объясняющими переменными, либо ошибками измерений этих переменных. Она проявляется в системах одновременных уравнений, а также в авторегрессионных моделях с автокоррелированными остатками. Оценка вектора β *методом наименьших квадратов* не является состоятельной, поскольку:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) = \beta + (\frac{1}{n} X^T X)^{-1} (\frac{1}{n} X^T \varepsilon)$$

В этом случае используют М.и.п., идея которого состоит в подборе таких вспомогательных (инструментальных) переменных

$$Z_1, Z_2, ..., Z_l, \quad l \geq k,$$

которые бы практически не коррелировали с регрессионными остатками и одновременно достаточно сильно коррелировали с исходными регрессорами. Инструментальные переменные – экзогенные переменные (внешние по отношению к модели), а исходные регрессоры – эндогенные, поскольку они связаны с ошибкой в

этом уравнении. Оценка вектора β М.и.п. определяется по формулам:

при $l = k$ $\hat{\beta}_{IP} = (Z^T X)^{-1} Z^T Y$,

при $l > k$

$\hat{\beta}_{IP} = (X^T PX)^{-1} X^T PY$,

$P = Z(Z^T Z)^{-1} Z^T$ (обобщённый М.и.п.).

Эта оценка является состоятельной, асимптотически несмещенной, не является эффективной. Осн. идея метода проявляется в его геометрической интерпретации: М.и.п. минимизирует расстояние между проекциями векторов Y и $\hat{Y}$ на подпространство, натянутое на вектора $Z_1, Z_2, ..., Z_l$. Он использует при оценивании только ту часть информации об объясняющих переменных, которая остается после проецирования их на вспомогательное подпространство. Поэтому качество оценки существенно зависит от выбора инструментальных переменных, в основе которого лежит опыт и интуиция исследователя. М.и.п. можно рассматривать как двухшаговый метод наименьших квадратов: на первом шаге строится регрессия каждого регрессора X_k, $(k = 1,2,..., p+1)$ на $Z_1, Z_2, ..., Z_l$ и определяется матрица «проекций» $\hat{X} = PX$. На втором шаге оценивается регрессия Y на $\hat{X}$. При этом

$\hat{\beta}_{IP} = (\hat{X}^T X)^{-1} \hat{X}^T Y$,

т.е. инструментами для каждого X_k в М.и.п. выступают наиболее близкие к нему (в смысле эвклидовой метрики) линейные комбинации исходных переменных. При обосновании М.и.п. можно также использовать *метод моментов*, рассматривая условие $E(Z^T \varepsilon) = 0$ и переходя к его выборочному аналогу. Для проверки гипотез о коэффициентах, в частности, о наличии сверхидентифицирующих ограничений, используют тест Вальда. Для обоснования применения М.и.п. применяют тест Хаусмана.

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (МНК)

осн. метод в теории ошибок для оценки неизвестных величин на основе результатов измерений, выполненных со случайными ошибками; предложен К. Гауссом в 1794. МНК применяется для аппроксимации результатов наблюдений с помощью их приближенного представления в виде функции из заданного параметрического семейства. В задаче выбора «наилучшей» аппроксимации массиву наблюдений $\{(x_i, y_i), i = 1,2,..., n\}$ сопоставляется действительная функция $f(x, \beta)$, зависящая от аргумента x и параметров β, при которых (в соответствии с МНК) значение средней квадратической ошибки оказывается миним.:

$$Q(\beta) = \sum_{i=1}^{n} \left(y_i - f(x_i, \beta)\right)^2 \to \min_{\beta} \qquad (1)$$

В частности, если параметрическая функция $f(x, \beta)$ является линейной относительно векторного параметра $\beta = (\beta_1, \beta_2, ..., \beta_k)^T$ и имеет вид

$$f(x, \beta) = \sum_{j=1}^{k} \beta_j \varphi_j(x),$$

где $\varphi_j(x)$ – известные функции, то значения параметров выбираются из условия минимизации функционала (1). Искомые k значений параметров β_j находят из системы k линейных («нормальных») уравнений (необходимые условия экстремума):

$$\frac{\partial Q}{\partial \beta_j} = 0 \quad (j = 1,..., k) \qquad (2)$$

Особую роль МНК играет в решении задачи статистического исследования зависимостей, являясь краеугольным методом получения оценок параметров регрессионных моделей. В рамках статистического подхода считается, что объекты исследуемой ген. совокупности характеризуются признаками (наблюдаемыми величинами) $x^{(1)}, x^{(2)}, ..., x^{(k)}, y$, где $x^{(1)}, x^{(2)}, ..., x^{(k)}$ – объясняющие переменные, а y – результирующая переменная, которая может быть представлена в виде функции от объясняющих переменных (функция регрессии) с добавлением аддитивной остаточной компоненты, выражающей влияние на y дополнительных (ненаблюдаемых) факторов: $y = f(x^{(1)}, x^{(2)}, ..., x^{(k)}) + \varepsilon$.

По результатам наблюдений $\{(x_i^{(1)}, x_i^{(2)}, ..., x_i^{(k)}, y_i), i = 1,..., n\}$, снятых с n объектов, случайно выбранных из ген. совокупности, необходимо при любом наборе значений объясняющих переменных построить

приближенное значение функции регрессии f (оценить f).

Часто указанная регрессионная модель конкретизируется в виде линейной модели множественной регрессии (ЛММР):

$$y_i = \beta_1 x_i^{(1)} + ... + \beta_k x_i^{(k)} + \varepsilon_i, \quad i = \overline{1,n}, \qquad (3)$$

в которой постулируется линейная зависимость между результирующей и объясняющими переменными. Коэффициенты $\beta_1, \beta_2, ..., \beta_k$ модели (3) – постоянные величины, независящие от наблюдений, а возмущения (регрессионные ошибки) $\varepsilon_1, ..., \varepsilon_n$ интерпретируются как случайные величины, выражающие отклонения фактических значений результирующей переменной от её модельных значений. Задача оценивания неизвестных значений параметров $\beta_1, \beta_2, ..., \beta_k$ состоит в том, чтобы построить такие функции от имеющегося массива наблюдений (оценки), которые давали бы «хорошие» приближения для неизвестных коэффициентов (оценки обычно считаются «хорошими», если они обладают свойствами несмещенности и/или состоятельности).

Решение задачи удобно представить, используя матричную форму записи ЛММР:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (4)$$

Здесь $\mathbf{y} = (y_1, ..., y_n)^{\mathrm{T}}$ – $n \times 1$ вектор наблюдений переменной y,

$$\mathbf{X} = \left\{ x_i^{(j)} \right\}_{\substack{i=\overline{1,n} \\ j=\overline{1,k}}}$$

– $n \times k$ матрица наблюдений объясняющих переменных (матрица плана), $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^{\mathrm{T}}$ – $k \times 1$ вектор коэффициентов модели, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^{\mathrm{T}}$ – $n \times 1$ вектор регрессионных ошибок.

Функционал (1) для ЛММР в матричной форме представляется в виде:

$$Q(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\mathbf{b}) \to \min_{\mathbf{b}},$$

а условие первого порядка (2) – в виде

$$\frac{\partial Q}{\partial \mathbf{b}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = \mathbf{O}_k .$$

При решении последнее соотношение относительно $\mathbf{b}$ в предположении, что матрица $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ – невырожденная, получается формула для вектора МНК-оценок параметров ЛММР:

$$\hat{\boldsymbol{\beta}}_{\text{мнк}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \qquad (5)$$

Статистический анализ, основанный на МНК-оценивании, в базовом варианте проводится в рамках *модели классической линейной множественной регрессии* (МКЛМР). В ней предполагается, что $\mathbf{X}$ – детерминированная матрица с полным рангом (объясняющие переменные имеют неслучайную природу и являются линейно независимыми); случайные величины $\varepsilon_1, ..., \varepsilon_n$ имеют нулевые средние значения $(E\varepsilon_i = 0)$, одинаковые дисперсии ($D\varepsilon_i = \sigma^2$ – свойство гомоскедастичности регрессионных возмущений) и нулевые корреляции $\left(E\varepsilon_i\varepsilon_j = 0, \quad i \neq j, i, j = \overline{1,n}\right)$. Т.о., МКЛМР представляется в виде:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\ , \ E\boldsymbol{\varepsilon} = O_n, V(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n ,$$

где $V(\boldsymbol{\varepsilon}) = \left\{ \sigma_{ij} \right\}_{i,\, j=\overline{1,n}}$ – ковариационная матрица с элементами $\sigma_{ij} = \mathrm{cov}(\varepsilon_i, \varepsilon_j) = E\varepsilon_i\varepsilon_j$, которые в классических условиях равны σ^2 при $i = j$ и нулю при $i \neq j$.

В рамках МКЛМР оценки $\hat{\boldsymbol{\beta}}_{\text{мнк}}$ обладают рядом «хороших» свойств, используемых для получения статистических выводов: из условия $E\boldsymbol{\varepsilon} = O_n$ следует несмещенность МНК-оценок: $E\hat{\beta}_{j,\,МНК} = \beta_j$, $j = \overline{1,n}$; МНК-оценки состоятельны тогда и только тогда, когда наименьшее собственное значение матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ стремится к бесконечности при $n \to \infty$; теорема Гаусса-Маркова. МНК-оценки эффективны (оптимальны), т.е. имеют наименьшие дисперсии в классе несмещенных оценок, являющихся линейными функциями от $y_1, ..., y_n$: $D\hat{\beta}_{j,\,МНК} \le D\tilde{\beta}_j$ (для любой оценки $\tilde{\beta}_j = c_1 y_1 + ... + c_n y_n, E\tilde{\beta}_j = \beta_j$), $j = \overline{1,n}$; ковариационная матрица вектора МНК-оценок вычисляется по формуле:

$$V(\hat{\beta}_{МНК}) = \sigma^2 \left(X^T X \right)^{-1} \qquad (6)$$

Поэтому

$$D(\hat{\beta}_{j,МНК}) = \sigma^2 \left[\left(X^T X \right)^{-1} \right]_{jj}, \quad j = \overline{1,k},$$

где $\left[A\right]_{ij}$ – *(i,j)*-элемент матрицы A; стандартные ошибки МНК-оценок $\hat{\beta}_{j,МНК}$ вычисляются по формулам:

$$s.e.(\hat{\beta}_{j,MHK}) = \sqrt{s^2\left[\left(X^T X\right)^{-1}\right]_{jj}}, \quad j = 1,...,k, \qquad (7)$$

где $s^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\beta}_{MHK})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\beta}_{MHK})\{\ \}$ – несмещённая оценка параметра σ^2; стандартные ошибки являются состоятельными оценками стандартных отклонений:

$s.d.(\hat{\beta}_{j,MHK}) = \sqrt{D\hat{\beta}_{j,MHK}}, \quad j = 1,...,k$;

при дополнительном предположении о нормальности регрессионных ошибок $\varepsilon_1,...,\varepsilon_n \square N(0,\sigma^2)$ (и, как следствие, их одинаковой нормальной распределенности и независимости) вектор МНК-оценок подчиняется многомерному нормальному распределению с вектором средних значений β и матрицей ковариаций $\sigma^2\left(X^T X\right)^{-1}$, а статистика

$$\frac{(n-k)s^2}{\sigma^2}$$

имеет χ^2-распределение с $n-k$ степенями свободы, причем $\hat{\beta}_{MHK}$ и s^2 – статистически независимы; t-статистика

$$t_j = \frac{\hat{\beta}_{j,MHK} - \beta_j}{s.e.(\hat{\beta}_{j,MHK})}$$

подчиняется распределению Стьюдента с $n-k$ степенями свободы; для выборок большого объёма (при $n \to \infty$) вне зависимости от закона распределения регрессионных ошибок оценки $\hat{\beta}_{j,MHK}$ имеют асимптотически нормальные распределения, а t-статистики t_j подчиняются (в асимптотике) стандартному нормальному распределению $N(0,1)$.

В случае невыполнения классических условий линейная модель множественной регрессии называется *обобщённой линейной моделью множественной регрессии* (ОЛММР). Эта модель имеет вид $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $E\boldsymbol{\varepsilon} = O_n, V(\boldsymbol{\varepsilon}) = \Omega$, где Ω – симметричная, положительно определённая $n \times n$ - матрица, $\Omega \neq \sigma^2\mathbf{I}_n$.

Заметим, что дисперсии регрессионных ошибок, стоящие на гл. диагонали матрицы Ω могут быть неодинаковыми, а среди недиагональных элементов Ω могут встречаться ненулевые числа.

Поскольку МНК – сугубо алгебраический метод, формулы (5) для вычисления МНК-оценок коэффициентов ЛММР остаются неизменными при любых статистических свойствах регрессионных ошибок.

В рамках ОЛММР обычные МНК-оценки коэффициентов β остаются несмещёнными и состоятельными, но не являются оптимальными: в классе линейных по у, несмещённых оценок наилучшими (в смысле величины дисперсии) оказываются оценки, полученные не обычным, а *обобщённым методом наименьших квадратов*.

Тем не менее, МНК-оценки можно использовать для получения статистических выводов (для проверки гипотез и построения доверительных интервалов), надо только иметь в виду, что формула (6) для вычисления ковариационной матрицы вектора $\hat{\beta}_{MHK}$ в условиях ОЛММР становится неверной. Можно показать, что теперь:

$$V(\hat{\beta}_{MHK}) = \left(X^T X\right)^{-1} X^T \Omega X \left(X^T X\right)^{-1}$$

Соответственно, и стандартные ошибки МНК-оценок, посчитанные по формулам (7), перестают быть состоятельными по отношению к величинам стандартных отклонений $s.d.(\hat{\beta}_{j,MHK}), \quad j = 1,...,k$

Г. Уайтом в 1980 предложен способ состоятельного оценивания элементов матрицы ковариаций $V(\hat{\beta}_{MHK})$ для случая гетероскедастичных $\left(D\varepsilon_i = \sigma_i^2 \neq const,\ i = 1,...,n\right)$, но некоррелированных $\left(E\varepsilon_i\varepsilon_j = 0,\ i \neq j\right)$ регрессионных ошибок. В этом случае, характерном для пространственных (перекрестных) данных, матрица Ω имеет диагональный вид $\Omega = diag\left\{\sigma_1^2,...,\sigma_n^2\right\}$. Для получения состоятельной оценки ковариационной матрицы $V(\hat{\beta}_{MHK})$ необходимо в формуле (8) заменить матрицу Ω на её оценку $\hat{\Omega} = diag\left\{e_1^2,...,e_n^2\right\}$, где на гл. диагонали стоят квадраты регрессионных МНК-остатков, вычисляемых по формулам $e_i = y_i - x_i^T\hat{\beta}_{MHK}$, $x_i^T = \left(x_i^{(1)},...,x_i^{(k)}\right)$. Стандартные ошибки МНК-оценок β_{MHK}, посчитанные т.о. являются состоятельными и назы-

ваются стандартными ошибками в форме Уайта или стандартными ошибками, учитывающими гетероскедастичность регрессионных возмущений:

$$s.e.(\hat{\beta}_{j,МНК}) = \sqrt{\left(X^T X\right)^{-1} X^T diag\left\{e_1^2,...,e_n^2\right\} X \left(X^T X\right)^{-1}} \qquad (8)$$

Существуют также формулы, предложенные Ньюи и Уэстом в 1987 для состоятельного оценивания матрицы $V(\hat{\beta}_{МНК})$ в более общей ситуации, характерной для временных рядов, когда в матрице $\Omega = (\sigma_{ij})$ ненулевые элементы могут встречаться не только на гл. диагонали, но и на ближайших к ней L – соседних диагоналях $\left(\sigma_{ij} = 0, |i-j| > L\right)$. Оценки стандартных отклонений, основанные на них, называются стандартными ошибками в форме Ньюи-Уэста.

См. также *Двухшаговый метод наименьших квадратов*, *Трёхшаговый метод наименьших квадратов*.

МОДЕЛЬ АВТОРЕГРЕССИИ – СКОЛЬЗЯЩЕГО СРЕДНЕГО (АРСС)

наиболее полно и экономно выражает автокорреляционные свойства стационарного *временного ряда*, используя небольшое число оцениваемых параметров; в англоязычной литературе называется ARMA-модель (от англ. – Auto-Regressive-Moving Average model); имеет вид:

$$y_t = \alpha_1 \cdot y_{t-1} + \alpha_2 \cdot y_{t-2} + ... + \alpha_p \cdot y_{t-p} + \varepsilon_t - \theta_1 \cdot \varepsilon_{t-1} - \theta_2 \cdot \varepsilon_{t-2} - ... - \theta_q \cdot \varepsilon_{t-q}$$

Такая модель АРСС(p,q) может интерпретироваться как линейная модель множественной регрессии, в которой в качестве объясняющих переменных выступают прошлые значения самой зависимой переменной, а в качестве регрессионного остатка – скользящие средние из элементов *белого шума* ε_t с нулевым математическим ожиданием и дисперсией σ_0^2. Параметры p и q определяют соответственно порядок авторегрессионной составляющей и порядок скользящих средних. Для выполнения условия стационарности необходимо и достаточно, чтобы вне единичного круга лежали все корни характеристического уравнения: $1 - \alpha_1 \cdot z - \alpha_2 \cdot z^2 - ... - \alpha_p \cdot z^p = 0$.

Применение модели АРСС(p,q) распространено на случай нестационарных временных рядов, характеризующихся наличием полиномиального тренда. Для описания таких рядов используется *модель авторегрессии-проинтегрированного скользящего среднего в остатках* (АРПСС(p, d, q)) или в англоязычном варианте AutoRegressive Integrated Moving Average model (ARIMA-model). В специальной литературе она также известна как модель Бокса-Дженкинса. В модели АРПСС(p, d, q) наряду с параметрами p и q указывается параметр d, определяющий порядок разностей ряда. Переход к разностям временного ряда предполагает исключение *тренда,* а затем для ряда разностей строится модель АРСС.

МОДЕЛЬ АВТОРЕГРЕССИИ СО СКОЛЬЗЯЩИМ СРЕДНИМ В ОСТАТКАХ

смешанная *модель авторегрессии–скользящего среднего* (АРСС); наиболее полно и экономно выражает автокорреляционные свойства стационарного временного ряда x_t; записывается модель АРСС(p,q) следующим образом:

$$\tilde{x}_t = \varphi_1 \tilde{x}_{t-1} + ... + \varphi_p \tilde{x}_{t-p} + \varepsilon_t - Q_1 \varepsilon_{t-1} - ... - Q_q \varepsilon_{t-q},$$

где $\tilde{x}_t = x_t - m_x$ – отклонение от средней; ε_t – случайная компонента, интерпретируемая как ошибка прогнозирования на 1 шаг вперед, со средним значением (математическим ожиданием) ноль и дисперсией σ^2; p – порядок авторегрессии; q – порядок скользящей средней, φ_1, $\varphi_2,...,\varphi_p$, $Q_1,...Q_q$ – коэффициенты, подлежащие оцениванию. Фактически модель скользящей средней в данном случае описывает ошибку модели $u_t = \varepsilon_t - Q_1 \varepsilon_{t-1} - \cdots - Q_q \varepsilon_{t-q}$. Она добавлена в модель авторегрессии с целью лучшего согласования динамики ряда и модели

и является корректирующим элементом, формирующим поправку прогноза с учетом ошибок, допущенных в предыдущие q моментов времени. Отметим, что сумма коэффициентов Q не обязательно должна равняться 1.

Модель предложена и изучена Боксом и Дженкинсом в 1970; является частным случаем рассмотренной ими *модели авторегресии-проинтегрированного скользящего среднего в остатках.*

МОДЕЛЬ АВТОРЕГРЕСИИ-ПРОИНТЕГРИРОВАННОГО СКОЛЬЗЯЩЕГО СРЕДНЕГО В ОСТАТКАХ (МОДЕЛЬ БОКСА-ДЖЕНКИНСА)

расширение *модели авторегрессии со скользящим средним в остатках* (АРСС) на случай нестационарных рядов, характеризующихся наличием полиномиального тренда. В этом случае от нестационарного ряда переходят к стационарному путём построения модели АРСС для разностей исходного ряда соответствующего порядка d. Порядок разностей d зависит от порядка полинома. Такую модель называют проинтегрированной моделью авторегрессии-скользящего среднего и кратко записывают как АРПСС(p,d,q) (англ. ARIMA(p,d,q)). Предложена и изучена Боксом и Дженкинсом в 1970. Большое внимание авторы уделили проблеме идентификации структуры модели, прежде всего, определению порядка разности d и порядка авторегрессии p . Для этого авторы предлагали использовать автокорреляционную и частную автокорреляционную функции и определять характер и момент (лаг) затухания этих функций. Однако впоследствии для определения порядков модели p,d,q были разработаны статистические критерии.

В частности, для нахождения порядка разности d, необходимого для перехода к стационарности, был предложен критерий Дикки-Фуллера (от англ. – Dickey-Fuller Test) и его более общий вариант – расширенный критерий Дикки-Фуллера (от англ.– Augmented Dickey-Fuller Test). Авторы критерия начали с построения процедуры проверки *нулевой гипотезы* H_0: ρ=1 в модели $y_t = \alpha + \rho y_{t-1} + u_t$, где u_t – *белый шум*. Ясно, что если нулевая гипотеза верна, то временной ряд нестационарен (или, как говорят, имеется единичный корень) и нужно переходить к первым разностям $\Delta y_t = y_t - y_{t-1}$. Этот тест выполняется путём оценивания уравнения, получаемого после перехода к разностям, когда модель записывается как $\Delta y_t = \gamma y_{t-1} + u_t$, где $\gamma=\rho-1$, и проверки гипотезы H_0: γ=0, которая эквивалента гипотезе, что ρ=1. Затем авторами рассмотрены три различных регрессионных уравнения, в которых проверяется наличие единичного корня: $\Delta y_t = \gamma y_{t-1} + u_t$ (1),
$\Delta y_t = \alpha + \gamma y_{t-1} + u_t$ (2),
$\Delta y_t = \alpha + \gamma y_{t-1} + \beta t + u_t$ (3). Отметим, что при $\gamma=\rho-1$ модель (3) может быть приведена к виду $y_t = \alpha + \rho y_{t-1} + \beta t + u_t$. Различие в моделях (1), (2), (3) состоит в наличии детерминированных элементов α и βt. Первая модель – модель чисто случайного блуждания, во вторую модель добавляется свободный член, который является параметром дрейфа, а в третью – включены и дрейф, и линейный временной тренд. Во всех трёх регрессиях интерес представляет параметр γ. Если γ=0, то ряд y_t характеризуется единичным корнем, т.е. для достижения стационарности требуется переход к первым разностям. Процедура проверки гипотезы предполагает оценивание *методом наименьших квадратов* (МНК) одного или нескольких записанных выше уравнений с целью получения оценки γ и её стандартной ошибки. Делением оценки на её стандартную ошибку получают t-статистику, которая, однако, при нулевой гипотезе о наличии единичного корня имеет распределение, отличное от *распределения Стьюдента*. Авторами получены критические значения для t, путём модельных статистических испытаний. Сравнение полученных t-статистик с критическим значением, приведённым в табл. Дикки-Фуллера (обычно оно выводится программой, предусматривающей процедуру проверки наличия корней), позволяет принять или отвергнуть нулевую гипотезу H_0: γ=0.

Расширенный критерий Дикки-Фуллера предназначен для более общего случая, когда возмущения u_t автокоррелированы, что предлага-

ется учесть, добавлением в уравнение регрессии лаговых разностей y_t. Другими словами, расширенный критерий Дикки-Фуллера используется для проверки наличия единичного корня в авторегрессионных уравнениях более высокого порядка, имеющих вид:

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=2}^{p} \varphi_i \Delta y_{t-i+1} + u_t \quad (4),$$

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{i=2}^{p} \varphi_i \Delta y_{t-i+1} + u_t \quad (5),$$

$$\Delta y_t = \alpha + \gamma y_{t-1} + \beta t + \sum_{i=2}^{p} \varphi_i \Delta y_{t-i+1} + u_t \quad (6)$$

, где нулевая гипотеза H_0: γ=0, и критические значения находят по таблицам Дикки-Фуллера. Если нулевая гипотеза принята, то целесообразно проверить наличие единичного корня у первых разностей и т.д. пока нулевая гипотеза не будет отклонена. Т.о. находят порядок разности d необходимый для получения стационарности. Величину d называют порядком интегрированности ряда.

Далее определяют порядки авторегрессии и скользящего среднего p и q. Для этого используется *информационный критерий Акайка* (от англ. – Akaike Information Criterion – AIC) или байесовский критерий Шварца (от англ. – Schwarz Bayesian Criterion – SBC):

$$AIC = \ln(\sigma^2) + \frac{2k}{n} \quad (7),$$

$$SBC = \ln(\sigma^2) + \frac{k \ln n}{n} \quad (8),$$

где σ^2 – дисперсия остатков регрессии, *k*- число коэффициентов в регрессии, *n*-объем выборки. Чем меньше значение AIC или SBC, тем лучше модель. Вторые слагаемые в обоих критериях играют роль «штрафа» за введение дополнительной переменной, в том числе лагового значения, в правую часть уравнения. Оптимальные значения порядков p и q, найденные по этим двум критериям могут различаться. Как правило, при использовании AIC порядки будут либо такими же, либо больше, чем при SBC, так как «штраф» в SBC более суровый. Отметим, что в ряде программ эти критерии могут быть построены так, что требуется их минимизация.

Модели Бокса-Дженкинса были расширены и на случай динамических рядов с сезонными эффектами. Тогда предлагается переходить к разностям порядка L, где L – число фаз в сезонном цикле. Напр., если исходные данные ежемесячные, то, полагая, что сезонность связана с вращением Земли вокруг Солнца, L=12, т .е. изучаются приросты между одноименными фазами сезонного цикла.

МОДЕЛЬ АВТОРЕГРЕССИОННЫХ УСЛОВНО ГЕТЕРОСКЕДАСТИЧНЫХ ОСТАТКОВ

модель зависимости текущей дисперсии остатков модели временного ряда от квадратов остаточных членов предыдущих периодов; в англоязычной литературе она называется *моделью-ARCH* (от англ. – AutoRegressive Conditional Heteroscedasticity). Модель отражает инерционность изменчивой дисперсии исследуемого процесса и может быть представлена в виде

$$\sigma_t^2 = \overline{w} + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2,$$

где параметры $\overline{w}$ и α_j $(j = 1,2,\ldots, p)$ – неотрицательные величины, p – порядок процесса авторегрессионной условной гетероскедастичности. Влияние остатка модели, отстоящего от настоящего момента времени на j, на текущую дисперсию определяется коэффициентом α_j, который для обеспечения стационарности процесса должен быть меньше единицы. На текущую дисперсию не влияют остатки, удалённые более чем на p тактов времени. Наиболее широко М.а.у.г.о. используются в моделировании финансовых временных рядов, которые характеризуются инерционной волатильностью. Для тестирования на условную авторегрессионную *гетероскедастичность* можно использовать *тест Бреуша-Пагана на гетероскедастичность остатков*. При обнаружении гетероскедастичности целесообразно использовать более эффективные методы оценивания параметров модели временного ряда, чем *метод наименьших квадратов*, такие как *обобщённый метод наименьших квадратов, метод макс. правдоподобия.*

МОДЕЛЬ АВТОРЕГРЕССИОННЫХ УСЛОВНО ГЕТЕРОСКЕДАСТИЧНЫХ ОСТАТКОВ ОБОБЩЁННАЯ

модификация *модели авторегрессионных условно гетероскедастичных остатков*; в англоязычной литературе они именуются *моде-*

лями-GARCH (от англ. – Generalized Auto-Regressive Conditional Heteroscedasticity). Существует большое число модификаций М.а.у.г.о.о., которые можно разделить на две большие группы: симметричные, в которых значение имеют только абсолютные значения возмущений временного ряда, и асимметричные. Симметричная модель может быть представлена в виде:

$$\sigma_t^2 = \overline{w} + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 ,$$

где параметры $\overline{w}$, α_j, β_j – неотрицательные параметры, σ_t^2 – текущая дисперсия, ε_t – текущий остаток модели временного ряда. Модель текущих остатков при этом в случае $p=1$ и $q=1$ представляет собой модель авторегрессии и скользящего среднего АРСС(1;1): $\varepsilon_t^2 = \overline{w} + (\alpha+\beta)\varepsilon_{t-j}^2 + \nu_t - \beta\nu_{t-1}$, в которой ошибка $\nu_t = \varepsilon_t^2 - \sigma_t^2$ является гетероскедастичной. Спецификация М.а.у.г.о.о. текущей дисперсии σ_t^2 порядка (1;1) эквивалентна спецификации модели условно гетероскедастичных остатков бесконечного порядка. Поэтому обобщённая модель более компактна и экономна. Несимметричные М.а.у.г.о.о. отражают тот факт, что изменение признака в одном направлении может иметь большее влияние на будущую дисперсию, чем такое же по абсолютной величине изменение в противоположном направлении. Напр., неожиданное снижение цены имеет большее воздействие на будущую волатильность, чем её неожиданное увеличение. К несимметричным относится, в частности, экспоненциальная М.а.у.г.о.о. Нельсона:

$$\log\sigma_t^2 = \overline{w} + \beta_j \log\sigma_{t-j}^2 + \gamma\frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \alpha\frac{\left|\varepsilon_{t-1}^2\right|}{\sigma_{t-1}} ,$$

которая асимметрична при $\gamma \neq 0$.

МОДЕЛЬ АДАПТИВНАЯ

в широком смысле это самокорректирующаяся, самонастраивающаяся модель, способная отражать изменяющиеся во времени условия, учитывать информационную ценность различных членов временной последовательности (временного ряда) и давать оценки будущих членов исследуемого ряда. М.а. предназначается, прежде всего, для краткосрочного статистического прогнозирования, а также для анализа на выборке долгосрочных тенденций.

В узком смысле под М.а. в статистике принято понимать модель, процедура корректировки параметров которой основывается на использовании формулы экспоненциально-взвешенной скользящей средней (EWMA – Exponentially Weighted Moving Average). Разработано большое число моделей адаптивного типа, способных отражать полиномиальные и экспоненциальные тренды, сезонные колебания аддитивного и мультипликативного типа, эволюционирующие законы распределения вероятностей, линейные и нелинейные регрессионные зависимости, нестационарные корреляционные связи.

Среди М.а. можно выделить класс моделей адаптивных комбинированных – комбинирование означает использование нескольких адаптивных моделей. М.а. предполагают формирование базового набора наиболее перспективных моделей с относительно простой структурой, по которым на каждом шаге продвижения во времени рассчитываются пробные прогнозы, которые затем сравниваются с реальными наблюдениями. Можно выделить два класса комбинированных моделей – селективные и гибридные модели. В селективной модели ошибки прогнозов используются для формирования адаптивного критерия, позволяющего выбирать из базового набора моделей наилучшую в текущий момент времени, по которой и делают реальный прогноз. Т.о., этот подход предполагает последовательное переключение с одной модели на другую, то есть имеет место не только адаптации параметров, но и структуры модели. В гибридной модели производится объединение прогнозов, полученных по моделям, входящих в базовый набор. Объединение прогнозов может осуществляться с различными весами в зависимости от точности или надежности прогнозов. В адаптивных моделях эти веса также могут иметь адаптивный характер и меняться на каждом шаге продвижения во времени, в соответствии с новыми данными о возрастании или ухудшении точности прогнозов

по соответствующему методу или модели. Объединение прогнозов целесообразно осуществлять тогда, когда нет решительных данных в пользу какого-либо одного метода прогнозирования. При этом ставится задача избежать наибольшей ошибки прогноза.

См. также *Модель Arch*, *Модель Garch*, *Модель Брауна обобщённая*, *Модель авторегрессии*, *Модель авторегрессии со скользящим средним в остатках*, *Модель скользящего среднего*, *Адаптивные методы прогнозирования*, *Параметр адаптации*, *Экспоненциальное сглаживание*.

МОДЕЛЬ БИНАРНОГО ВЫБОРА

регрессионная модель с зависимой переменной, принимающей одно из двух возможных значений, характеризующих, как правило, качественное состояние объекта. Примерами таких качественных состояний могут служить функционирование и банкротство пр-тия, приобретение или не приобретение покупателем данного товара, статус работающего или безработного. Наблюдаемые значения качественного признака удобно обозначать нулем и единицей. При этом модельное значение зависимой, в общем случае не целое, интерпретируется как вероятность принятия i-м объектом состояния, соответствующего единичному значению зависимой переменной y_i: $P(y_i = 1 | X_i) = \varphi(\theta^T X_i)$, где X_i – вектор количественных признаков, характеризующих i-й объект. Обычная линейная регрессионная модель в качестве М.б.в. неприменима вследствие возможного выхода модельных значений за пределы интервала [0;1]. Поэтому в М.б.в. используют монотонно возрастающие функции, приближённо линейные в окрестности нулевого значения аргумента, и стремящиеся к нулю и к единице по мере устремления её аргумента соответственно к -∞ и к ∞ влево. Такими свойствами обладают логистическая функция

$$e^{\theta^T X_i} / \left(1 + e^{\theta^T X_i}\right)$$

и стандартная нормальная вероятностная функция распределения $\Phi(\theta^T X_i)$, где

$$\Phi(z) = \left(1/\sqrt{2\pi}\right) \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx .$$

Соответствующие им модели называют *логит-* и *пробит-моделями бинарного выбора*. Важное условие корректного построения модели – равенство ковариационных матриц признаков для совокупностей объектов с нулевым и единичным наблюдаемыми значениями зависимой переменной. Выполнение этого условия можно проверить с помощью соответствующих статистических тестов. При наличии множества альтернатив интерпретация результатов моделирования затруднена, но такие модели часто могут быть сведены к М.б.в.

МОДЕЛЬ БРАУНА

модель адаптивная рекуррентная полиномиальная, основывающаяся на многократном экспоненциальном сглаживании временного ряда.

Простейшая адаптивная модель основывается на вычислении экспоненциально-взвешенной скользящей средней, называемой также просто экспоненциальной средней (англ. – Exponentially Weighted Moving Average – EWMA).

Предположим, что исследуется временной ряд x_t. Выявление и анализ тенденции динамического ряда часто производится с помощью его выравнивания или сглаживания. Экспоненциальное сглаживание – простой и распространённый приём выравнивания ряда.

Экспонециальное сглаживание ряда осуществляется по рекуррентной формуле

$$S_t = \alpha x_t + \beta S_{t-1} \qquad (1),$$

где S_t – значение экспоненциальной средней в момент t, α – параметр сглаживания, α=const, $0<\alpha<1$, $\beta=1-\alpha$. Выражение (1) можно переписать следующим образом

$$S_t = \alpha x_t + (1-\alpha) S_{t-1} = S_{t-1} + \alpha (x_t - S_{t-1}) \qquad (2).$$

Экспоненциальная средняя на момент t здесь выражена как экспоненциальная средняя предшествующего момента плюс доля α разницы текущего наблюдения и экспоненциальной средней прошлого момента.

Если последовательно использовать рекуррентное соотношение (1), то экспоненциальную среднюю S_t можно выразить через прошлые значения временного ряда x_t:

$$S_t = \alpha x_t + \beta S_{t-1} = \alpha x_t + \beta(\alpha x_{t-1} + \beta S_{t-2}) = \alpha x_t + \alpha\beta x_{t-1} + \beta^2 S_{t-2} = \ldots =$$

$$\alpha x_t + \alpha\beta x_{t-1} + \alpha\beta^2 x_{t-2} + \ldots + \alpha\beta^i x_{t-i} + \ldots + \beta^N S_0 = \alpha\sum_{i=0}^{N-1}\beta^i x_{t-i} + \beta^N S_0 \quad (3),$$

где N – число членов ряда, S_0 – некоторая величина, характеризующая начальные условия для первого применения рекуррентной формулы (1) при t=1. Так как 0<β<1, то при $N \to \infty$ $\beta^N \to 0$, а сумма коэффициентов

$$\alpha\sum_{i=0}^{N-1}\beta^i \to 1 \quad (4).$$

Тогда

$$S_t = \alpha\sum_{i=0}^{\infty}\beta^i x_{t-i} \quad (5).$$

Таким образом, величина S_t оказывается взвешенной средней суммой всех членов ряда. Причем веса падают экспоненциально в зависимости от давности («возраста») наблюдения. Это и объясняет, почему величина S_t названа экспоненциальной средней.

Для теоретического изучения свойств экспоненциальной средней рассмотрим ряд, генерированный моделью $x_t = a_1 + \varepsilon_t$ (6), где a_1=const, ε_t – случайные неавтокоррелированные отклонения, белый шум со средним значением 0 и постоянной дисперсией σ^2. Применим к нему процедуру экспоненциального сглаживания (1). Тогда

$$S_t = \alpha\sum_{i=0}^{\infty}\beta^i x_{t-i} = \alpha\sum_{i=0}^{\infty}\beta_i(a_1 + \varepsilon_{t-i}) = a_1 + \alpha\sum_{i=0}^{\infty}\beta^i\varepsilon_{t-i} \quad (7)$$

Найдем математическое ожидание экспоненциальной средней $E(S_t) = E(x_t) = a_1$ (8) и дисперсию

$$D(S_t) = E[(S_t - a_1)^2] = E\left[\left(\alpha\sum_{i=0}^{\infty}\beta^i\varepsilon_{t-i}\right)^2\right] = \alpha^2\sum_{i=0}^{\infty}\beta^{2i}\sigma^2 = \frac{\alpha}{2-\alpha}\sigma^2 \quad (9)$$

Так как 0<α<1, то $D(S_t) < D(x_t) = \sigma^2$.

Т.о., экспоненциальная средняя S_t имеет то же математическое ожидание, что и исходный ряд x_t, но меньшую дисперсию. Как видно из (9), при высоком значении α, близком к 1 , дисперсия экспоненциальной средней незначительно отличается от дисперсии ряда x_t. Чем меньше α, тем в большей степени сокращается дисперсия экспоненциальной средней. Следовательно, экспоненциальную среднюю можно представить как фильтр, на вход которого в виде потока последовательно поступают члены исходного ряда, а на выходе формируются текущие значения экспоненциальной средней. И чем меньше α, тем в большей степени фильтруются, подавляются колебания исходного ряда.

После появления работ Р.Г.Брауна (1959, 1963) экспоненциальная средняя часто используется для краткосрочного прогнозирования. В этом случае предполагается, что ряд генерируется моделью $x_t = a_{1,t} + \varepsilon_t$ (10), где $a_{1,t}$ – варьирующий во времени средний уровень ряда, ε_t – случайные неавтокоррелированные отклонения с нулевым математическим ожиданием и дисперсией σ^2. Прогнозная модель имеет вид $\hat{x}_\tau(t) = \hat{a}_{1,t}$ (11), где $\hat{x}_\tau(t)$ – прогноз, сделанный в момент t на τ единиц времени вперед, $\hat{a}_{1,t}$ – оценка $a_{1,t}$. Средством оценки единственного параметра модели служит экспоненциальная средняя $\hat{a}_{1,t} = S_t$ (12).

Следовательно, все свойства экспоненциальной средней распространяются на прогнозную модель. В частности, если S_t рассматривать как прогноз на один шаг вперед, то в выражении (2) величина (x_t-S_{t-1}) есть ошибка этого прогноза, а новый прогноз S_t получается как результате корректировки предыдущего прогноза с уче-

том его ошибки. В этом и состоит существо адаптации.

При краткосрочном прогнозировании желательно как можно быстрее отразить изменения в $a_{1,t}$ и в то же время как можно лучше «очистить» ряд от случайных колебаний. Таким образом, с одной стороны, следует увеличивать вес более свежих наблюдений, что может быть достигнуто повышением α, с другой стороны для сглаживания случайных отклонений величину α нужно уменьшить. Как видим, эти два требования находятся в противоречии. Теоретически вопрос о выборе оптимального значения α рассмотрел Д.Мат (Muth J.F.). Поиск компромиссного значения α и составляет задачу оптимизации модели. Такую модель будем называть адаптивной экспоненциального типа или адаптивным полиномом нулевого порядка, а величину α – параметром адаптации. Гл. достоинство прогнозной модели, основанной на экспоненциальной средней, состоит в том, что она способна последовательно адаптироваться к новому уровню процесса без значительного реагирования на случайные отклонения.

Адаптивные модели линейного роста. Можно показать, что экспоненциальная средняя приводит к смещённым прогнозам, т.е. даёт систематическую ошибку, когда временной ряд имеет устойчивую тенденцию линейного роста. Для этого случая разработано несколько вариантов адаптивных моделей, также использующих процедуру экспоненциального сглаживания. Мы рассмотрим две из них.

В основе моделей лежит гипотеза о том, что прогноз может быть получен по модели $\hat{x}_\tau(t) = \hat{a}_{1,t} + \tau\hat{a}_{2,t}$ (13), где $\hat{a}_{1,t}$, $\hat{a}_{2,t}$ – текущие оценки коэффициентов адаптивного полинома первого порядка. Задача сводится к получению оценок этих коэффициентов.

Одной из первых моделей этого типа была двухпараметрическая модель Хольта (Holt, 1957), в которой оценка коэффициентов производится следующим образом:

$$\hat{a}_{1,t} = \alpha_1 x_t + (1-\alpha_1)(\hat{a}_{1,t-1} + \hat{a}_{2,t-1}) \quad (14),$$

$$\hat{a}_{2,t} = \alpha_2(\hat{a}_{1,t} - \hat{a}_{1,t-1}) + (1-\alpha_2)\hat{a}_{2,t-1} \quad (15),$$

где α_1 и α_2 – параметры экспоненциального сглаживания ($0<\alpha_1$, $\alpha_2<1$), которые также будем называть параметрами адаптации. Эти уравнения могут быть переписаны в виде:

$$\hat{a}_{1,t} = \hat{a}_{1,t-1} + \hat{a}_{2,t-1} + \alpha_1 e_t \quad (16),$$

$$\hat{a}_{2,t} = \hat{a}_{2,t-1} + \alpha_1\alpha_2 e_t \quad (17),$$

где $e_t = x_t - \hat{x}_1(t-1)$ – ошибка прогноза на 1 шаг вперед. Частным случаем модели Хольта является модель линейного роста Р.Г.Брауна:

$$\hat{a}_{1,t} = \hat{a}_{1,t-1} + \hat{a}_{2,t-1} + (1-\beta^2)e_t \quad (18),$$

$$\hat{a}_{2,t} = \hat{a}_{2,t-1} + (1-\beta)^2 e_t \quad (19),$$

где параметр β – коэффициент дисконтирования, характеризующий обесценение наблюдения за единицу времени, $0<\beta<1$.

Аппроксимация полиномиальных трендов с помощью многократного сглаживания**.** Р.Г.Брауном и Р.Ф.Майером модель линейного роста была развита путём включения в неё большего количества полиномиальных членов. Обновление во времени оценок коэффициентов полинома производится с помощью многократного экспоненциального сглаживания исходного ряда. Экспоненциальная средняя порядка *p* в момент *t* определяется как

$$S_t^{[p]} = \alpha S_t^{[p-1]} + \beta S_{t-1}^{[p]} \quad (20),$$

где α – постоянная сглаживания $0 < \alpha < 1$, $\beta = 1 - \alpha$. Фундаментальная теорема метода экспоненциального сглаживания и прогнозирования, доказанная Р.Г.Брауном и Р.Ф.Майером, говорит о том, что коэффициенты предсказывающего полинома порядка *n* линейно связаны с экспоненциальными средними до порядка *p*=*n*+1 включительно. Д.А.Д'Эзопо доказал, что для любой последовательности наблюдений полином *P* степени *n*, полученный с помощью многократного экспоненциального сглаживания, является решением, которое минимизирует взвешенную сумму квадратов ошибок

$\alpha \sum_{t=0}^{\infty} \beta^t (x_{t-1} - P_{t-1})^2$.

См. также *Модель Брауна обобщённая*, *Адаптивные методы прогнозирования*, *Параметр адаптации*.

МОДЕЛЬ БРАУНА ОБОБЩЁННАЯ

модель адаптивная, представляющая временной ряд в виде взвешенной суммы некоторых известных, выбранных заранее детерминиро-

ванных функций от времени, и наложенной аддитивной независимой случайной составляющей с нулевым средним (математическим ожиданием) и постоянной дисперсией. Р.Г.Браун разработал рекуррентную процедуру адаптации коэффициентов (весов) модели при каждом получении новой фактической точки ряда для случаев, когда функциями, входящими в модель, являются полиномы, экспоненты и синусоиды или их произведения. См. также *Адаптивные методы прогнозирования*, *Параметр адаптации*.

МОДЕЛЬ ГРАВИТАЦИОННАЯ

модель взаимодействия между пространственными объектами (городами, регионами, странами) в региональном анализе и пространственном анализе экономики. В различных модификациях такие модели используются при исследовании процессов урбанизации, размещения пром-сти, экспортно-импортных взаимосвязей, миграции нас. Общая черта этих моделей заключается в том, что сила взаимодействия (интенсивность потоков) в них зависит от значимости (величины) объектов и расстояния между ними. Общая форма М.г.:

$$I_{ij} = \frac{A \cdot P_j^{\alpha} \cdot P_i^{\beta}}{D_{ij}^{\gamma}},$$

где I_{ij} – объём взаимодействия между объектами i и j; A – коэффициент соответствия; P – некоторая мера значимости объекта (напр., численность нас. города i – P_i и города j – P_j); D_{ij} – расстояние между ними; степенные показатели α, β, γ – параметры модели.

Приведённая формула аналогична физической формуле притяжения между телами, поэтому она получила назв. «М.г.».

Подобные модели применяются при исследовании товарных потоков между парами стран. В них учитываются социально-экономические факторы, определяются экспортные возможности и импортные потребности торговых партнеров, факторы, относящиеся к продвижению товарного потока (расстояние, наличие таможенных барьеров и т. п.).

Адекватность М.г., экономическая обоснованность их применения оспаривается рядом специалистов.

Первые идеи о рыночном взаимодействии в пространстве были высказаны немецким ученым А. Шеффле в последней четверти 19 в. Он предположил взаимодействие торг. и пром-сти и сформулировал схему М.г. Шеффле утверждал, что пром-сть развивается преимущественно в больших городах или поблизости от них. Большие города притягивают к себе пром. пр-тия, причём сила их притяжения обратно пропорциональна квадрату расстояния между ними. Эта модель получила развитие в работах ряда учёных –регионалистов в 20 в. Американский ученый У. Рейли сформулировал «закон гравитации розничной торг.», согласно которому городской рынок притягивает покупателей региона окружающего рынок пропорционально численности городского нас. и обратно пропорционально квадрату расстояния от покупателей до города. Американский экономист Д. Рэй предложил расширенный вариант модели пространственного взаимодействия рыночных потенциалов с учётом влияния финансового капитала. Согласно Рэю, рыночный потенциал терр. является интегральным показателем, характеризующий степень экономического взаимодействия терр. с рассматриваемыми региональными рынками. Значение рыночного потенциала терр. зависит от расстояния до региональных рынков, транспортных издержек, а также от размеров финансового потенциала регионов, с которыми имеется рыночное взаимодействие. В 80 – 90-х гг. 20 в. теория рыночных потенциалов и пространственного взаимодействия получила дальнейшее развитие в работах М. Биркина, Ф. Фоулджера, Х. Уильямса и др., в которых уделяется внимание прогнозированию региональных рынков товаров и услуг на основе моделей пространственного взаимодействия.

МОДЕЛЬ ИМИТАЦИОННАЯ

логико-математическое описание объекта, которое может быть использовано для экспериментирования на компьютере в целях проекти-

рования, анализа и оценки функционирования объекта. Такую модель можно «проиграть» во времени как для одного испытания, так и заданного их множества. При этом результаты будут определяться случайным характером процессов. По этим данным можно получить достаточно устойчивую статистику.

Имитационное моделирование – метод исследования, при котором изучаемая система заменяется моделью, достаточной точностью описывающей реальную систему и с ней проводятся эксперименты с целью получения информации об этой системе. Экспериментирование с моделью называют имитацией (имитация – это постижение сути явления, не прибегая к экспериментам на реальном объекте). Имитационное моделирование – частный случай математического моделирования. Существует класс объектов, для которых по различным причинам не разработаны аналитические модели, либо не разработаны методы решения полученной модели. В этом случае математическая модель заменяется имитатором или М.и.

Цель имитационного моделирования – воспроизведение поведения исследуемой системы на основе результатов анализа наиболее существенных взаимосвязей между её элементами или другими словами – разработке симулятора исследуемой предметной области для проведения различных экспериментов. Выделяются две разновидности имитации: *метод Монте-Карло* (метод статистических испытаний); метод имитационного моделирования (статистическое моделирование).

Различают виды имитационного моделирования: агентное моделирование – относительно новое (1990–2000) направление в имитационном моделировании, которое используется для исследования децентрализованных систем, динамика функционирования которых определяется не глобальными правилами и законами (как в других парадигмах моделирования), а наоборот, когда эти глобальные правила и законы являются результатом индивидуальной активности членов группы. Цель агентных моделей – получить представление об этих глобальных правилах, общем поведении системы, исходя из предположений об индивидуальном, частном поведении ее отдельных активных объектов и взаимодействии этих объектов в системе. Агент – некая сущность, обладающая активностью, автономным поведением, может принимать решения в соответствии с некоторым набором правил, взаимодействовать с окружением, а также самостоятельно изменяться; дискретно-событийное моделирование – подход к моделированию, предлагающий абстрагироваться от непрерывной природы событий и рассматривать только осн. события моделируемой системы, такие как: «ожидание», «обработка заказа», «движение с грузом», «разгрузка» и другие. Дискретно-событийное моделирование наиболее развито и имеет огромную сферу приложений – от логистики и систем массового обслуживания до транспортных и производственных систем. Этот вид моделирования наиболее подходит для моделирования производственных процессов; метод основан Дж. Гордоном в 1960-х гг.

Системная динамика – парадигма моделирования, где для исследуемой системы строятся графические диаграммы причинных связей и глобальных влияний одних параметров на другие во времени, а затем созданная на основе этих диаграмм модель имитируется на компьютере. По сути, такой вид моделирования более всех других парадигм помогает понять суть происходящего выявления причинно-следственных связей между объектами и явлениями. С помощью системной динамики строят модели бизнес-процессов, развития города, модели производства, динамики популяции, экологии и развития эпидемии. Метод основан Форрестером в 1950-х гг.

МОДЕЛЬ КЛАССИЧЕСКАЯ ЛИНЕЙНАЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ (МКЛМР)

модель объясняемой переменной в виде суммы линейной комбинации регрессоров и случайной составляющей, удовлетворяющей определенным требованиям. В общем случае МКЛМР для i-го объекта может быть представлена в виде

$$y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \cdots + \theta_k x_{ki} + \varepsilon_i ,$$

где y_i, x_{li}, ε_i – соответственно значения зависимой переменной, l-го регрессора ($l=1,\ldots,k$) и случайного отклонения (остатка) для i-го наблюдения ($i=1,\ldots,n$). Число наблюдений n должно превышать число регрессоров, кроме того, между регрессорами не должно быть линейной связи – каждый из них вносит в модель определенную информации о зависимой переменной, не содержащуюся в других регрессорах. В классической модели, в отличие от *обобщённой линейной модели множественной регрессии*, остатки должны быть подчинены нормальному закону распределения с нулевым математическим ожиданием и дисперсией $\sigma_i^2=\sigma^2=const$, не зависящей от значений регрессоров (требование гомоскедастичности), и некоррелированы $M(\varepsilon_i\varepsilon_j)=0$ для i≠j (требование хаотичности). При представлении матрицы регрессоров в виде стоки векторов $X=(\underline{1}\quad X_1\quad\cdots\quad X_k)$, первый из которых является вырожденным регрессором в виде столбца единичных значений $\underline{1}=(1\quad 1\quad\cdots\quad 1)^T$ как множителей при коэффициенте θ_0, а остальные – векторы-столбцы значений регрессоров, м.к.м.р. для всех наблюдаемых объектов можно представить в матричном виде $Y=X\Theta+\mathrm{E}$, где $Y=(y_i\quad\cdots\quad y_n)^T$ и $E=(\varepsilon_i\quad\cdots\quad\varepsilon_n)^T$ – векторы зависимой переменной и случайных отклонений, а $\Theta=(\theta_0\quad\theta_1\quad\cdots\quad\theta_k)^T$ – вектор коэффициентов модели. В матричном представлении требования к остаткам МКЛМР могут быть сформулированы более компактно. Вектор остатков должен иметь n-мерное нормальное распределение с нулевым вектором математических ожиданий и скалярной ковариационной матрицей $\Sigma=\sigma^2 I$, где I – единичная матрица. Для оценки коэффициентов МКЛМР, как правило, используют *метод наименьших квадратов*. Вектор оценок коэффициентов определяется выражением $\widehat{\Theta}=(X^TX)^{-1}X^TY$. Все коэффициенты модели должны быть значимы.

МКЛМР широко применяется для моделирования социально-экономических явлений и является стандартной процедурой всех статистических пакетов программ. Т.к. соответствие остатков предъявляемым к ним требованиям изначально неизвестно, оно, как правило, постулируется и затем после построения модели проверяется с помощью различных статистических тестов.

МОДЕЛЬ КЛАССИЧЕСКАЯ ЛИНЕЙНАЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ С ПЕРЕМЕННОЙ СТРУКТУРОЙ

общая линейная модель множественной регрессии для объектов с различными качественными состояниями. Примерами могут служить регрессионные модели заработной платы мужчин и женщин, или работников, имеющих высшее образование и остальных, потребление определённого товара в холодное и теплое время года. Можно строить уравнения внутри каждой категории, а затем изучать различия между ними, но часто более информативна модель в виде единого уравнения сразу по всем категориям, содержащая т.н. фиктивные переменные (или «дамми», от англ. – dummy variable), которые количественным образом отражают качественные изменения моделируемого объекта или явления. Построение отдельного уравнения для каждого качественного состояния часто затруднено или невозможно вследствие недостаточного числа соответствующих наблюдений. Кроме того, статистическая надёжность получаемых оценок коэффициентов единого уравнения в общем случае будет выше, чем оценок, которые могут быть получены отдельно по каждой однородной выборке. Наиболее удобны для использования и интерпретации результатов моделирования бинарные фиктивные переменные, принимающие единичное значение при одном качественном состоянии признака и нулевое – при другом. Коэффициент δ при бинарной фиктивной переменной d в аддитивной линейной модели $y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k+\delta d+\varepsilon$ интерпретируется как отличие среднего значения зависимой переменной y для объектов с единичным значением качественного признака по сравнению со средним для остальных объектов при неизменных значениях прочих регрессоров $x_1,x_2\cdots x_k$. В роли мультипликатора при j-й $(1\le j\le k)$ количественной переменной

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta dx_j + \varepsilon$, фиктивная переменная отражает различное влияние j-го регрессора на зависимую переменную. Если качественная переменная, разделяющая объекты на однородные группы, имеет k градаций, то для отражения её влияния на структуру регрессионной связи, целесообразно вместо неё использовать k-1 бинарных фиктивных переменных. Большее число бинарных переменных приводит к появлению линейной взаимосвязи между ними и вследствие этого – к строгой *мультиколлинеарности* регрессоров. Может быть существенным эффект влияния на зависимую переменную не только различных фиктивных переменных, но и их взаимодействия. Это взаимодействие отражается введением в уравнение регрессии дополнительных фиктивных переменных в виде множителей при произведении взаимодействующих переменных. Если коэффициенты при них окажутся не значимыми, то влиянием соответствующего взаимодействия можно пренебречь.

МОДЕЛЬ КОРРЕКЦИИ ОСТАТКОВ (ОШИБОК)

модель приращений Δy_t *временного ряда* y_t с лагированными (взятыми в предыдущий момент времени) переменными, отражающая одновременно краткосрочную и долгосрочную динамику процесса, Для модели авторегрессионной распределённых лагов с одним лагом эндогенной и одним − экзогенной переменных $y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + \beta_4 y_{t-1} + \varepsilon_t$, $t=1,2,\ldots,n.$, модель коррекции ошибок имеет вид:

$$\Delta y_t = \beta_2 \Delta x_t - (1-\beta_4)\left(y_{t-1} - \frac{\beta_1}{1-\beta_4} - \frac{\beta_2+\beta_3}{1-\beta_4} x_{t-1} \right) + \varepsilon_t ,$$

т.е. изменение переменной y, состоит из двух компонент. Первая пропорциональна текущему изменению Δx_t, и отражает отклик на краткосрочное изменение x. Вторая компонента, пропорциональная выражению в больших скобках, является поправкой, частичной коррекцией на имевшее место в предыдущий момент (t−1) отклонение от долгосрочного равновесия, т.е. «подтягивает» процесс y_t к долгосрочному соотношению с процессом x_t (отсюда и название модели − «коррекции ошибок»). Принципиально такая лаговая структура модели Δy_t сохраняется при увеличении числа лагов эндогенной и экзогенной переменных. Коэффициенты двух моделей − исходной y_t и коррекции ошибок Δy_t связаны невырожденным линейным преобразованием и при условии стационарности переменных (необходимым условием которого является выполнение неравенства $|\beta_4|<1$) могут быть идентично оценены с помощью *метода наименьших квадратов*. Поэтому выбор модели определяется содержательной задачей исследования, а не особенностями процедуры оценивания.

МОДЕЛЬ МАКРОЭКОНОМИЧЕСКАЯ

формализованные (логически, графически и алгебраически) описания различных экономических явлений и процессов с целью выявления функциональных взаимосвязей между ними. Любая модель (теория, уравнение, график и т.д.) является упрощенным, абстрактным отражением реальности, так как все многообразие конкретных деталей не может быть одновременно принято во внимание при проведении исследования. Поэтому ни одна М.м. не абсолютна. Она не даёт единственно правильных ответов, адресованных конкретным странам в конкретный период времени. Однако с помощью таких обобщенных моделей определяется комплекс альтернативных способов управления динамикой уровней занятости, выпуска, инфляции, инвестиций, потребления, процентных ставок, валютного курса и других внутренних (эндогенных) экономических переменных, вероятностные значения которых устанавливаются в результате решения модели. В качестве внешних (экзогенных) переменных, величина которых определяется вне модели, нередко выступают осн. инструменты фискальной политики правительства и монетарной политики ЦБ

РФ – изменения в величинах гос. расходов, налогов и денежной массы.

Многовариантность способов разрешения экономических проблем, обеспечиваемая с помощью М.м., позволяет добиваться необходимой альтернативности и гибкости макроэкономической политики. Использование М.м. даёт возможность оптимизировать сочетания инструментов бюджетно-налоговой, кредитно-денежной, валютной и внешнеторговой политики, успешно координировать меры правительства и ЦБ РФ по управлению циклическими колебаниями экономики. Наиболее перспективными с этой точки зрения являются модели, учитывающие динамику инфляционных ожиданий экономических агентов. Их использование в макроэкономическом прогнозировании позволяет снизить риск возникновения феномена неожиданной инфляции, которая оказывает наиболее разрушительное влияние на экономику, а также смягчить являющуюся одной из самых сложных в макроэкономике проблему недоверия к политике правительства и Центрального Банка.

Такие обобщенные макроэкономические модели, как модель круговых потоков, AD-AS, крест Кейнса, IS-LM, кривые Филлипса, Лаффера, модель Солоу и т.д. представляют собой общий инструментарий макроэкономического анализа и не имеют какой-либо национальной специфики. Специфическими могут быть значения эмпирических коэффициентов и конкретные формы функциональных зависимостей между экономическими переменными в разных странах. Оценка любой М.м. должна даваться по критерию её полезности в процессе познания экономической динамики и управления её показателями.

МОДЕЛЬ МНОЖЕСТВЕННОГО ВЫБОРА

(от англ. – multiple choice model) позволяет моделировать зависимость между результирующей, качественной переменной, определяющей более двух возможных состояний анализируемого объекта, от объясняющих переменных.

М.м.в., имеющие более чем две альтернативы, строятся на основе моделей бинарного выбора (когда зависимая переменная может принимать только два значения). При этом множественный выбор может быть представлен как последовательность бинарных выборов. Обобщением биномиального распределения на случай более чем двух возможных исходов является полиномиальный (мультиномиальный) закон распределения. Полиномиальное распределение используется при статистической обработке выборок большой совокупности, элементы которой разделяются более чем на две категории, применяются в социологических, социально-экономических и медицинских выборочных обследованиях.

В М.м.в. проводится оценка *условной вероятности* выбора альтернативы: модели с неупорядоченными альтернативами (unordered models). Эти модели основаны на предположении, что каждая альтернатива имеет случайный уровень полезности и выбирается альтернатива, приносящая наибольшую полезность; модели с упорядоченными альтернативами (ordered models). Эти модели используются в случаях, когда дискретная зависимая переменная является порядковой, т.е. когда альтернативы естественным образом упорядочены.

МОДЕЛЬ РАСПРЕДЕЛЁННЫХ ЛАГОВ

тип динамических эконометрических моделей, в которых значения переменной за прошлые периоды времени (лаговые переменные) непосредственно включены в модель:

$$y_t = a + b_0 \cdot x_t + b_1 \cdot x_{t-1} + b_2 \cdot x_{t-2} + \ldots + b_p \cdot x_{t-p} + \varepsilon_t .$$

Коэффициент регрессии b_0 (краткосрочный мультипликатор) при переменной x_t характеризует среднее абсолютное изменение y_t при изменении x_t на единицу своего измерения в некоторый момент времени t, без учёта воздействия лаговых значений фактора x.

Промежуточный мультипликатор характеризуется суммой $(b_0 + b_1 + \ldots + b_i)$ и показывает со-

вокупное воздействие факторной переменной x_t на результат y_t в момент времени $(t+i)$, при $i<p$. Долгосрочный мультипликатор характеризуется суммой $(b_0+b_1+...+b_l=b)$с учетом конечной величины лага и отображает абсолютное изменение y_t через l моментов времени при воздействии переменной x_t в момент времени t на 1 единицу.

Относительный коэффициент М.р.л. определяется по формуле:

$$\beta_j = \frac{b_j}{b}.$$

С экономической точки зрения все коэффициенты b_j должны иметь одинаковые знаки: воздействие одного и того же фактора на результат должно быть однонаправленным независимо от того, с каким временным лагом измеряется сила или теснота связи между этими признаками. В этом случае для любого $j=0\div1$ справедливо: $0<\beta_j<1$ и

$$\sum_{j=0}^{l}\beta_j = 1.$$

Тогда относительные коэффициенты β_j являются весами для соответствующих коэффициентов b_j, каждый из которых измеряет долю общего изменения результативного признака в момент $(t+j)$.

Построение М.р.л. имеет свою специфику. Оценка параметров модели в некоторых случаях не может быть произведена с помощью обычного *метода наименьших квадратов* ввиду нарушения его предпосылок.

Текущие и лаговые значения независимой переменной, как правило, тесно связаны друг с другом, оценка параметров происходит в условиях *мультиколлинеарности*. При большой величине *лага* снижается число наблюдений, по которому строится модель, и увеличивается число её факторных признаков, что приводит к потере числа степеней свободы. В М.р.л. часто возникает проблема автокорреляции остатков. Исследователю приходится решать проблему выбора оптимальной величины лага и определения его структуры.

МОДЕЛЬ РЕКУРСИВНАЯ

система рекурсивных уравнений, в которой зависимая переменная (y) одного уравнения выступает в виде фактора (x) в другом. Система рекурсивных уравнений имеет вид:

$$\begin{cases} y_1 = a_{11}\cdot x_1 + a_{12}\cdot x_2 + ... + a_{1m}\cdot x_m + \varepsilon_1, \\ y_2 = b_{21}\cdot y_1 + a_{21}\cdot x_1 + a_{22}\cdot x_2 + ... + a_{2m}\cdot x_m + \varepsilon_2, \\ y_3 = b_{31}\cdot y_1 + b_{32}\cdot y_2 + a_{31}\cdot x_1 + a_{32}\cdot x_2 + ... + a_{3m}\cdot x_m + \varepsilon_3, \\ ..... \\ y_n = b_{n1}\cdot y_1 + ... + b_{n,n-1}\cdot y_{n-1} + a_{n1}\cdot x_1 + a_{n2}\cdot x_2 + ... + a_{nm}\cdot x_m + \varepsilon_n, \end{cases}$$

В общем виде модель имеет вид:

$$\beta_{i1}y_{1t} + \beta_{i2}y_{2t} + ... + \beta_{iG}y_{Gt} + \gamma_{i1}x_{1t} + ... + \gamma_{ik}x_{kt} = \varepsilon_{it},$$

где y_{it} – значения эндогенной переменной в момент времени t; x_{jt} – значение предопределенной переменной, т.е. экзогенной (объясняющей) переменной в момент t или лаговой эндогенной переменной; ε_{it} – случайные возмущения, имеющие нулевые средние.

Систему одновременных уравнений (СОУ) называют рекурсивной, если выполняются следующие условия: матрица значений эндогенных переменных является нижней треугольной матрицей, т.е. $\beta_{ij}=0$ при $j>i$ и $\beta_{ii}=1$; случайные ошибки независимы между собой, т.е. $\sigma_{ii}>0$, $\sigma_{ij}=0$ при $i\neq j$. Отсюда следует, что ковариационная матрица ошибок $M\varepsilon_t\varepsilon_t^T=\Sigma_{(\varepsilon)}$ диагональна; каждое ограничение на структурные коэффициенты относится к отдельному уравнению.

В данной системе каждое последующее уравнение включает в качестве факторов все зависимые переменные предшествующих уравнений наряду с набором собственно факторов x.

Каждое уравнение этой системы может рассматриваться самостоятельно, и его параметры определяются методом наименьших квадратов (МНК). Процедура оценивания коэффициентов рекурсивной системы с помощью метода наименьших квадратов, применённого к отдельному уравнению, приводит к состоятельным оценкам.

Для оценивания СОУ наиболее часто используют *двухшаговый метод наименьших квадратов*, применяемый к каждому уравнению системы в отдельности, и *трёхшаговый метод наименьших квадратов*, предназначенный для оценивания всей системы в целом.

МОДЕЛЬ СКОЛЬЗЯЩЕГО СРЕДНЕГО

модель, которая строится с помощью метода усреднения наблюдений временного ряда, попавших в «окно» шириной в n точек ряда. Это окно движется, «скользит» по оси времени, т.е. вдоль временного ряда, и каждый раз при перемещении «окна» на одну единицу времени параметры модели тренда пересчитываются. Прогноз получают путём экстраполяции модели тренда за пределы «окна» на τ шагов вперед. Расчёт скользящей средней предполагает локальную оценку тренда сначала на первых n=2m+1 точках, где m фиксированное целое число, определяющее ширину «окна» усреднения, m<<n где n – объём выборки. Полученную оценку тренда относим к средней точке периода усреднения, то есть к точке m+1. Далее период усреднения сдвигаем на одну единицу времени вперед и оценку тренда производим на n наблюдениях, начиная со второго и кончая (2m+2)-м. Полученную оценку тренда также относим к средней точке периода усреднения, то есть к точке m+2. Этот процесс продолжаем до конца выборки. Как видим, мы не получаем оценок тренда для первых m и для последних m точек. Далее можно изучать отклонения наблюдений от полученных оценок тренда и, если есть основания предполагать сезонность, можно их трактовать как сезонные или циклические колебания, наблюдаемые в условиях флуктуации. В этом случае период усреднения полагают равным предполагаемому сезонному циклу.

Скользящая средняя часто используется на фондовом рынке в техническом анализе, когда рассчитывают сразу две скользящих средних: медленную (с большим периодом усреднения n) и быструю (с коротким периодом усреднения). Пересечение быстрой скользящей средней медленной сверху вниз интерпретируется как сигнал к продаже, а снизу вверх как сигнал к покупке акций. Отметим одно свойство скользящей средней, на которое указали Е. Слуцкий и Г. Юл, два статистика, которые независимо изучали этот эффект, названный эффектом Слуцкого-Юла. Если в исходном ряде отклонения от стабильного среднего некоррелированы, то рассчитанные скользящие средние характеризуются положительной корреляцией тем большей, чем больше период усреднения (так как соседние периоды усреднения перекрываются), и это может приводить к возникновению видимости циклических колебаний в скользящей средней, которых в исходном ряде на самом деле нет.

Рассмотрим статистические свойства скользящей средней и процедуру оценки полиномиального тренда на n=2m+1 точках. Перенумеруем моменты времени внутри периода усреднения n как t=-m, -(m-1),…,-1, 0, 1,…,m. Если полином имеет нулевой порядок, то исходный ряд локально представляется моделью $x_t = a + u_t$ (1), где a – среднее значение ряда, ut – независимый случайный член с нулевым математическим ожиданием e(ut)=0, дисперсией σ2. Скользящая средняя для текущего периода усреднения вычисляется:

$$S = \frac{1}{2m+1}\sum_{t=-m}^{m} x_t =$$

$$= \frac{1}{2m+1}\sum_{t=-m}^{m}(a+u_t) = a + \frac{1}{2m+1}\sum_{t=-m}^{m} u_t \quad .$$

Математическое ожидание e(s)=a. Т.о., величина s – *оценка несмещённая* среднего уровня ряда. Дисперсия:

$$D(S) = E\left(\frac{1}{2m+1}\sum_{t=-m}^{m} u_t^2\right)^2 = \frac{2m+1}{(2m+1)^2}\sigma^2 = \frac{\sigma^2}{2m+1}. \qquad (3)$$

Т.о., дисперсия скользящей средней будет существенно меньше, чем у исходного ряда xt: при m=1 в 3 раза, а при m=2 в 5 раз. Это означает, что скользящая средняя сглаживает ряд, отфильтровывает случайные колебания.

Реальный прогноз на одну единицу времени вперед за пределами выборочного периода равен последней оценке скользящей средней:

$$\hat{x}_1(n) = S_n = \frac{1}{2m+1}\sum_{t=-m}^{m} x_t = \frac{1}{2m+1}(x_{-m} + x_{-m+1} + ... + x_m) = \frac{1}{2m+1}(x_{n-N+1} + x_{n-N+2} + ... + x_n) \qquad (4)$$

.

В последнем равенстве от индексов времени внутри периода усреднения мы перешли к номерам точек в выборке. Из формулы (4) видно, что здесь прогнозной моделью является авторегрессия порядка N с одинаковыми весами, равными

$$\frac{1}{2m+1},$$

при лаговых значениях переменной. Наилучший порядок авторегрессии N, точнее величина m, определяется методом проб и сопоставления значений критерия суммы квадратов ошибок прогнозов на τ единиц времени вперед. Выбор τ определяется содержанием задачи. При необходимости получения долгосрочных прогнозов, т.е. при больших значениях τ, период усреднения N, очевидно, будет больше, а при краткосрочном прогнозировании – меньше.

В случае, когда тренд имеет приблизительно линейный характер и локально на периоде усреднения N может быть представлен моделью:

$$x_t = a + bt + u_t \qquad (5),$$

где t-m, -(m-1),…,-1, 0, 1,…, m, оценки параметров a и b находятся *методом наименьших квадратов*, т.е. путём минимизации:

$$\sum_{t=-m}^{m}(x_t - a - bt)^2 \qquad (6)$$

.

Дифференцирование по параметрам a и b и приравнивание первых производных нулю дает два уравнения

$$-2\sum_{t=-m}^{m}(x_t - a - bt) = 0,$$

$$-2\sum_{t=-m}^{m} t(x_t - a - bt) = 0,$$

которые нетрудно привести к виду

$$aN + b\sum_{t=-m}^{m} t = \sum_{t=-m}^{m} x_t \;,\; a\sum_{t=-m}^{m} t + b\sum_{t=-m}^{m} t^2 = \sum_{t=-m}^{m} x_t t\,.$$

Но

$$\sum_{t=-m}^{m} t = 0,$$

благодаря выбранной нами системе отсчёта времени от средней точки периода усреднения, и система нормальных уравнений принимает вид:

$$aN = \sum_{t=-m}^{m} x_t \;,\quad b\sum_{t=-m}^{m} t^2 = \sum_{t=-m}^{m} x_t t\,,$$

откуда

$$a = \frac{1}{N}\sum_{t=-m}^{m} x_t \qquad (7),$$

$$b = \frac{\sum_{t=-m}^{m} x_t t}{\sum_{t=-m}^{m} t^2} = \sum_{t=-m}^{m}\left(x_t \frac{t}{\sum_{t=-m}^{m} t^2}\right). \qquad (8)$$

Полученные выражения показывают, что коэффициенты a и b – линейные функции от значений xt, попавших в интервал оценивания, с постоянными параметрами, зависящими только от ширины интервала усреднения N.

В приложении к книге М.Кендэла (1981) можно найти табл. весов для расчёта полиномиальных трендов до пятого порядка включительно и до N=21. Однако в экономических исследованиях практическая полезность трендов выше первого порядка сомнительна.

Скользящая средняя может использоваться для корректировки авторегрессионных и регрессионных моделей путём усреднения ошибок модели, полученных в предыдущие моменты времени.

См. также *Модель авторегрессии со скользящим средним в остатках* (модель Бокса-Дженкинса).

МОДЕЛЬ СМЕСИ РАСПРЕДЕЛЕНИЙ

одна из моделей, получающаяся в результате конструирования тех или иных комбинаций модельных распределений. М.с.р. заданного типа $f_j(x, \theta_j)$ описывается формулой

$$f(x)=\sum_{i=1}^{k} \pi_j f(x_i, \theta_j),$$

в которой $f_j(x, \theta_j)$ и $f(x)$ – плотности (в непрерывном случае) или *полигоны вероятностей* (в дискретном случае) соответственно *j*-ой компоненты смеси и результирующего закона распределения, π_j – априорная вероятность появления в случайной выборке наблюдения с законом распределения $f_j(x, \theta_j)$ (т.е. удельный вес таких наблюдений в общей ген. совокупности), k- число компонент смеси. Законы распределения подобной структуры могут возникнуть, напр., при анализе *ген. совокупности*, объединяющей в себе несколько совокупностей, каждая из которых в определенном смысле однородна (напр., имеет унимодальный закон распределения $f_j(x, \theta_j)$, отличающийся от других параметром θ_j, интерпретируемым как параметр сдвига (центра группирования соответствующих наблюдений) или масштаба (меры случайного рассеивания наблюдений).

МОДЕЛЬ ТЕОРЕТИКО-ЭКОНОМИЧЕСКАЯ

Применение математических методов в экономике началось именно в теоретико-экономических исследованиях. Модели в экономике используются, начиная с 18 в. В «Экономических таблицах» Ф. Кенэ впервые была сделана попытка формализации всего процесса общественного воспроиз-ва. Огромное влияние на экономическую науку оказали схемы воспроиз-ва, созданные К. Марксом и развитые В. И. Лениным. Непосредственное следствие этого подхода – теория межотраслевого баланса.

Процесс экономического исследования с помощью моделей можно условно подразделить на ряд этапов: а) формулировка общей задачи, определение объекта исследования, требований к характеру исходной информации, которая может быть статистической (получаемой в результате наблюдений за ходом экономических процессов) или нормативной (коэффициенты затрат выпуска, рациональные нормы потребления). Изучаются исходные свойства объекта и выдвигаются гипотезы о характере его развития; б) создание модели экономической системы. Модели народно-хозяйственные (балансовые, оптимизационные, равновесные, игровые и др.) исследуют наиболее общие закономерности развития экономики. *Модели макроэкономические* – анализ и прогноз динамики и соотношения различных синтетических показателей (национального дохода, занятости, потребления, сбережений, инвестиций и т.п.). Микроэкономической модели – исследование конкретных хоз. ситуаций (модели произ-ва, транспорта, торг., снабжения и сбыта и т.п). Экономические модели классифицируются по: целям и задачам, объекту, применяемому аппарату исследования, характеру исходной информации (статистические и нормативные модели). Все классификации условны, реальные модели могут занимать промежуточное положение; в) математический анализ модели – средство получения количественных и качественных выводов. Качественные выводы позволяют обнаружить неизвестные ранее свойства экономической системы: структуру, динамику развития, устойчивость, соотношения макроэкономических параметров, свойства ценностных показателей и т.п. Количественные выводы – оптимальные планы развития тех или иных хозяйственных ячеек, прогнозы экономической динамики, расчёты цен; г) проверка полученных результатов. Часто применяемое в естественных науках сопоставление полученных результатов эксперимента, с характеристиками реальных процессов – не всегда применимо. Важна теоретическая проверка правильно-

сти исходных предпосылок модели; д) внедрение: этот этап должен приводить (в случае положительного исхода предшествующего этапа) к совершенствованию экономической теории и методов управления экономическими процессами, цен, планов развития. В противном случае необходимо уточнить исходные предпосылки модели.

Использование моделей в экономике имеет определённые границы применения: ограничения по формализации и доступности информации, возможности проведения теоретического анализа. На практике применение моделей в экономике дополняется использованием оценок экспертов.

МОДЕЛЬ ЭКОНОМЕТРИЧЕСКАЯ

экономико-математическая модель, которая выражает статистические зависимости между показателями**,** характеризующими состояние и развитие конкретной социально-экономической системы. М.э. используется для *прогнозирования* состояния анализируемой системы, а также имитации возможных сценариев её развития. В основе её построения лежит предположение о том, что поведение экономической системы определяется с помощью совместных и одновременных операций с некоторым числом экономических соотношений. М.э. – совокупность уравнений регрессии и тождеств, описывающих связи между анализируемыми экономическими показателями. Наиболее распространенным выражением М.э. является линейная *система одновременных уравнений* (СОУ), в которой одни и те же показатели в разных уравнениях могут быть как объясняемыми, так и объясняющими переменными**.** Все переменные, участвующие в М.э., подразделяют на экзогенные, эндогенные и предопределённые. *Экзогенные переменные* задаются вне модели, несут смысловую нагрузку управляемых (планируемых) показателей, значения эндогенных переменных формируются внутри функционирования анализируемой системы, а множество предопределённых переменных содержит все экзогенные переменные и лаговые (относящихся к прошлым моментам времени) *эндогенные переменные*. Т.о., М.э. служит для объяснения поведения эндогенных переменных в зависимости от значений предопределенных переменных.

Исходная форма представления М.э. в виде линейной СОУ – её структурная форма, которая отражает связи между переменными в соответствии с их экономической сущностью. Путём разрешения системы относительно эндогенных переменных её можно свести к приведённой форме**.** Структурная форма линейной СОУ, содержащей m_1 уравнений регрессии и m_2 тождеств, включающей *m* эндогенных переменных $m = m_1 + m_2$ и (*p*+1) предопределённых переменных, имеет вид: $BY_t + CX_t = \overline{\Delta}_t$, $t = 1,2,...,n$,

t – номер наблюдения;

$Y_t = (y_t^{(1)}, y_t^{(2)},..., y_t^{(m)})^T$ – вектор эндогенных переменных;

$X_t = (x_t^{(1)}, x_t^{(2)},..., x_t^{(p+1)})^T \quad (x_t^{(1)} = 1)$ – вектор предопределённых переменных;

$\overline{\Delta}_t = \begin{pmatrix} \Delta_t \\ 0_{m_2} \end{pmatrix}$ – составной вектор, в котором:

$\Delta_t = (\delta_t^{(1)}, \delta_t^{(2)},..., \delta_t^{(m_1)})$ – вектор случайных составляющих в уравнениях регрессии,

$0_{m_2} = (\underbrace{0,0,...,0}_{m_2})$ – вектор размерности m_2 из нулей, стоящих в правых частях тождеств;

$B = \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix}_{m \times m}$, $C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}_{m \times (p+1)}$

– матрицы коэффициентов при переменных Y_t и X_t.

$$B_1 = \left\| \beta_{ij}^{(1)} \right\|_{m_1 \times m_1}, \; B_2 = \left\| \beta_{ij}^{(2)} \right\|_{m_1 \times m_2},$$
$$B_3 = \left\| \beta_{ij}^{(3)} \right\|_{m_2 \times m_1}, \; B_4 = \left\| \beta_{ij}^{(4)} \right\|_{m_2 \times m_2},$$
$$C_1 = \left\| c_{ij}^{(1)} \right\|_{m_1 \times (p+1)}, \; C_2 = \left\| c_{ij}^{(2)} \right\|_{m_2 \times (p+1)}.$$

При этом предполагается, что коэффициент при *i*-й эндогенной переменной в *i*-м уравнении $(i = 1,2,...,m)$ равен единице, а матрицы B_4 и *B* невырождены.

Соответствующая данной структурной форме приведённая форма М.э. имеет вид:

$$Y_t = \Pi X_t + \varepsilon_t, \quad t = 1,2,...,n,$$

$$\Pi = \left\| \pi_{ij} \right\|_{m\times(p+1)} = -B^{-1}C\,,\ \varepsilon_t = B^{-1}\bar{\Delta}_t\,.$$

Элементы матриц B_3, B_4, C_2 являются известными и определяются содержательным смыслом соответствующих тождеств системы. Элементы матриц B_1, B_2, C_1 оцениваются на основе *исходных статистических данных,* представляемых в виде матриц:

$$Y = \begin{pmatrix} Y_1^T \\ Y_2^T \\ \dots \\ Y_n^T \end{pmatrix}_{n\times m}\ ,\ X = \begin{pmatrix} X_1^T \\ X_2^T \\ \dots \\ X_n^T \end{pmatrix}_{n\times(p+1)}\ .$$

В процессе эконометрического моделирования обычно приходится решать следующие проблемы. Первая из них, проблема спецификации модели, включает в себя: а) определение конечных целей моделирования; б) определение списка экзогенных и эндогенных переменных; в) определение состава анализируемой системы уравнений и тождеств, их структуры и соответственно списка предопределенных переменных; г) формулировку исходных предпосылок и априорных ограничений относительно стохастической природы остатков и числовых значений отдельных коэффициентов.

Спецификация опирается на имеющиеся экономические теории, знания и интуицию исследователя об анализируемой системе. Обычно принимаются допущения о том, что все случайные остатки $\delta_t^{(j)},\ \ (j = 1,2,...,m_1)$ – имеют нулевые средние значения:

$$\mathrm{E}\delta_t^{(j)} \equiv 0,\ \ t = 1,2,...,n;$$

не коррелируют друг с другом:

$$\mathrm{E}(\delta_t^{(i)} \cdot \delta_t^{(j)}) = 0,\ \ i \neq j;$$

не имеют автокорреляций:

$$\mathrm{E}(\delta_{t_1}^{(j)} \cdot \delta_{t_2}^{(j)}) = 0,\ \ t_1 \neq t_2;$$

не коррелируют ни с одной из предопределенных переменных.

Обычным также является предположение о линейности связей между переменными. В тех случаях, когда оно отсутствует, и анализируемые зависимости между переменными не являются линейными, то либо используют процедуру *линеаризации* модели путём подбора подходящего преобразования исходных переменных, либо решают задачу регрессионного анализа в общем виде, реализуя базовую идею *метода наименьших квадратов*.

Проблема *идентифицируемости* состоит в определении возможности восстановления структурной формы по оцененной приведённой форме, которая связана с необходимостью получения характеристик внутренней (истинной) структуры анализируемой системы. Эта проблема обусловлена существенным превышением числа неизвестных коэффициентов в матрицах *B* и *C* структурной формы над числом неизвестных элементов в матрице П приведённой формы и не может быть решена без выполнения некоторых дополнительных ограничений (условий идентифицируемости). Проблема идентификации заключается в выборе и реализации методов *статистического оценивания параметров*. В зависимости от вида М.э. используют различные методы: *метод наименьших квадратов* (МНК), *обобщённый МНК*, косвенный МНК, *двухшаговый МНК, трёхшаговый МНК, метод макс. правдоподобия* и др.

Проблема верификации модели заключается в оценке качества М.э. в смысле её адекватности реальной действительности и возможности использования в практических целях, а также оценке точности прогноза и имитационных расчетов, полученных по модели. Методы верификации М.э. основаны на процедурах статистической проверки гипотез и на анализе показателей точности методов статистического оценивания. Наиболее распространённым и эффективным подходом к верификации М.э. является принцип ретроспективных расчётов, идея которого состоит в разделении исходного массива данных на обучающую и экзаменующую выборки, построения модели по первой выборке и сравнении модельных (прогнозных) значений эндогенных переменных со значениями из второй выборки. Специфика М.э., связанная с необходимостью её «настраивания» на конкретную социально-экономическую систему, накладывает повышенные требования к качеству и количеству исходных статистических данных и обуславливает дополнение классического набора методов регрессионного анализа широким спектром методов многомерного статистического анализа и, в первую очередь, ме-

тодами *распознавания образов*, их типологизации и снижения размерности исследуемого факторного пространства.

См. также *Системы одновременных уравнений (СОУ)*.

МОДЕЛЬ ЭКОНОМИКО-МАТЕМАТИЧЕСКАЯ

модель экономического явления или процесса, записываемая с помощью одного или нескольких математических выражений (уравнений, функций, неравенств, тождеств), характеризующих важнейшие взаимосвязи явлений и процессов, условия и закономерности их развития, ограничения, требования и т.д. М.э.-м. обобщает существенную качественную и количественную информацию об объекте анализа и служит базой для проведения расчётных экспериментов, которые позволяют получить различные характеристики и параметры изучаемого объекта для заданных условий его развития.

Построение и использование М.э.-м. существенно расширяет возможности экономического анализа. Оно обеспечивает одновременный учёт большого числа требований, условий и предположений, а также известную свободу в пересмотре этих условий в ходе работы с моделью, непротиворечивость получаемых систем показателей, возможность получения вариантов поведения изучаемого явления для широкого диапазона и сочетания исходных условий и предположений. М.э.-м. по назначению делятся на *модели теоретико-экономические* и прикладные. Многие прикладные модели являются *моделями экономико-статистическими* или включают последние в качестве составных частей.

МОДЕЛЬ ЭКОНОМИКО-СТАТИСТИЧЕСКАЯ

система математических соотношений, которая описывает некоторый экономический объект, и параметры которой определяются с помощью статистических методов на основании фактических данных. Структура и конкретный вид М.э.-с. определяются спецификой моделируемого объекта, теоретическими представлениями исследователя об объекте, целями исследования, доступной информацией, используемыми статистическими методами и информационными технологиями. Процесс моделирования состоит из нескольких этапов: на первом этапе определяется общий вид модели, входящие в неё *экзогенные переменные* и *эндогенные переменные*, соотношения между ними; на втором этапе производится статистическое оценивание неизвестных параметров модели на основе данных наблюдений; далее с помощью модели решают задачи анализа состояния моделируемого объекта, *прогнозирования* его развития. См. также *Модель имитационная*, *модель эконометрическая*, *модель экономико-математическая*.

МОДЕЛЬ ARCH

(модель авторегрессионная условной гетероскедастичности, от англ. –autoregressive conditional heteroscedasticity) – модель для дисперсии ошибок регрессионного уравнения, в которой безусловная дисперсия ошибок (т.е. на длительных отрезках времени) предполагается постоянной, а на коротких отрезках времени – переменной (отсюда и название – условная гетероскедастичность, т.е. условная изменчивость дисперсии). Если регрессионное уравнение имеет вид

$$y_t = \beta_1 + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + u_t \qquad (1),$$

где u_t – ошибка уравнения, то, если условную дисперсию ошибки в момент t обозначить как h_t^2, модель ARCH в простейшем случае записывается как $h_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2$ (2). Предложена эта модель Р. Энглом в 1982. Последнее уравнение означает следующее: если ошибка в предыдущий момент была большой, то и дисперсия текущей ошибки будет большой. Если же коэффициент α_1=0, то это означает, что эффекта ARCH не наблюдается, то есть дисперсия постоянна – имеет место гомоскедастичность и можно для оценивания первого (регрессионного) уравнения обычный метод наименьших квадратов (МНК). При $\alpha_1 \neq 0$ для оценивания модели используются более сложные методы максимального правдоподобия. Проверка нуле-

вой гипотезы H_0: $\alpha_1=0$ выполняется с помощью оценивания МНК первого уравнения, получения остатков u_t и оценивания регрессии квадратов остатков u_t^2 на их лаговые значения u_{t-1}^2 (со свободным членом α_0) и проверки значимости коэффициента при u_{t-1}^2.

Если гипотеза о гетероскедастичности принимается, то полученные оценки дисперсии h_t^2 применяются для новой оценки первого уравнения модели уже с учетом характера гетероскедастичности. Далее могут быть получены уточненные оценки дисперсии и заново оценено первое уравнение. Этот итерационный процесс может быть повторен до достижения сходимости.

Эффекты ARCH обнаружены у курсов акций и у других активов, с которыми производят спекулятивные операции. Эффекты ARCH признаны полезными в моделях инфляции, когда последовательности больших и малых ошибок прогнозирования образуют кластеры, то есть следуют «гроздьями». Все это объясняется краткосрочным ажиотажным спросом, возникающим при очередном повышении цен и ожиданиях их дальнейшего роста.

Уточненная (локальная, текущая) дисперсия остаточного члена позволяет гибко строить доверительные интервалы для будущих значений эндогенной переменной, т.е. для y_{t+1}.

Модель Энгла расширенная, в общем виде она включает большее число *лагов* остаточного члена: $h_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + ... + \alpha_a u_{t-a}^2$, называется моделью ARCH(q) порядка q. Безусловная дисперсия u_t постоянна и равна

$$\mathrm{var}(u_t) = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i} > 0.$$

Используется, в частности, для прогнозирования волатильности ценных бумаг на фондовом рынке, как показателя риска. См. также *Модель GARCH*.

МОДЕЛЬ GARCH

обобщённая модель авторегрессионная условной гетероскедастичности (от англ. – generalized autoregressive conditional heteroscedasticity); используется для дисперсии ошибок регрессионного уравнения, в которой безусловная дисперсия ошибок (т.е. на длительных отрезках времени) предполагается постоянной, а на коротких отрезках времени – переменной (отсюда и название – условная гетероскедастичность, т.е. условная изменчивость дисперсии); является расширением модели ARCH. Введена Т. Болерслевым в 1986. Если регрессионное уравнение имеет вид:

$$y_t = \beta_1 + \beta_2 x_{2t} + ... + \beta_k x_{kt} + u_t \qquad (1),$$

где u_t – ошибка уравнения, то, если условную дисперсию ошибки в момент t обозначить как h_t^2, модель GARCH для дисперсии включает авторегрессию порядка p и скользящую среднюю порядка q:

$$h_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + ... + \alpha_q u_{t-q}^2 + \varphi_1 h_{t-1}^2 + ... + \varphi_p h_{t-p}^2 \quad (2)$$

что записывается кратко как GARCH(p,q). Безусловная дисперсия постоянна и равна:

$$\mathrm{var}(u_t) = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{i=1}^{p} \phi_i} .$$

Разработано целое семейство М. GARCH, в которых различным образом специфицируется уравнение, определяющее h_t^2, и учитывается влияние на эндогенную переменную y_t.

Модель GARCH-M (от англ. – GARCH-in-mean), т.е. обобщённая модель ARCH в среднем значении, кратко обозначается как GARCH(p,q)-M и имеет спецификацию:

$$y_t = \beta_1 + \beta_2 x_{2t} + ... + \beta_k x_{kt} + \gamma h_t^2 + u_t \quad (4), \quad (3)$$

где модель для условной дисперсии h_t^2 определена выражением (2). Т.о., среднее значение зависимой переменной y_t в момент t зависит от условной дисперсии остаточного члена u_t, представляемой в свою очередь *моделью авторегрессии-скользящего среднего*, откуда и происходит общее название модели. Необходимое условие ковариационной стационарности (1):

$$\sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \phi_i < 1 .$$

В дополнение к ограничениям (3) и (5) Болерслев предполагает также, что $\alpha_i \geq 0$, $i=1, 2, \ldots, q$ и $\phi_i \geq 0$, $i=1, 2, \ldots, p$. Этих дополнительных ограничений достаточно для того, чтобы условная дисперсия была положительной, но они не являются необходимыми.

Модель AGARCH – модель GARCH(p,q), в которой условная стандартная ошибка u_t в (1) определяется:

$$h_t = \sqrt{\mathrm{var}(u_t)} = \alpha_0 + \sum_{i=1}^{q} \alpha_i \, | u_{t-i} | + \sum_{i=1}^{p} \phi_i h_{t-i} \qquad (6)$$

где

$$\mu = E\left(\left| \frac{u_t}{h_t} \right| \right).$$

Значение μ зависит от функции плотности вероятностей, которая, по предположению, описывает стандартизованные возмущения

$$\varepsilon_t = \frac{u_t}{h_t}.$$

Эта модель, разработанная Нельсоном в 1991, допускает асимметричные воздействия прошлых ошибок на условные дисперсии ошибок.

Модель EGARCH-M – модель EGARCH(p,q) в среднем значении (EGARCH*(p,q)*-in-mean), специфицируемая уравнениями (4) и (7). См. также *Модель ARCH*.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

тесная корреляционная зависимость между объясняющими признаками в множественной линейной регрессионной модели. Тогда один из таких признаков можно представить в виде линейной комбинации других. Равенство выполняется тем точнее, чем сильнее М. В социально-экономических исследованиях часто возникает из-за дублирования информации, содержащейся в различных показателях. При этом определитель *ковариационной* и *корреляционной матрицы* стремится к нулю. Тогда миним. собственное число этой матрицы стремится к нулю и увеличивается различие между макс. и миним. собственными числами. Наличие почти функциональной связи одного из объясняющих признаков с другими приводит тому, что соответствующий множественный коэффициент приближается к единице. М. может проявляться явно, если в матрице *парных коэффициентов корреляции* ряд коэффициентов между объясняющими признаками принимает значения, по модулю больше, чем 0,8. Но отсутствие таких коэффициентов не является гарантией отсутствия самой М. При большом числе регрессоров может возникнуть эффект, когда все парные коэффициенты между регрессорами указывают не слабую связь, но в то же время хотя бы для одного регрессора множественный коэффициент корреляции его со всеми остальными регрессорами окажется близок к 1. Поэтому требуется строгая проверка гипотезы о том, можно ли считать ортогональной систему координат, построенную на исходных признаках. Если система существенно неортогональна, то имеет место М. Наличие М. приводит к существенному ухудшению статистических свойств построенной регрессионной модели. Уравнение становится неустойчивым, т.е. незначительное изменение состава выборки (добавление, удаление или замена всего нескольких объектов) может привести к существенному изменению коэффициентов уравнения (хотя вычисленные по модели значения результативного признака могут мало измениться). Т.е. коэффициенты регрессии перестают адекватно отражать роль регрессоров для объяснения результативного признака. Поэтому от М. избавляются до начала построения регрессионной модели (на

Модель AGARCH-M – модель AGARCH в среднем значении (AGARCH-in-mean) специфицируется уравнениями (1) и (6).

Модель EGARCH – экспоненциальная модель GARCH(p,q), в которой логарифм условной дисперсии ошибки уравнения (1) имеет спецификацию:

$$\ln h_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \left(\frac{u_{t-i}}{h_{t-i}} \right) + \sum_{i=1}^{q} \alpha_i^* \left(\left| \frac{u_{t-i}}{h_{t-i}} \right| - \mu \right) + \sum_{i=1}^{p} \phi_i \ln h_{t-1}^2 \qquad (7),$$

называется абсолютной GARCH и обозначается AGARCH(p,q). Она также содержит авторегрессионную часть порядка p и скользящую среднюю порядка q.

этапе содержательного анализа данных). Если полностью избавиться от этого негативного явления не удалось, то исследователь применяет при построении регрессионной модели пошаговые процедуры регрессионного анализа, позволяющие отобрать для модели только наиболее информативные регрессоры, сведя дублирование информации в них к минимуму. Если содержательная интерпретация построенного уравнения не вполне устраивает исследователя (некоторые «полезные» регрессоры не включены в модель), то он применяет методы снижения размерности (метод главных компонент, а если этого недостаточно, то факторный анализ), а затем строит уравнение регрессии на несколько первых, наиболее информативных обобщенных показателей.

При классификации объектов М. также может существенно исказить результаты, поэтому используется переход из косоугольной исходной системы координат в ортогональную, построенную на главных компонентах или общих факторах. Другой способ – использование метрики Махалонобиса, учитывающей косоугольность пространства и различный масштаб по координатным осям с помощью ковариационной матрицы.

Н

НЕИДЕНТИФИЦИРУЕМОСТЬ

невозможность перехода от приведённой формы записи модели одновременных эконометрических уравнений к структурной форме. Модель неидентифицируема, если не все её структурные коэффициенты определяются однозначно по коэффициентам приведённой модели. В этом случае, число приведённых коэффициентов меньше числа структурных коэффициентов, и структурные коэффициенты не могут быть оценены через коэффициенты приведённой формы. Структурная запись модели – система совместных уравнений, каждое из которых должно быть проверено на *идентифицируемость*. Если хотя бы одно из уравнений системы неидентифицируемо, то и вся модель считается неидентифицируемой. Если в любом уравнении системы число отсутствующих предопределенных переменных меньше числа *эндогенных переменных*, то уравнение и вся система в целом будут неидентифицируемы. В этом случае число источников независимой информации в уравнении меньше общего числа регрессоров. Проблема Н. – это проблема структуры модели (т.е. числа уравнений, соотношения количеств эндогенных и предопределённых переменных в системе в каждом уравнении, *мультиколлинеарности* анализируемых переменных, некоторых свойств матриц структурных коэффициентов), но никак не связаны со статистическими свойствами исходных наблюдений, напр. их количеством. Н. не исчезает с ростом количества наблюдений и означает, что существует бесконечное число структурных моделей, имеющих одну и ту же приведённую форму. Н. не является редким явлением и довольно часто распространена в *моделях эконометрических*. Н. модели приводит к невозможности применения косвенного *метода наименьших квадратов* для оценки параметров структурной модели.

О

ОБОБЩЁННАЯ ЛИНЕЙНАЯ МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ (ОЛММР)

(от англ. – Generalized Linear Multiple Regression model) – модель множественной регрессии в условиях нарушения требований, предъявляемых к случайным регрессионным остаткам в классической модели, т.е. когда регрессионная модель рассматривается в условиях *гетероскедастичности* и/или взаимной коррелированности случайных регрессионных остатков. В матричном виде ОЛММР можно записать:

$$\begin{cases} Y = X\beta + \varepsilon \\ M\varepsilon = 0_n \\ \Sigma_\varepsilon = \sigma^2 \Sigma_0 \end{cases},$$

где

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

– вектор значений зависимой переменной,

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

– матрица значений объясняющих переменных,

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

– вектор параметров модели,

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

– случайный вектор регрессионных остатков,

$$0_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

– нулевой вектор, Σ_0 – некоторая положительно определённая матрица

порядка n x n , где n – число статистических наблюдений, σ^2 – скаляр.

($x^{(1)}$, $x^{(2)}$, …, $x^{(p)}$) – объясняющие переменные, рассматриваемые как неслучайные переменные, ранг матрицы X равен $p+1 < n$.

ОЛММР отличается от классической линейной модели множественной регрессии (КЛММР) только видом *ковариационной матрицы*. На месте единичной матрицы E_n в описании ковариационной матрицы остатков Σ_ε содержится матрица $\sigma^2\Sigma_0$. Это означает, что дисперсии и корреляции остатков могут быть произвольными при условии невырожденности матрицы Σ_0. Тогда как в КЛММР остатки некоррелированные и гомоскедастичные.

ОБОБЩЁННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (ОМНК)

модификация обычного *метода наименьших квадратов* (МНК), позволяющая получать эффективные оценки коэффициентов линейной модели множественной регрессии (ЛММР). ОМНК применяется в условиях обобщенной ЛММР, т.е. при нарушении классической «скалярной» структуры ковариационной матрицы вектора регрессионных отклонений:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \quad ,$$
$$E\boldsymbol{\varepsilon} = O_n, V(\boldsymbol{\varepsilon}) = E\varepsilon\varepsilon^T = \Omega \neq \sigma^2\mathbf{I}_n \qquad (1)$$

Наличие в матрице Ω несовпадающих элементов на гл. диагонали (свойство *гетероскедастичности* регрессионных ошибок) и/или ненулевых элементов вне диагонали (свойство коррелированности ошибок) приводит к невыполнению теоремы Гаусса-Маркова: МНК-оценки коэффициентов регрессии, оставаясь несмещёнными и состоятельными перестают быть наилучшими. ОМНК решает задачу нахождения наилучших оценок, суть этого метода состоит в таком преобразовании переменных модели, при котором ковариационная матрица регрессионных ошибок становится «скалярной», и, далее, применении обычного МНК к преобразованной модели.

Пользуясь известным фактом о существовании, для любой положительно определённой матрицы Ω, невырожденной матрицы P, такой, что $P\Omega P^T = I$ (или, что эквивалентно, $\Omega^{-1} = P^T P$), находятся новые регрессионные переменные $\tilde{X} = PX, \tilde{y} = Py$ и преобразованные регрессионные ошибки $\tilde{\varepsilon} = P\varepsilon$, при этом умножение слева обоих частей уравнения (1) на P приводит к соотношению

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}} \ , \ E\tilde{\boldsymbol{\varepsilon}} = O_n, V(\tilde{\boldsymbol{\varepsilon}}) = \mathbf{I}_n \qquad (2)$$

Применяя к (2) обычный МНК, получается вектор ОМНК-оценок

$$\hat{\beta}_{ОМНК} = \left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}^T\tilde{y},$$

который в терминах исходных переменных задается формулой:

$$\hat{\beta}_{ОМНК} = \left(X^T\Omega^{-1}X\right)^{-1}X^T\Omega^{-1}y \quad (3).$$

Вектор ОМНК-оценок – решение оптимизационной задачи:

$$\left(y - Xb\right)^T\Omega^{-1}\left(y - Xb\right) \to \min_b \qquad (4)$$

ОМНК-оценки являются *оценками несмещёнными* регрессионных коэффициентов модели (1).

Роль обобщённой характеристики точности полученных оценок играет ковариационная матрица, которая в рамках ОЛММР имеет вид:

$V(\hat{\beta}_{МНК}) = \left(X^T \Omega^{-1} X \right)^{-1}$.

ОМНК-оценки коэффициентов β в условиях ОЛММР – более точны, чем МНК-оценки. Более того, ОМНК-оценки являются лучшими (эффективными, в смысле минимума дисперсии) в классе линейных по $y_1, y_2, \ldots, y_n$ несмещённых оценок коэффициентов.

Для реализации ОМНК требуется знание матрицы Ω, что в реальных ситуациях случается не часто. На практике неизвестную матрицу Ω в соотношении (3) заменяют какой-либо состоятельной (при $n \to \infty$) оценкой $\hat{\Omega}$ и, т.о., приходят к практически реализуемому (или доступному) ОМНК, позволяющему получать асимптотически эффективные оценки коэффициентов регрессии. Способ получения оценки $\hat{\Omega}$ зависит от того, предполагается ли наличие гетероскедастичности, коррелированности или их сочетания. В любом случае исходят из знания структуры матрицы ковариаций Ω, т.е. представления всех её элементов в виде функций с фиксированным (независящим от объема наблюдений) количеством параметров.

Важным примером является ситуация, когда все элементы Ω известны с точностью до одного неизвестного множителя: $\Omega = \sigma^2 \Omega_0$, где σ^2 – неизвестный параметр, Ω_0 – известная матрица. В этом случае вектор ОМНК-оценок равен:

$$\hat{\beta}_{ОМНК} = \left(X^T \Omega_0^{-1} X \right)^{-1} X^T \Omega_0^{-1} y, \qquad (5)$$

а его ковариационная матрица

$V(\hat{\beta}_{МНК}) = \sigma^2 \left(X^T \Omega_0^{-1} X \right)^{-1}$.

Для оценки матрицы $V(\hat{\beta}_{МНК})$ достаточно заменить параметр σ^2 на его несмещенную оценку:

$$s^2 = \frac{1}{n-k} (\mathbf{y} - \mathbf{X}\hat{\beta}_{ОМНК})^{\mathrm{T}} \Omega_0^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}_{ОМНК})$$

Весьма распространённые частные случаи применения практически реализуемого ОМНК – две ситуации: 1. регрессионные ошибки являются гетероскедастичными и некоррелированными. Тогда $\Omega = diag\left\{\sigma_1^2, \ldots, \sigma_n^2\right\}$, где $D\varepsilon_i = \sigma_i^2 \neq const$ – дисперсии ошибок. Можно показать, что в этом случае матрица преобразования P, приводящая обобщённую модель (1) к модели (2) с классическими ошибками, имеет вид $P = diag\left\{1/\sigma_1, \ldots, 1/\sigma_n\right\}$, что соответствует делению *i*-го регрессионного уравнения модели (1) на

σ_i: $y_i/\sigma_i = \beta_1 x_i^{(1)}/\sigma_i + \ldots + \beta_k x_i^{(k)}/\sigma_i + \varepsilon_i/\sigma_i$, $i = 1, \ldots, n$.

Применение обычного МНК к преобразованной системе соответствует решению задачи оптимизации, в которой минимизируемый функционал (4) записывается в виде:

$$\sum_{i=1}^{n} \left[y_i/\sigma_i - \left(\beta_1 x_i^{(1)}/\sigma_i + \ldots + \beta_k x_i^{(k)}/\sigma_i \right) \right]^2 = \sum_{i=1}^{n} 1/\sigma_i^2 \left[y_i - \left(\beta_1 x_i^{(1)} + \ldots + \beta_k x_i^{(k)} \right) \right]^2,$$

т.е. представляет собой взвешенную сумму квадратов. Поэтому в случае гетероскедастичных и некоррелированных регрессионных отклонений ОМНК называют взвешенным. Как видно, наблюдения, соответствующие большим значениям дисперсий отклонений, будут вносить меньший вклад в формирование оценок (входят в функционал с меньшим весом), чем наблюдения, соответствующие меньшим дисперсиям отклонений. Для реализации взвешенного ОМНК требуется либо знание дисперсий отклонений, либо замена этих дисперсий состоятельными оценками; 2. регрессионные ошибки описываются моделью авторегрессии первого порядка.

ОБЪЯСНЯЮЩАЯ ПЕРЕМЕННАЯ

в эконометрических моделях переменная, значение которой известно исследователю и которая выступает в модели в роли фактора, влияющего на результат моделирования. О.п. также называют независимой, входной, предсказывающей, предикторной, *экзогенной переменной*, фактором, регрессором, факторными признаком. В любой *модели эконометрической* значения О.п. задаются как бы «извне» (отсюда ещё одно название «экзогенная переменная), автономно, в определённой степени их значения управляемые (планируемые), варьируются исследователем. Напр., формирующийся на рын-

ке спрос на некоторый товар рассматривается как функция его цены. В этом случае переменная, характеризующая цену на товар, будет выступать в виде О.п. – фактора по отношению к зависимой переменной, характеризующей спрос на товар.

П

ПАНЕЛЬНЫЕ ДАННЫЕ

совокупность результатов наблюдений одних и тех же экономических единиц, осуществленных в последовательные периоды времени, или совокупность нескольких временных рядов для некоторого множества объектов, напр., стран, регионов, фирм или случайной выборки домашних хоз-в.

Использование П.д. предоставляет исследователю ряд существенных преимуществ. П.д. характеризуются большим числом наблюдений и большей вариацией переменных, чем выборочные данные за один период или временной ряд для одного объекта. Т.к. П.д. содержат больше информации, появляется возможность получать более точные оценки и тестировать более сложные модели при менее жестких ограничениях. Кроме того, П.д. позволяют контролировать присутствие ненаблюдаемых, но постоянных во времени или меняющихся во времени, но постоянных для объектов факторов. П.д. предоставляют лучшие возможности для обнаружения эффектов, которые невозможно исследовать по выборочным данным за один период или одиночному временному ряду. Наконец, П.д. позволяют исследовать динамические аспекты поведения отдельных объектов.

Для анализа П.д. предложено богатое семейство моделей. Несмотря на то, что в случае П.д. могут использоваться модели со случайными коэффициентами, в большинстве случае исследователи ограничиваются моделью линейной регрессии с компонентами ошибок:

$$y_{it} = X_{it}\beta + \alpha_i + \lambda_t + \varepsilon_{it},\ i = 1,...,n,\ t = 1,...,T,$$

где i – номер объекта, t – время, X_{it} – вектор значений k независимых переменных, β – вектор неизвестных коэффициентов регрессии, α_i – ненаблюдаемый индивидуальный эффект, λ_t – ненаблюдаемый временной эффект, ε_{it} – независимые случайные величины с нулевым средним и дисперсией σ_ε^2.

Если α_i и λ_t рассматривают как неизвестные параметры подлежащие оценке, то получают модель с фиксированными эффектами. Если α_i и λ_t случайные величины нулевым средним и дисперсиями σ_α^2 и σ_λ^2 – модель со случайными эффектами. Оценки коэффициентов модели с фиксированными эффектами получаются *методом наименьших квадратов* после перехода к отклонениям от средних $\tilde{y}_{it} = y_{it} - y_{i.} - y_{.t} - y_{..}$, и аналогичного преобразования для X_{it}. Для проверки значимости индивидуальных и временных эффектов может использоваться стандартный F-критерий.

Оценки коэффициентов модели со случайными эффектами также могут быть получены методом наименьших квадратов путём регрессии $\tilde{y}_{it} = y_{it} - \theta_1 y_{i.} - \theta_2 y_{.t} - \theta_3 y_{..}$, где $\theta_1, \theta_2, \theta_{3..}$ – выражаются через компоненты дисперсии σ_ε^2, σ_α^2 и σ_λ^2. Для проверки значимости индивидуальных или временных эффектов может использоваться тест множителей Лагранжа (критерий Бреуша-Пагана). Для проверки гипотезы о корреляции индивидуальных или временных эффектов используется тест Хаусмана.

Набор П.д. называют сбалансированным, если состав объектов полностью сохраняется между раундами, в противном случае – несбалансированным. Если состав исследуемой совокупности быстро обновляется, напр., при исследовании безработицы, то постоянную долю наблюдений исключают и заменяют новыми. Такие панели называют ротационными.

Если панельные обследования отсутствуют, но доступны данные независимых выборочных обследований за несколько периодов применяют так называемые псевдопанельные данные. В этом случае выделяют когорты по некоторым социально-демографическим признакам и в качестве переменных исследуют изменения средних значений признаков. При этом необходимо оптимально выбирать число и размер когорт.

ПАРАМЕТРИЗАЦИЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

часть предварительной обработки исходного массива данных:

$$X = \begin{pmatrix} x_1^{(1)}(t) & x_1^{(2)}(t) & \dots & x_1^{(p)}(t) \\ x_2^{(1)}(t) & x_2^{(2)}(t) & \dots & x_2^{(p)}(t) \\ \dots & \dots & \dots & \dots \\ x_n^{(1)}(t) & x_n^{(2)}(t) & \dots & x_n^{(p)}(t) \end{pmatrix},$$

при $t=t_1, t_2, \dots, t_N$, где $x_i^{(j)}(t_k)$ – значение j анализируемого признака, характеризующего состояние i объекта в момент времени t_k, которая включает в себя вычисление осн. числовых характеристик распределения: *медианы*, *моды*, *математического ожидания* (среднего значения), *дисперсии*, *коэффициентов ассиметрии*, *эксцесса* и *вариации*.

В многомерном случае также определяются элементы выборочной ковариационной матрицы. Результаты анализа этих числовых характеристик могут позволить сформулировать одну или несколько конкурирующих гипотез об общем (параметрическом) виде закона распределения вероятностей, задающего эту *ген. совокупность*. Знание общего вида вероятностного распределения в исследуемой ген. совокупности, позволяет сделать наилучший выбор метода статистического оценивания параметров этого распределения, а также метода последующей статистической обработки массива исходных данных.

ПАРАМЕТРИЗАЦИЯ РЕГРЕССИОННОЙ МОДЕЛИ

один из этапов *регрессионного анализа*, который заключается в выборе параметрического семейства функций (класса допустимых решений) $F = \{f(X;\Theta)\}$, в рамках которого производится дальнейший поиск неизвестной функции регрессии. П.р.м. – одновременно наиболее важный и наименее теоретически обоснованный этап регрессионного анализа.

Цель данного этапа регрессионного анализа – определение общего вида, структуры искомой связи между признаками. Осн. моменты при выборе общего вида функции регрессии: использование априорной информации о сущности зависимости; предварительный анализ геометрической структуры исходных данных; использование различных статистических приёмов обработки исходных данных, позволяющих сделать наилучший выбор. Поиск аппроксимации $\hat{f}(X)$ сводится к наилучшему (с точки зрения заданного критерия адекватности) подбору неизвестного параметра $\hat{\Theta}$, что в свою очередь осуществляется с помощью полностью формализованного алгоритма решения соответствующей оптимизационной задачи (статистическим оцениванием параметров). В качестве класса допустимых решений использьзуются:

линейные функции: $f(X;\Theta) = \theta_0 + \sum_{k=1}^{p} \theta_k \cdot x^{(k)}$;

степенные функции: $f(X;\Theta) = \theta_0 \cdot \left(x^{(1)}\right)^{\nu_1} \cdot \left(x^{(2)}\right)^{\nu_2} \cdot \dots \cdot \left(x^{(p)}\right)^{\nu_p}$;

алгебраические полиномы:

$$f(X;\Theta) = \theta_0 + \sum_{k=1}^{p} \theta_k \cdot x^{(k)} + \sum_{k_1=1}^{p}\sum_{k_2=1}^{p} \theta_{k_1 k_2} \cdot x^{(k_1)} \cdot x^{(k_2)} + \dots + \sum_{k_1=1}^{p} \dots \sum_{k_m=1}^{p} \theta_{k_1 k_2 \dots k_m} \cdot x^{(k_1)} \cdot x^{(k_2)} \cdot \dots \cdot x^{(k_m)}.$$

От выбора общего вида функции регрессии зависит точность восстановления неизвестной функции регрессии. В тоже время не существует системы стандартных рекомендаций и методов, которые образовывали бы строгую теоретическую базу для его наиболее эффективной реализации.

ПРИВЕДЁННАЯ ФОРМА МОДЕЛИ

уравнение, явно разрешённое относительно объясняемой переменной, т.е. уравнение вида: Y = f(X, β), где Y – объясняемая переменная (возможно многомерная), X = ($X^{(1)}$,…, $X^{(k)}$) – *объясняющие переменные* (регрессоры), β – параметры.

Исходно модель задаётся в структурной форме F(X, Y, θ)=0 (*структурная форма модели*), т.е. в форме, явно отражающей структуру (экономического) процесса. Параметры θ называются структурными. Именно их оценки представляют интерес.

Наиболее распространён случай, когда функция F – линейная вектор-функция, а Y=(Y_1,… Y_m). В этом случае структурная форма модели принимает вид: BY+ГX=ε, где

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_m \end{pmatrix},\ X = \begin{pmatrix} X_1 \\ \dots \\ X_l \end{pmatrix},\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_m \end{pmatrix},$$

Y_i, ε_i – n-мерные векторы, X_i – (n×k)-матрица, а B и Г – блочные матрицы соответствующих размерностей. Элементы этих матриц – структурные параметры модели.

Оценки *методом наименьших квадратов* (МНК), полученные по структурной форме, несостоятельны. Приведённая форма *Y*=*X*β + ν, получаемая разрешением уравнений модели в структурной форме относительно Y, есть один из способов получения *оценок состоятельных*. Структурные параметры B и Г следует выразить через параметры β и подставить вместо β их МНК-оценки b. Такая процедура называется косвенным МНК.

Структурный параметр называется идентифицируемым, если он может быть однозначно выражен через параметры приведённой формы. Если таковое выражение отсутствует, параметр называется неидентифицируемым. Его оценку приходится получать непосредственно из структурной формы, напр., *методом макс. правдоподобия*.

См. также *Идентифицируемость*.

ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ

устанавливает количественную связь между результатом (эффектом) некоторого процесса и условиями его получения. Под результатом чаще всего понимается выпуск продукции (фактической или максимально возможной) некоторой производственной единицы (пр-тия, отрасли, региона, народного хоз-ва в целом) в натуральном или денежном выражении, а под условиями – ресурсы (затраченные, использованные или наличные). Концепция П.ф. базируется в первую очередь на идее замещения между факторами, т.е. на гипотезе о том, что один и тот же выпуск может, быть получен при разных комбинациях используемых ресурсов. При этом речь идёт о замещении как между различными ресурсами в рамках одной и той же технологии, так и между различными технологиями произ-ва одного и того же продукта или между различными продуктами, имеющими разную ресурсоёмкость.

П.ф. записывается в виде: $Y = F(x_1, x_2, ..., x_n)$, где Y – объём выпуска, x_i – объём i-го фактора произ-ва. Вид функции F и значения её параметров определяются из теоретических представлений и имеющейся конкретной информации о моделируемом объекте. Оценка параметров производится методами регрессионного анализа, поэтому любая оцененная П.ф. представляет собой уравнение регрессии.

П.ф. народного хоз-ва чаще всего имеет вид: $Y = F(K, L)$ или $Y = F(K, L, t)$, где K и L характеризуют ресурсы (затраты) осн. фондов и живого труда, а t – время, вводимое для описания воздействия прочих факторов (неучтённых в K и L), среди которых важную роль играет научно-технический прогресс.

Первая, наиболее простая П.ф. между макс. объёмом выпуска и комбинацией факторов его создающих при имеющемся уровне знаний и технологий была построена Ч. Коббом и П. Дугласом в 1928 для обрабатывающей пром-сти США и имела вид:

$$Ln(Q) = \ln(1{,}01) + 0{,}73\ln(L) + 0{,}27\ln(K).$$

Затем данная функция была обобщена Р. Солоу и К. Арроу (США) с учётом влияния масштаба

произ-ва, технического прогресса и других факторов.

К наиболее часто используемым П.ф. относятся: *П.ф. Кобба-Дугласа, П.ф. с постоянной эластичностью замещения* (CES-функция), а также *П.ф. леонтьевского типа* и *П.ф. линейная*. В теоретических работах в некоторых случаях используются П.ф., записанные в неявном виде, иногда и многофакторные.

ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ КОББА-ДУГЛАСА

функция, устанавливающая количественную связь объёма произ-ва Y с затратами капитала (K) и затратами труда (L) и имеет вид: $Y = AK^{\alpha}L^{\beta}\varepsilon$. П.ф.К.-Д. наряду с линейной производственной функцией – частный случай *производственных функций* с постоянной эластичностью замещения факторов (CES), когда её эластичности стремятся к единице ($\sigma \to 1$, $\rho \to 0$). Автономная зависимость от времени выражена в коэффициенте научно-технического прогресса A. Показатели α и β – коэффициенты частной эластичности объёма произ-ва Y соответственно по затратам капитала K и труда L. Это означает, что при увеличении затрат капитала (труда) на 1% объём произва увеличивается на α % (β %).

Сумма коэффициентов – важный экономический показатель, название которого – отдача от масштаба. При $\alpha + \beta > 1$ – возрастающая отдача от масштаба (увеличение объёма выпуска больше увеличения затрат ресурсов). При $\alpha + \beta < 1$ имеет место убывающая отдача от масштаба (увеличение объёма выпуска меньше увеличения затрат ресурсов). При $\alpha + \beta = 1$ говорят о постоянной отдаче от масштаба (во сколько раз увеличиваются затраты ресурсов, во столько же раз увеличивается выпуск).

П.ф. К.-Д. также представляют в виде:

$$\frac{Y}{L} = \frac{AK^{\alpha}}{L^{\alpha}}\varepsilon .$$

Т.о., получается зависимость производительности труда (Y/L) от его капиталовооруженности (K/L).

Удовлетворяя большинству теоретических требований, эта функция сочетает в себе простую математическую запись и небольшое количество параметров, численные значения которых могут быть легко оценены. Для оценки параметров данной модели её логарифмируют с целью приведения к линейному виду (для i-го наблюдения):

$$\ln(Y/L)_i = \ln A + \alpha \ln(K/L)_i + \ln \varepsilon_i, i = 1,2,...,n.$$

Параметры находят *методом наименьших квадратов*. В непрерывном времени это равенство является точным, а в дискретном – приближённым. П.ф.К.-Д. – нелинейная модель относительно оцениваемых параметров, т.к. в параметры α и β входят в неё мультипликативно. Однако её можно считать внутренне линейной, т.к. логарифмирование уравнения приводит его к линейному виду.

В общем виде П.ф.К.-Д. имеет вид: $Y = Ax_1^{a_1} \cdot ... \cdot x_n^{a_n}$, где A, $a_1,\ldots,a_n$ – параметры. Частная эластичность выпуска по каждому ресурсу в рамках П.ф.К.-Д. – постоянна и равна соответствующему показателю степени, а эластичность замены между любыми двумя ресурсами равна единице.

ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ ЛЕОНТЬЕВСКОГО ТИПА

производственная функция с постоянными пропорциями потребления факторов, описывающая жесткие производственные процессы. В жестком технологическом процессе невозможна замена одного фактора другими и недостаток одного фактора не может быть компенсирован избытком другого. П.ф.л.т. наряду с *производственной функцией Кобба-Дугласа* и *линейной производственной функцией* – частный случай производственных функций с постоянной эластичностью замещения факторов (CES) и имеет вид:

$$Y = \min\left(\frac{x_1}{a_1} \cdot .... \cdot \frac{x_n}{a_n}\right),$$

где $a_1,\ldots, a_n$ – параметры. Предельная норма замены между двумя любыми ресурсами равна бесконечности, а эластичность замещения – нулю. Т.о., предполагается, что ресурсоёмкость

произ-ва по каждому ресурсу фиксирована и, следовательно, объём выпуска однозначно определяется количеством лимитирующего фактора. Также П.ф.л.т. может быть записана в виде:

$$\lim\left(F(K,L)\right) = \min(K\delta / a, L\delta / b)$$
$$\sigma \Rightarrow 0$$

где К – затраты капитала, L – затраты труда, δ – степень однородности функции. Функция Леонтьева для δ=1 имеет вид: F(K,L) = min(K/a, L/b). Изокванта состоит из прямых параллельных осей координат.

ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ ЛИНЕЙНАЯ

функция, устанавливающая количественную связь объёма произ-ва Y с определяющими его факторами $x_1, x_2, \ldots, x_k$, вида: $Y = a_1x_1 + ... + a_nx_n$, где $a_1, \ldots, a_n$ – параметры. П.ф.л. применяется для гибких производственных систем, характеризующихся возможностями компенсации одних факторов другими и полного замещения факторов крупномасштабных произ-в (все предельные нормы замены постоянны, а эластичности равны бесконечности). П.ф.л. – наиболее простая из *производственных функций*; является предельным случаем *производственной функции с постоянной эластичностью замещения факторов* (CES), когда все эластичности стремятся к бесконечности. Используется на практике редко, т.к. гипотеза о линейности в большинстве случаев не является адекватной. Предельные продукты факторов равны коэффициентам ПФ. Изокванты П.ф.л. пересекают оси координат. Предельная норма замещения труда (L – затраты труда) капиталом (К – затраты капитала) составляет $\Gamma_{LK} = b / a$. Коэффициенты a и b показывают пропорции, в которых один фактор может быть заменён другим. Если, напр., a = b = 1, то этот означает, что один час труда может быть заменён одним часом машинного времени.

ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ С ПОСТОЯННОЙ ЭЛАСТИЧНОСТЬЮ ЗАМЕЩЕНИЯ

однородная *производственная функция* степени δ класса CES (constant elasticity of substitution), которая в случае двух факторов K и L имеет вид:

$$F(K,L) = (a_1K^{-\rho} + a_2L^{-\rho})^{-\delta/\rho} \qquad (1),$$

где a_1 и a_2 – константы; $\rho = (1- \sigma_{LK})/ \sigma_{LK}$, δ – степень однородности функции, σ – эластичность замещения труда капиталом. Уравнение (1) имеет смысл в тех случаях, когда $\sigma_{LK} \neq 1$ и $\sigma_{LK} \neq 0$. Эластичность замещения в рамках этой функции равна

$$\sigma = \frac{1}{1+\rho}.$$

Впервые производственные функции класса CES были введены американскими экономистами Эрроу и Солоу в 1961. П.ф. с п.э.з. (CES) считается наиболее гибкой и теоретически содержательной. *Производственная функция Кобба-Дугласа* и *производственная функция линейная* – её предельные случаи: первая при $\rho \to 0$ (т.е. при $\sigma \to 1$), вторая при $\rho \to -1$ (т.е. при $\sigma \to \infty$). Для отражения влияния научно-технического прогресса используется модификация функции:

$$\text{CES: } Y = \left[a_1(e^{\nu_1 t}x_t)^{-\rho} + a_2(e^{\nu_2 t}x_2)^{-\rho}\right]^{-\delta/\rho},$$

где t – время.

В соответствии с различными определениями эластичности замещения существуют различные обобщения функции (1) на случай n-ресурсов. Наиболее известны среди них функции:

$$Y = \left[a_1x_1^{-\rho} + \ldots + a_nx_n^{-\rho}\right]^{-\delta/\rho}$$

$$Y = \left(\sum_{k\in N_1} a_k x_k^{-\rho_1}\right)^{-\delta_1/\rho_1} \ldots \left(\sum_{k\in N_s} a_k x_k^{-\rho_s}\right)^{-\delta_s/\rho_s},$$

где $N_1, \ldots, N_s$ – непересекающиеся подмножества индексов, в сумме составляющие множество $\{1, \ldots, n\}$. Для первой функции

$$\sigma_{ij} = \frac{1}{1+\rho}$$

при всех i, j, а для второй –

$$\sigma_{ij} = \begin{cases} 1, \ \textit{если } i, j \textit{ принадлежат разным подмножествам из числа } N_1,...,N_s \\ \dfrac{1}{1+\rho_l}, \ \textit{если } i, j \in N_l, \ l = 1,2,...,s. \end{cases}$$

Свойство функций класса (CES): асимптоты, проведённые к изоквантам такой функции параллельны осям координат, но их не касаются. Экономически это означает, для таких производственных систем невозможно полностью заменить труд капиталом. Существуют критические значения затрат факторов, ниже которых произ-во невозможно.

ПРОСТРАНСТВЕННО-ВРЕМЕННАЯ ВЫБОРКА

исходные статистические данные, содержащие сведения об одном и том же множестве объектов за ряд последовательных периодов времени; такие данные также называются *панельными данными* или просто панелью. В качестве панели могут выступать индивидуумы, группы лиц, пртия, домохоз-ва, регионы, страны и т.д., сведения о которых собраны в течение нескольких периодов времени. Этот метод сбора данных используется при изучении потребительского поведения, занятости, безработицы, доходов и заработной платы, производственных функций и политики дивидендов фирм, в междунар. и межрегиональных сопоставлениях. Исходные статистические данные обычно записывают в виде табл. (матриц) «объект-свойство»: по строкам располагаются объекты, по столбцам – признаки в определённый момент времени:

$$\begin{pmatrix} x_1^{(1)}(t) & x_1^{(2)}(t) & \cdots & x_1^{(p)}(t) \\ x_2^{(1)}(t) & x_2^{(2)}(t) & \cdots & x_2^{(p)}(t) \\ \cdots & \cdots & \cdots & \cdots \\ x_n^{(1)}(t) & x_n^{(2)}(t) & \cdots & x_n^{(p)}(t) \end{pmatrix}$$

$t = t_1, t_2, \ldots, t_N,$

где $x_i^{(j)}(t_k)$ – значение j-го анализируемого признака, характеризующего состояние i-го объекта в момент времени t.

Такие данные образуют т.н. П.-в.в., при формировании которой статистическому обследованию подвергаются n объектов (как-то размещённых в пространстве), причём на каждом из объектов регистрируется значение p характеризующих его признаков в N последовательные моменты времени $t_1, t_2, \ldots, t_N$.

Приведённая форма записи исходных данных определяет в действительности целую последовательность (N) матриц «объект-свойство». Для экономических приложений типична ситуация, когда моменты времени $t_1, t_2, \ldots t_N$, в которые производится регистрация значений анализируемых признаков, являются равноотстоящими, т.е. $t_2-t_1=t_3-t_2= \ldots = t_N - t_{N-1}=\Delta t$. В этом случае время удобнее считать и обозначать в числе «тактов» Δt. Соответственно тогда вместо $t_1, t_2, \ldots t_N$ записывают $t=1,2, \ldots, N$. Панельные данные бывают сбалансированными и несбалансированными. Если данные присутствуют по всем объектам за все периоды времени, то панель называется сбалансированной. Достаточно часто из-за технических, организационных или иных причин в некоторые периоды времени не удается собрать сведения для всех объектов, включенных в выборку первоначально (смерть, болезнь, отъезд индивидуума и т.п.). Чтобы сохранить репрезентативность, отсутствующие объекты заменяются другими. Такие данные называют несбалансированной панелью. При исследовании проблем занятости и безработицы в междунар. практике распространены т.н. ротационные панели. Объект (человек трудоспособного возраста) участвует в шести последовательных ежеквартальных опросах, а затем исключается из панели. Т.о., 1/6 часть всей выборки обновляется. Возможны и иные модификации панельных данных. Но наибольшее распространение получили сбалансированные и несбалансированные панели. На практике в большинстве случаев число объектов достаточно велико (несколько десятков, сотен или тыс.), а число моментов наблюдений ограничено. Преимущества П.-в.в.: во-первых, большое число наблюдений обеспечивает большую эффективность оценивания параметров *модели эконометрической*; во-вторых, появляется воз-

можность контроля над неоднородностью объектов; в-третьих, возможность идентифицировать эффекты, недоступные в анализе пространственных данных.

Р

РЕГРЕССИЯ ТИПОЛОГИЧЕСКАЯ

вид статистической модели, при построении которой используется сочетание методов классификации многомерных наблюдений и множественной регрессии. Последовательное применение этих методов обеспечивает выделение однородных классов объектов (наблюдений) и построение в каждом из выделенных кластеров регрессионных зависимостей. Выделение однородных групп объектов достигается за счёт использования методов многомерной классификации (напр., *кластерного анализа*). Р.т. позволяет расширить сферу применения методов статистического исследования зависимостей за счёт преодоления ограничений, связанных с требований однородности исходной совокупности данных. При этом необходимо учитывать, что одна и та же совокупность может быть качественно однородной в одном статистическом исследовании и разнородной в другом. Так, напр., совокупность пр-тий является однородной в случае анализа производительности труда, и неоднородной в случае, если изучается налогообложение предприятий. Процесс построения моделей Р.т. является итерационным: на каждом шаге уточняется классификационная структура совокупности объектов (наблюдений) и параметров внутриклассовых регрессий. Выбор окончательного результата производится на основании критериев, определяемых постановкой задачи (напр., по достижению наилучшей точности прогнозных оценок) и целью исследования. В ряде задач возникает необходимость отнесения объектов не участвующих ранее в классификации к одному из классов. Р.т. находит широкое применение в прикладных задачах статистического анализа.

РЕГРЕССИОННЫЙ АНАЛИЗ

раздел *математической статистики*, объединяющий методы анализа зависимости среднего значения *случайной величины* результативного признака y от переменных x1, x2, …., xn (факторов или регрессоров), которые могут рассматриваться как случайные, так и неслучайные величины, независимо от истинного закона их распределения. В качестве формы зависимости выбирается определённый класс функций. Подобный выбор осуществляется экспертным путём на основе соображений, касающихся изучаемой зависимости (экономических, социальных и т.п.). В случае неизвестной формы зависимости выбирают такую функцию, которая давала бы значения результативного признака, близкие к полученным реализациям случайной величины y при наблюдаемых значениях регрессоров. Кроме того, можно использовать графическое изображение наблюдаемых переменных, а также руководствоваться по возможности простой формой зависимости. Наиболее простыми видами зависимости являются линейные, или приводимые к ним с помощью *линеаризации*.

После выбора класса функций регрессии, отражающих зависимость *математических ожиданий* результативного признака от значений независимых аргументов, задачей Р.а. становится оценка неизвестных параметров. Самый распространённый метод оценки параметров регрессионной модели – *метод наименьших квадратов*, дающий при определённых условиях *оценки несмещённые* с наименьшей дисперсией. После того как модель построена, необходимо проверить её адекватность исходным данным, а также полученную точность. Проблема точности построенной модели наиболее эффективно разрешается при допущении, что вектор наблюдений Y распределён нормально. Условие нормальности используется для построения доверительных интервалов и проверки значимости как отдельных коэффициентов регрессии, так и самого уравнения регрессии при фиксированных значениях аргументов $x_1^0,\ldots,x_n^0$. Р.а. является одним из наиболее распространённых методов обработки данных при изучении зависимостей в различных областях знаний.

РЕЗУЛЬТИРУЮЩАЯ ПЕРЕМЕННАЯ

случайная переменная, которая в экономико-статистических моделях является результатом моделирования, и среднее значение которой формируется в зависимости от значений заранее известных переменных-факторов (объясняющих переменных). Р.п. также называют функцией отклика, объясняемой, выходной, зависимой, эндогенной переменной и результативным признаком. В экономико-статистических моделях Р.п. рассматриваются как случайные величины, средние значения которых моделируются в зависимости от значений объясняющих переменных. В эконометрической модели значения результативных переменных формируются в процессе функционирования анализируемой социально-экономической системы под воздействием заранее известных объясняющих переменных. Напр., при моделировании величины потребительских расходов в зависимости от среднедушевого дохода и пола, расходы респондента, будут выступать в виде Р.п. при объясняющих переменных-факторах – доход и пол.

С

СИСТЕМА ОДНОВРЕМЕННЫХ УРАВНЕНИЙ (СОУ)

набор взаимосвязанных *уравнений регрессии*, в которых одни и те же переменные могут одновременно играть роль, в различных уравнениях системы, и результирующих показателей и объясняющих переменных (предикторов).

В любой *модели эконометрической,* в зависимости от конечных прикладных целей её использования, все участвующие в ней *переменные* подразделяются на: *экзогенные*, т.е. задаваемые как бы «извне», автономно, и, в большинстве своём, в определённой степени управляемые (планируемые); *эндогенные*, т.е. такие переменные, значения которых формируются в процессе и внутри функционирования анализируемой социально-экономической системы в существенной мере под воздействием экзогенных переменных и, конечно, во взаимодействии друг с другом; в эконометрической модели они являются предметом объяснения; предопределённые, т.е. выступающие в системе в роли факторов–аргументов, или объясняющих переменных.

Множество предопределённых переменных формируется из всех экзогенных переменных (которые могут быть «привязаны» к прошлому, текущему или будущим моментам времени) и т.н. лаговых эндогенных переменных, т.е. таких эндогенных переменных, значения которых входят в уравнения анализируемой эконометрической системы измеренными в прошлые (по отношению к текущему) моменты времени, а следовательно, являются уже известными, заданными.

Т.о., можно сказать, что эконометрическая модель служит для объяснения поведения эндогенных переменных в зависимости от значений экзогенных и лаговых эндогенных переменных.

При построении и анализе эконометрической модели следует различать её структурную и приведённую формы. Для пояснения этих понятий будем обозначать латинской буквой X вектор-столбец всех предопределённых переменных (включает в себя свободный член, все экзогенные переменные и все участвующие в модели лаговые эндогенные переменные). Пусть общее число эндогенных переменных равно m, а общее число предопределённых переменных – $p+1$. Общее число уравнений и тождеств в эконометрической модели равно числу эндогенных переменных, т.е. равно m. И пусть из общего числа m соотношений модели имеется m_1 уравнений, включающих случайные остаточные компоненты, и m_2 тождеств $(m_1+m_2=m)$. Разобьём вектор эндогенных переменных $Y_t=(y_t^{(1)},y_t^{(2)},\ldots,y_t^{(m)})^{\mathrm{T}}$ на два подвектора $Y_t^1=(y_t^{(1)},y_t^{(2)},\ldots,y_t^{(m_1)})^{\mathrm{T}}$ и $Y_t^2=(y_t^{(m_1+1)},\ldots,y_t^{(m_1+m_2)})^{\mathrm{T}}$, при этом порядок, в котором перенумерованы эндогенные переменные, не имеет значения.

Тогда общий вид линейной эконометрической модели представлен в форме:

$$\begin{cases} \mathbf{B}_1 Y_t^{(1)} + \mathbf{B}_2 Y_t^{(2)} + \mathbf{C}_1 X_t = \Delta_t \\ \mathbf{B}_3 Y_t^{(1)} + \mathbf{B}_4 Y_t^{(2)} + \mathbf{C}_2 X_t = 0, \quad t = 1,2,\ldots,n, \end{cases} \tag{1}$$

где

$\mathbf{B}_1 = (\beta_{ij})_{i,\,j=\overline{1,m_1}}$

– матрица размерности $(m_1 \times m_1)$ из коэффициентов при $y_t^{(1)},\ldots,y_t^{(m_1)}$ m_1 первых уравнениях;

$\mathbf{B}_2 = (\beta_{ij})_{\substack{i=\overline{1,m_1} \\ j=\overline{m_1+1,m}}}$

– матрица размерности $m_1 \times (m - m_1)$ из коэффициентов при $y_t^{(m_1+1)},\ldots,y_t^{(m)}$ в m_1 первых уравнениях; $X_t = (x_t^{(0)}, x_t^{(1)}, x_t^{(2)},\ldots,x_t^{(p)})^{\mathrm{T}}$ – вектор-столбец предопределённых переменных (в нём $x_t^{(0)} \equiv 1$);

$\mathbf{C}_1 = (c_{ij})_{\substack{i=\overline{1,m_1}, \\ j=0,p}}$

– матрица размерности $m_1 \times (p+1)$ из коэффициентов при предопределённых переменных в первых m_1 уравнениях (очевидно, коэффициенты c_{i0} играют роль свободных членов уравнений);

$\mathbf{B}_3 = (\beta_{ij})_{\substack{i=\overline{m_1+1,m}, \\ j=\overline{1,m_1}}}$

– матрица размерности $(m - m_1)$ из коэффициентов при $y_t^{(1)},\ldots,y_t^{(m_1)}$ в $m_2 = m - m_1$ тождествах системы;

$\mathbf{B}_4 = (\beta_{ij})_{\substack{i=\overline{m_1+1,m}, \\ j=\overline{m_1+1,m}}}$

– матрица размерности $(m - m_1)$ из коэффициентов при $y_t^{(m_1+1)},\ldots,y_t^{(m)}$ в $m_2 = m - m_1$ тождествах системы;

$\mathbf{C}_2 = (c_{ij})_{\substack{i=\overline{m_1+1,m}, \\ j=0,p}}$

– матрица размерности $m_2 \times (p+1)$ из коэффициентов при предопределённых переменных в m_2 тождествах системы;

$\Delta_t = (\varepsilon_t^{(1)}, \varepsilon_t^{(2)},\ldots,\varepsilon_t^{(m_1)})^{\mathrm{T}}$ – вектор-столбец размерности m_1 случайных остаточных составляющих m_1 первых уравнений системы и $\mathbf{O}_{m_2} = (0,0,\ldots,0)^{\mathrm{T}}$ – вектор-столбец размерности m_2, состоящий из нулей.

Исходными статистическими данными, необходимыми для проведения статистического анализа системы (1) (а именно, для оценки неизвестных коэффициентов β_{ij} и c_{ij}, проверки статистических гипотез, напр., о линейном характере исследуемых зависимостей и т. п.), являются матрицы:

$$Y = \begin{pmatrix} Y_i^{\mathrm{T}} \\ \vdots \\ Y_n^{\mathrm{T}} \end{pmatrix} \quad \text{и} \quad X = \begin{pmatrix} Y_i^{\mathrm{T}} \\ \vdots \\ Y_n^{\mathrm{T}} \end{pmatrix} \tag{2}$$

соответственно размерностей $n \times m$ и $n \times (p+1)$, а все элементы матриц B_3, B_4 и C_2 – известны: их числовые значения определяются содержательным смыслом соответствующих тождеств системы.

Система (1) может быть записана также в виде:

$$\mathbf{B} Y_t + \mathbf{C} X_t = \bar{\Delta}_t, \quad t = 1,2,\ldots,n, \tag{1'}$$

или в виде:

$$Y \cdot \mathbf{B}^{\mathrm{T}} + X \cdot \mathbf{C}^{\mathrm{T}} = \bar{\Delta}, \tag{1''}$$

Где

$$Y_t = \begin{pmatrix} Y_t^{(1)} \\ Y_t^{(2)} \end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{B}_1 & \mathbf{B}_2 \\ \mathbf{B}_3 & \mathbf{B}_4 \end{pmatrix}, \mathbf{C} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{pmatrix}, \bar{\Delta}_t = \begin{pmatrix} \Delta_t \\ O_{m_2} \end{pmatrix}, \bar{\Delta} = \begin{pmatrix} \bar{\Delta}_1^{\mathrm{T}} \\ \vdots \\ \bar{\Delta}_n^{\mathrm{T}} \end{pmatrix},$$

а матрицы Y и X определены в (2).

Система уравнений и тождеств вида (1) (или эквивалентных ей записей (1') или (1")) называется структурной формой линейной эконометрической модели. При этом предполагается, что коэффициент при i-й эндогенной переменной в 1-м структурном стохастическом уравнении $(i = 1,2,\ldots,m)$ равен единице (правило

нормировки системы), а матрицы B_4 и B невырождены (допускаются и другие способы нормировки системы).

Поскольку при реализации конечных прикладных целей эконометрического моделирования (т.е. при прогнозе значений эндогенных переменных и при различных имитационных расчётах) главный интерес представляют соотношения, позволяющие явно выразить все эндогенные переменные Y_t, через предопределённые X_t, то одновременно со структурной формой имеет смысл рассмотреть т.н. приведённую (редуцированную) форму линейной эконометрической модели. Требуемый результат получим, домножив слева обе части соотношений (1') на матрицу B^{-1} и уединив затем Y_i:

$$Y_t = -\mathbf{B}^{-1}\mathbf{C}X_t + \mathbf{B}^{-1}\overline{\Delta}_t, \quad t = 1,2,\ldots,n, \qquad (3)$$

или

$$Y_t = \mathbf{\Pi} \cdot X_t + \tilde{\varepsilon}_t, \; t = 1,2,\ldots,n, \qquad (3')$$

где $m \times (p+1)$ матрица $\mathbf{\Pi}$ и вектор остаточных случайных составляющих $\tilde{\varepsilon}_t$, определяются соотношениями:

$$\mathbf{\Pi} = -\mathbf{B}^{-1}\mathbf{C}, \qquad (4)$$

$$\tilde{\varepsilon}_t = \mathbf{B}^{-1}\overline{\Delta}_t. \qquad (5)$$

Система соотношений (3'), в которой все эндогенные переменные эконометрической модели явно линейно выражены через предопределённые переменные и случайные остаточные компоненты, называется приведённой формой линейной эконометрической модели.

Проблема спецификации модели заключается в определении: конечных целей моделирования (прогноз, имитация различных сценариев социально-экономического развития анализируемой системы, оценка определённых экономических характеристик); списка экзогенных и эндогенных переменных; состава анализируемой системы уравнений и тождеств, их структуры и соответственно списка предопределённых переменных; способа параметризации модели, т.е. нахождения общего вида искомых функциональных зависимостей, связывающих между собой анализируемые переменные; формулировки исходных предпосылок и априорных ограничений относительно: стохастической природы остатков Δ_t (в классических вариантах моделей постулируются их взаимная статистическая независимость или некоррелированность, нулевые значения их средних величин и, иногда, сохранение постоянными в процессе наблюдения значений их дисперсий – *гомоскедастичность*), числовых значений отдельных параметров модели.

Спецификация модели – первый и важнейший шаг эконометрического исследования. От того, насколько удачно решена проблема спецификации и, в частности, насколько реалистичны наши решения и предположения относительно состава эндогенных, экзогенных и предопределённых переменных, структуры и общего вида самой системы уравнений и тождеств, стохастической природы случайных остатков и конкретных числовых значений части неизвестных параметров модели, решающим образом зависит успех всего эконометрического исследования. Спецификация опирается как на имеющиеся положения экономической теории, специальные знания или интуитивные представления исследователя об анализируемой экономической системе, так и на специальные методы и приёмы, в т.ч., математико-статистические, т.н. разведочного анализа.

СИТУАЦИОННЫЙ АНАЛИЗ

в эконометрическом моделировании метод исследования зависимостей средних значений эндогенных переменных при различных вариантах (сценариях, ситуациях) значений предопределенных переменных. Другое название метода – сценарный анализ. Одна из прикладных целей эконометрического моделирования заключается в получении условного прогноза эндогенных переменных при определённых условиях, накладываемых на значения предопределённых. Поскольку значения предопределенных переменных заранее известны исследователю или он ими может управлять, то получив эконометрическую модель, представляющую собой, напр., систему одновременных эконометрических уравнений, в неё могут быть подставлены различные варианты значений факторов. Т.о., производятся многовариантные сценарные расчёты, показывающие, как будут

«себя вести» эндогенные переменные при различных условиях, касающихся значений предопределенных переменных. С.а. применяется на макро-уровне эконометрического моделирования с целью оптимального регулирования параметров функционирования анализируемой экономической системы. В этом случае речь идёт об оптимальном регулировании тех макропараметров национальной экономики, которые поддаются хотя бы частичном управлению и планированию (институциональные и структурные преобразования, налоговая и социальная политика, инвестиционная активность государства и т.п.). Построив и оценив статистические связи, существующие между этими переменными, можно отслеживать соответствующие реакции эндогенных переменных. Т.е. происходит как бы многократная модельная «прогонка» различных сценариев социально-экономического развития. Поэтому такой способ исследования и называются ситуационным или сценарным анализом. Реализуется этот подход в эконометрике, как правило, с помощью систем одновременных уравнений. Менее освоенным (но не менее правомерным и актуальным) является такой подход в задачах оптимального регулирования.

СТАТИСТИЧЕСКАЯ ЗАВИСИМОСТЬ (ВЕРОЯТНОСТНАЯ, СТОХАСТИЧЕСКАЯ)

зависимость между случайными величинами, которая выражается в изменении условных распределений любой из величин при изменении значений других величин. Виды С.з. многообразны: если случайные величины не являются взаимно независимыми, то им в той или иной степени свойственна случайная зависимость. Один из наиболее общих типов С.з. – корреляционная зависимость. При сопоставлении более чем двух случайных величин их называют взаимно независимыми. Законы их распределения не зависят от того, какие возможные значения приняли остальные случайные величины. Из статистической независимости следует некоррелированность случайных величин. Обратное утверждение не всегда имеет место. Для нормально распределённых случайных величин С.з. и коррелируемость эквивалентны. Предельный случай С.з. – функциональная зависимость, при которой каждой величине одного признака соответствует единственное значение другого признака. В общем случае С.з. каждому фиксированному значению одного признака соответствует не одно, а множество значений другого признака со своим законом распределения. При этом заранее нельзя сказать, какое именно значение примет второй признак. Среднее значение этих величин соответствует регрессии одной случайной величины на другую. См. также *Регрессионный анализ*.

СТАТИСТИЧЕСКАЯ НЕЗАВИСИМОСТЬ

две случайные величины называются независимыми, если закон распределения одной из них не зависит от того, какие возможные значения приняла другая величина. Напр., если дискретная случайная величина X может принимать значения x_i (i=1, 2, … , n), а случайная величина Y – значения y_j (j=1, 2, … , m), то независимость дискретных случайных величин X и Y означает независимость событий $X=x_i$ и $Y=y_j$ при любых i=1, 2, … , n и j=1, 2, … , m. Понятие о независимости случайных величин – одно из ключевых понятий теории вероятностей. Общее определение С.н.с.в. (и для дискретных, и непрерывных) можно дать в терминах функций распределений. Случайные величины X и Y называются независимыми, если их совместная функция распределения $F(x,y)$ представляется в виде произведения функций распределений $F_1(x)$ и $F_2(y)$, т.е. $F(x,y)= F_1(x) \cdot F_2(y)$. В противном случае, при невыполнении данного равенства, случайные величины называют зависимыми. Для независимых непрерывных случайных величин X и Y их совместная плотность $f(x,y)$ равна произведению плотностей вероятности $f_1(x)$ и $f_2(y)$ этих случайных величин, т.е. $f(x,y)= f_1(x) \cdot f_2(y)$. Условие независимости Y от X может быть также записано в виде: $f(y|x)=f_1(y)$ при любом y, и $f(x|y)=f_2(x)$ при любом x, т.е. условные плотности вероятности

каждой из величин совпадают с соответствующими безусловными плотностями. Зависимость или независимость случайных величин всегда взаимны: если величина *Y* не зависит от .*X*, то и величина *X* не зависит от *Y*.

СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ

раздел *математической статистики*, посвящённый математическим методам, направленным на выявление характера и структуры взаимосвязей между компонентами исследуемых явлений и процессов. Исследование зависимостей – одна из гл. целей прикладного статистического анализа. Различают объясняющие (входные, независимые, экзогенные) показатели $x_1, x_2, \ldots, x_p$, описывающие условия функционирования исследуемого явления, и результирующие (выходные, зависимые, эндогенные) $y_1, y_2, \ldots, y_k$, характеризующие результат функционирования системы. В этом случае перед исследователем возникает общая задача С.и.з., заключающаяся в том, чтобы на основе *n* измерений $\{x_i^{(1)},\ldots,x_i^{(p)};y_i^{(1)},\ldots,y_i^{(m)}\}, i=1,\ldots,n$ построить вектор

$$f(x^{(1)},\ldots,x^{(p)})=\begin{pmatrix} f^{(1)}(x^{(1)},\ldots,x^{(p)}) \\ \cdots \\ f^{(m)}(x^{(1)},\ldots,x^{(p)}) \end{pmatrix},$$

позволяющий наилучшим образом восстанавливать значения *Y* по заданным *X*.

Осн. задачи С.и.з.: а) установление самого факта наличия или отсутствия статистически значимой зависимости между результирующими и объясняющими переменными; б) установление формы зависимости (математического выражения) между исследуемыми показателями; в) выявление причинных связей между объясняющими переменными и результирующими показателями, частичное управление результативных переменных путём регулирования величин объясняющих признаков; г) прогноз (восстановление) неизвестных значений индивидуальных или средних значений исследуемых результирующих показателей по заданным значениям соответствующих объясняющих переменных.

На практике математическим аппаратом С.и.з. служат методы многомерного статистического анализа. К ним традиционно относят: *корреляционный*, *регрессионный* и *дисперсионный анализ*, а также методы снижения размерности – компонентный и факторный анализ.

СТРУКТУРНАЯ ФОРМА МОДЕЛИ

система одновременных уравнений (СОУ), в которой одни и те же зависимые переменные в одних уравнениях входят в левую часть, а в других уравнениях – в правую часть системы. С.ф.м. позволяет увидеть влияние изменений любой экзогенной переменной на значения эндогенной переменной. Общий вид С.ф.м.:

$$\begin{cases} y_1 = b_{12}\cdot y_2 + b_{13}\cdot y_3 + \ldots + b_{1n}\cdot y_n + a_{11}\cdot x_1 + a_{12}\cdot x_2 + \ldots + a_{1m}\cdot x_m + \varepsilon_1 \\ y_2 = b_{21}\cdot y_1 + b_{23}\cdot y_3 + \ldots + b_{2n}\cdot y_n + a_{21}\cdot x_1 + a_{22}\cdot x_2 + \ldots + a_{2m}\cdot x_m + \varepsilon_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = b_{n1}\cdot y_1 + b_{n2}\cdot y_2 + \ldots + b_{nn-1}\cdot y_{n-1} + a_{n1}\cdot x_1 + a_{n2}\cdot x_2 + \ldots + a_{nm}\cdot x_m + \varepsilon_n \end{cases}$$

где y_i – эндогенная переменная; x_i – экзогенная переменная; b_i и a_j – структурные коэффициенты модели при эндогенной и экзогенной переменными соответственно.

Все переменные в модели центрированы, выражены в отклонениях от среднего уровня, т.е. под x подразумевается $(x' - \overline{x})$, под y – соответственно $(y' - \overline{y})$, где x' и y' – наблюдаемые значения. Поэтому свободный член в каждом уравнении отсутствует.

Матричный вид системы эконометрических уравнений: BY+ΓX=E, где B – матрица коэффициентов при зависимых переменных; Y – вектор зависимых переменных; Γ – матрица параметров при объясняющих переменных; X –

вектор объясняющих переменных; E – вектор ошибок. Для модели вида:

$$\begin{cases} y_1 = a_{01} + b_{12} \cdot y_2 + a_{11} \cdot x_1 + a_{12} \cdot x_2 + \varepsilon_1 \\ y_2 = a_{02} + b_{21} \cdot y_1 + b_{23} \cdot y_3 + a_{23} \cdot x_3 + \varepsilon_2 \\ y_3 = a_{03} + b_{31} \cdot y_1 + a_{32} \cdot x_2 + a_{33} \cdot x_3 + \varepsilon_3 \end{cases}$$

матрица коэффициентов при зависимых переменных имеет вид:

$$B = \begin{bmatrix} 1 & -b_{12} & 0 \\ -b_{21} & 1 & -b_{23} \\ -b_{31} & 0 & 1 \end{bmatrix}$$

В СОУ отдельно взятое уравнение не может рассматриваться самостоятельно. Для нахождения параметров уравнения традиционный *метод наименьших квадратов* (МНК) неприменим. Наибольшее распространение получили следующие методы оценивания коэффициентов С.ф.м.: косвенный метод наименьших квадратов (КМНК); *двухшаговый метод наименьших квадратов* (2МНК); *трёхшаговый метод наименьших квадратов* (3МНК); *метод макс. правдоподобия* (ММП) с полной информацией ($ММП_f$); метод макс. правдоподобия при ограниченной информации ($ММП_s$).

КМНК применяется для идентифицируемой СОУ (все структурные коэффициенты модели определяются однозначно, единственным образом по коэффициентам приведённой формы модели, т.е. число параметров С.ф.м. равно числу параметров приведённой формы модели). 2МНК используется для оценки коэффициентов сверхидентифицируемой модели (число приведённых коэффициентов больше числа структурных коэффициентов, возможно получение двух и более значений одного структурного коэффициента). ТМНК – используется для всех видов уравнений С.ф.м. ММП – наиболее общий метод оценивания, его результаты при нормальном распределении признаков совпадают с МНК. Однако он трудоёмок в вычислениях при большом числе уравнений в системе, поэтому используется его модификация – $ММП_s$.

Наиболее широко СОУ применяются для построения макроэкономических моделей функционирования экономики страны.

Т

ТРЁХШАГОВЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (3МНК)

оценки параметров *системы одновременных уравнений* (СОУ), согласно которым первоначально с целью оценки параметров каждого структурного уравнения, применяют *двухшаговый метод наименьших квадратов* (2МНК), а затем определяют оценку для ковариационной матрицы случайных ошибок. После этого, с целью оценивания коэффициентов всей системы, применяется *обобщённый метод наименьших квадратов* (ОМНК). Рассмотрим СОУ, содержащую G *эндогенных* и K *экзогенных переменных*, принимаемых как неслучайные. Преобразуем i-е уравнение к виду:

$y_i = Z_i \delta_i + \varepsilon_i$, где, $Z_i = (Y_i X_i)$, $\delta_i = \begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix}$.

Умножив левую и правую части уравнения слева, на транспонированную матрицу X^T, значений всех экзогенных переменных модели, получим: $X^T y_i = X^T Z_i \delta_i + X^T \varepsilon_i$.

Записав т.о. все уравнения системы, получим:

$$\begin{pmatrix} X^T y_1 \\ \vdots \\ X^T y_i \\ \vdots \\ X^T y_G \end{pmatrix} = \begin{pmatrix} X^T Z_1 & 0 & \cdots & \cdots & 0 \\ 0 & X^T Z_2 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & X^T Z_G \end{pmatrix} \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_i \\ \vdots \\ \delta_G \end{pmatrix} + \begin{pmatrix} X^T \varepsilon_1 \\ \vdots \\ X^T \varepsilon_i \\ \vdots \\ X^T \varepsilon_G \end{pmatrix} .$$

Для применения ОМНК, построим ковариационную матрицу вектора возмущений:

$$\Sigma_{(U)} = \begin{pmatrix} \sigma_{11} X^T X & \cdots & \sigma_{1i} X^T X & \cdots & \sigma_{1G} X^T X \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{i1} X^T X & \cdots & \sigma_{ii} X^T X & \cdots & \sigma_{2G} X^T X \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{G1} X^T X & \cdots & \sigma_{Gi} X^T X & \cdots & \sigma_{GG} X^T X \end{pmatrix} = \Sigma \otimes X^T X .$$

Заменив матрицу $\Sigma = (\sigma_{ij})$ её оценкой $S = (s_{ij})$, получим оценку ковариационной матрицы вектора возмущений – $S_{(U)} = S \otimes X^T X$ и соответствующую обратную матрицу – $S^{-1}_{(U)} = S^{-1} \otimes (X^T X)^{-1}$.

Тогда, искомая оценка ТМНК, имеет вид: $\hat{\delta} = (A^T S^{-1}_{(U)} A)^{-1} A^T S^{-1}_{(U)} Z$, где,

$$A = \begin{pmatrix} X^T Z_1 & 0 & \cdots & 0 \\ 0 & X^T Z_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & X^T Z_G \end{pmatrix} ;$$

$$Z = \begin{pmatrix} X^T y_1 \\ \vdots \\ X^T y_i \\ \vdots \\ X^T y_G \end{pmatrix} .$$

В случае, когда матрица Σ не является диагональной, т.е. когда возмущения, входящие в различные структурные уравнения, зависимы, трёхшаговая процедура имеет лучшую асимптотическую эффективность по сравнению с двухшаговой.

У

УНИФИКАЦИЯ ТИПОВ ПЕРЕМЕННЫХ

задача унификации записи единичного многомерного наблюдения Xi, снятого с объекта i, где i = 1,2,…, n и Xi = (xi1, xi2,…, xik)T, когда среди компонент Xi имеются количественные, порядковые и номинальные признаки.

Наличие среди компонент x1, x2,… xk многомерного признака X показателей разных типов приводит к трудности автоматизированного анализа информации. В соответствии с одним из вариантов решения этой задачи i-е многомерное наблюдение в унифицированной форме записи представляется вектором-столбцом размерности m1+m2+…+mk, где mj – число градаций (интервалов группирования, однородных групп) признака xi, причём компонентами этого вектора-столбца могут быть только нули и единицы.

При таком подходе к достижению единообразия записи наблюдений многомерного признака смешанной природы, приходится мириться с субъективизмом выбора способов разбиения диапазонов варьирования анализируемых количественных признаков, а также с потерей информации при переходе от индивидуальных к группированным значениям признака.

Другой подход к унификации записи исходных данных основан на идее прямо противоположной рассмотренной выше. В частности, руководствуясь некоторыми дополнительными допущениями, исследователь пытается преобразовать качественные (порядковые) и классификационные (номинальные) признаки в количественные, используя процесс т.н. «оцифровки» или шкалирования неколичественных переменных. *Многомерное шкалирование* включает в себя методы обработки табл. неколичественных

данных, которая первоначально преобразуется в матрицу расстояний (или близости) между n объектами. Цель методов многомерного шкалирования состоит в том, чтобы на основании информации, содержащейся в матрице расстояний D={die}, где die=d(Oi,Oe) – расстояние между объектами Oi и Oe, получить данные о координатах n точек в многомерном геометрическом пространстве размерности m, т.е. восстановить векторы наблюдений X1′, X2′,…, Xn′. В случае многомерного шкалирования предполагается, что элементы матрицы D – измеренные с некоторой ошибкой, расстояния между объектами совокупности, которые рассматриваются как точки в некотором m-мерном (m<k) пространстве.

УРАВНЕНИЕ РЕГРЕССИИ (ФУНКЦИЯ РЕГРЕССИИ)

уравнение, описывающее зависимость условного среднего значения результативной переменной y от заданных объясняющих переменных $X=(x_1,x_2,\ldots,x_k)^T$:

$$\tilde{y}=M(y/X)=f(x_1,x_2,\ldots,x_k)$$

В этом соотношении объясняющие переменные X могут быть как случайными, так и неслучайными величинами, от значений которых зависит закон распределения вероятностей случайной результативной переменной y.

Если X – k-мерный случайный вектор, то в системе многомерной, (k+1) - мерной случайной величины (y, X) У.р. интерпретируется как *математическое ожидание* y, полученное по условному распределению y при заданном значении X* вектора X.

Если же $X=(x_1,x_2,\ldots,x_k)^T$ ряд неслучайных величин, от значений которых зависит одномерный закон распределения вероятностей результативной величины *y*, то У.р. интерпретируется как математическое ожидание y, полученное при значениях объясняющих переменных равных X^*.

В регрессионном анализе результативная переменная y выступает в роли функции, значения которой с точностью до случайной составляющей зависит от значений объясняющих переменных $X=(x_1,x_2,\ldots,x_k)^T$. Эту зависимость можно представить в виде:

$$y=f(X)+\varepsilon(X).$$

Случайная составляющая, регрессионные остатки $\varepsilon(X)$ характеризуют с одной стороны влияние на y не входящих в X факторов, а с другой – случайную погрешность в измерении значения результативного показателя y. При этом должно выполняться условие $M\varepsilon(X)=0$, т.к. Функция регрессии $f(X)$ есть условное математическое ожидание y при заданном X, т.е.: $f(X)=M(y/X)$.

Выбор метода статистического анализа модели зависит от требований к виду функции $f(X)$, природы объясняющих переменных X и вероятностной природы регрессионных остатков $\varepsilon(X)$.

В регрессионном анализе, как в любом статистическом исследовании вывод строится на основании имеющихся исходных статистических данных типа «объект-свойство», когда i-е наблюдение (i = 1,2,…,n) можно представить в виде строки $(y_i,x_{i1},x_{i2},\ldots,x_{ij},\ldots,x_{ik})$, где n – объём выборочной совокупности (выборки).

Т.о., исходные статистические данные содержат информацию об n наблюдениях и могут быть представлены в виде вектора значений результативной переменной $Y=(y_1,\ldots,y_i,\ldots,y_n)^T$ и матрицы X значений k объясняющих переменных.

С целью наилучшего восстановления по исходным статистическим данным условного среднего значения результативного показателя $y(X)$ и неизвестной функции регрессии $f(X)=M(y/X)$ наиболее часто используют следующие критерии адекватности (функции потерь): 1. *метод наименьших квадратов* (МНК), согласно которому минимизируется квадрат отклонения наблюдаемых значений результативного показателя $y_i(i=1,2,\ldots,n)$ от модельных значений $\tilde{y}_i=f(X_i,\beta)$, где $\beta=(\beta_0,\beta_1,\ldots,\beta_k)^T$ – коэффициенты У.р., Xi-вектор значений аргументов в i-м наблюдении $X_i=(x_{i1},x_{i2},\ldots,x_{ij},\ldots,x_{ik})^T$:

$$\sum_{i=1}^{n}(y_i-f(x_i,\beta))^2\to\min_{\beta}$$

Решается задача отыскания оценки $b=(b_0,b_1,...,b_k)^T$ вектора β. Получаемая регрессия называется среднеквадратической; 2. метод наименьших модулей, согласно которому минимизируется сумма абсолютных отклонений наблюдаемых значений результативного показателя от модельных значений $\tilde{y}_i = f(X_i,\beta)$, т.е.:

$$\sum_{i=1}^{n}|y_i - f(X_i,\beta)| \to \min_{\beta}.$$

Получаемая регрессия называется среднеабсолютной (медианной); 3. метод минимакса сводится к минимизации максимума модуля отклонения наблюдаемого значения результативного показателя y_i от модельного значения $f(X_i,\beta)$, т.е.:

$$\max_{1\le i\le n}|y_i - f(X_i,\beta)| \to \min_{\beta}.$$

Получаемая при этом регрессия называется минимаксной.

При этом необходимо решить осн. задачи: 1) определить наилучшие в определённом смысле точечные и интервальные оценки неизвестной функции регрессии $f(X)$ и её параметров, дать им содержательную экономическую интерпретацию; 2) построить точечный и интервальный прогноз для неизвестного значения результативной переменной $y(X)$ по заданным значениям X; 3) оценить удельный вес влияния каждой из объясняющих переменных $x_1,x_2,...,x_k$ на результативный показатель $y(X)$ и определить какие из объясняющих переменных можно исключить из модели, как практически не влияющие на процесс формирования $y(X)$.

После отбора объясняющих переменных $x_1,x_2,...,x_k$ для регрессионной модели результативного показателя у и сбора статистической информации ключевой становится задача выбора параметрического семейства функций $f(X,\beta)$, в рамках которого предполагается вести поиск наилучшей в определённом смысле оценки $\hat{f}(X;\hat{\beta})$ для $f(X,\beta)$, где $\beta=(\beta_1,\beta_2,...,\beta_k)^T$ – вектор неизвестных параметров модели.

Выбор формы зависимости может осуществляться на основании содержательного анализа исследуемого явления, а также по результатам анализа взаимосвязи переменных, входящих в модель.

По форме (структуре) виды уравнений регрессии, используемые в регрессионном анализе, можно условно разделить на классы: 1) линейные:

$$f(X)=\beta_0+\sum_{j=1}^{k}\beta_j x_j;$$

4) линейные по объясняющим переменным:

$$f(X)=\beta_0+\sum_{j=1}^{k}f_j(\beta_1,\beta_2,...,\beta_m)x_j,$$

напр.,

$$f(X)=\beta_0+\beta_1x_1+\beta_1^2x_2+\beta_2x_3+\beta_1\beta_4x_4;$$

3) линейные по параметрам:

$$f(X)=\beta_0+\sum_{j=1}^{k}\beta_j f(x_1,x_2,...,x_k),$$

напр., полиномиальное уравнение:

$$f(X)=\beta_0+\beta_1x+\beta_2x^2+...+\beta_kx^k.$$

Путём подстановки $x^j=x_j$ полиномиальное уравнение преобразуется в линейное;

4) нелинейные:

$$f(X)=\gamma(x_1,x_2,...,x_p,\beta_0,\beta_1,...,\beta_k),$$

напр., степенное:

$$f(X)=\beta_0x_1^{\beta_1}\cdot x_2^{\beta_2}\cdot...\cdot x_\beta^{\beta_p}.$$

Путём логарифмирования и замены переменных степенное уравнение может быть преобразовано в линейное.

Ф

ФИКТИВНАЯ ПЕРЕМЕННАЯ

искусственная переменная, используемая в *регрессионном анализе* для описания качественных или трудно квантифицируемых характеристик. Ф.п. является такой же равноправной переменной, как и любой из регрессоров $x_j,\ j=1,2,...,k$. Её фиктивность состоит только в том, что она количественным образом описывает качественный признак. Ф.п., как правило, принимает значения 0 или 1, так как в этом случае наиболее просто интерпретируется.

Напр., если регрессионная модель включает в себя два периода, причём в первом периоде процессы протекали при одних условиях, а во втором периоде – при других (напр., до ввода новых производственных мощностей и после ввода новых производственных мощностей или до проведения реформ и после проведения реформ). Тогда d=0, если t<k; d=1, если $t \geq k$, где k соответствует моменту смены условий. *Уравнение регрессии* записывается в виде: $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \gamma d + \varepsilon.$ После оценивания уравнения по знаку коэффициента регрессии при Ф.п. можно судить о качестве производимых изменений. Если d=0, то никаких изменений не произошло.

Если качественный признак имеет не одно, а несколько значений, целесообразно использовать несколько бинарных переменных. Типичный пример подобной ситуации – исследование *сезонных колебаний*, напр., квартальных. Уравнение регрессии в данном случае имеет вид: $y_t = \beta_0 + \beta_1 d_{t1} + \beta_2 d_{t2} + \beta_3 x_{t3} + \varepsilon_t$. При этом необходимо учитывать, что сумма бинарных переменных не должна равняться единице.

Ч

ЧАСТНАЯ АВТОКОРРЕЛЯЦИОННАЯ ФУНКЦИЯ

служит для измерения *автокорреляции*, существующей между разделенными τ тактами времени членами *временного ряда* x_t и $x_{t+\tau}$, при устраненном опосредованном влиянии на эту взаимосвязь всех промежуточных (т.е. расположенных между x_t и $x_{t+\tau}$ членов этого временного ряда).

Коэффициент частной автокорреляции первого порядка $\rho_{частн.}(2)$ при $\tau = 2$ будет определять корреляцию между уровнями временного ряда, разделенными двумя тактами времени, при условии, что значения промежуточных уровней зафиксированы на среднем уровне

$$\rho_{частн.}(2) = \rho(x_t, x_{t+2} \mid x_{t+1} = \mu).$$

Очевидно, что коэффициент частной автокорреляции $\rho_{частн.}(1)$ для лага $\tau = 1$ будет равен коэффициенту автокорреляции $\rho(1)$, так как при этом значении τ нет промежуточных лагов. Но при $\tau > 1$ уже появятся отличия в этих коэффициентах.

Выборочная оценка Ч.а.ф. при $\tau = 2$ определяется формулой:

$$r_{частн.}(2) = r(x_t, x_{t+2} \mid x_{t+1} = \bar{x}) = \frac{r(x_t, x_{t+2}) - r(x_t, x_{t+1}) r(x_{t+2}, x_{t+1})}{\sqrt{(1 - r^2(x_t, x_{t+1}))(1 - r^2(x_{t+2}, x_{t+1}))}} = \frac{r(2) - r^2(1)}{1 - r^2(1)},$$

где $r(x_t, x_{t+i})$ – *автокорреляционная функция*.

Частные автокорреляции более высоких порядков могут быть подсчитаны аналогичным образом. Напр., оценка Ч.а.ф. второго порядка при $\tau = 3$ может быть определена по формуле:

$$r_{частн.}(3) = r(x_t, x_{t+3} \mid x_{t+1} = x_{t+2} = \bar{x}) = \frac{r_{03/1} - r_{02/1} r_{32/1}}{\sqrt{(1 - r^2_{02/1})(1 - r^2_{32/1})}},$$

где

$$r_{03/1} = r(x_t, x_{t+3} \mid x_{t+1} = \bar{x}) = \frac{r(3) - r(1) r(2)}{\sqrt{(1 - r^2(1))(1 - r^2(2))}},$$

$$r_{02/1} = r(x_t, x_{t+2} \mid x_{t+1} = \bar{x}) = \frac{r(2) - r^2(1)}{1 - r^2(1)},$$

$$r_{32/1} = r(x_{t+3}, x_{t+2} \mid x_{t+1} = \bar{x}) = \frac{r(1) - r(2) r(1)}{\sqrt{(1 - r^2(2))(1 - r^2(1))}}$$

Полученные автокорреляции $r_{частн.}(1), r_{частн.}(2), r_{частн.}(3), \ldots$ можно нанести на график, в котором роль абсциссы выполняет величина сдвига τ. Значение автокорреляци-

онной функции $r(\tau)$ и Ч.а.ф. $r_{частн.}(\tau)$ оказывают существенную помощь в решении задач подбора и идентификации модели временного ряда в анализе временных рядов.

Э

ЭКЗОГЕННЫЕ ПЕРЕМЕННЫЕ

переменные *модели эконометрической*, задаваемые как бы извне, автономно, в определённой степени управляемые (планируемые).

В эконометрике в моделях регрессии *системы одновременных уравнений* (СОУ) Э.п. (в отличие от *эндогенных переменных*) относятся к объясняющим переменным и могут быть «привязаны» к прошлым, текущим и будущим моментам времени. В этих моделях в роли факторов – аргументов или объясняющих переменных выступают предопределённые переменные, которые формируются из всех экзогенных и лаговых эндогенных переменных.

В эконометрической модели СОУ поведение эндогенных переменных определяется значениями предопределённых переменных.

ЭКОНОМЕТРИКА

(от экономики и греч. metréō – измеряю) – научная дисциплина, позволяющая на базе положений экономической теории и результатов экономических измерений, с помощью подходящих математических методов, придавать конкретное количественное выражение общим (качественным) закономерностям, обусловленным экономической теорией. При этом осн. роль в математическом оснащении этой дисциплины играют методы *математической статистики,* и в первую очередь, методы многомерного статистического анализа.

Т.о., суть Э. – именно в синтезе экономической теории, экономической статистики и прикладного математического инструментария. Говоря об экономической теории в рамках Э., будем интересоваться не просто выявлением объективно существующих (на качественном уровне) экономических законов и связей между экономическими показателями, но и подходами к их формализации, включающими в себя методы спецификации и идентификации соответствующих моделей с учётом решения проблемы их идентифицируемости (см. схему 1.). При рассмотрении экономической статистики как составной части Э. наиболее интересен тот аспект этой самостоятельной дисциплины, который непосредственно связан с информационным обеспечением анализируемой эконометрической модели, хотя в этих рамках специалисту по Э. зачастую приходится решать полный спектр соответствующих задач: выбор необходимых экономических показателей и обоснование способа их измерения, определение плана статистического обследования и т. п. Наконец, прикладной математической инструментарий Э. в качестве своей осн. составляющей содержит ряд специальных разделов многомерного статистического анализа: линейные (классическая и обобщённая) и некоторые специальные модели регрессии, методы и модели анализа временных рядов, обобщённый метод моментов, т.н. *системы одновременных уравнений* (СОУ), статистические методы классификации и снижения размерности анализируемого признакового пространства. Однако Э. использует понятия, постановки и методы решения задач и из многих других разделов математики: *теории вероятностей*, математического программирования, численных методов решения задач линейной алгебры, систем нелинейных уравнений, анализа неподвижных точек отображений.

Представленная схема при всей своей условности и неполноте в целом даёт общее наглядное представление об Э. и её месте в ряду других экономических и статистических дисциплин.

Именно «приземление» экономической теории на базу конкретной экономической статистики и извлечение из этого приземления с помощью подходящего математического аппарата вполне определённых количественных взаимосвязей – ключевые моменты в понимании сущности Э. Это, в частности, обеспечивает разграничение Э. с такими дисциплинами как математическая экономика, описательная экономическая статистика и *математическая статистика*. Математическая экономика, которая часто определяется как математически сформулированная экономическая теория, изучает взаимосвязи между экономическими переменными на общем (неколичественном) уровне. Она преобразуется в Э., когда символически представленные в этих взаимосвязях коэффициенты заменяются конкретными численными оценками, полученными на базе соответствующих экономических данных.

Из определения Э. следует, что предмет этой дисциплины – экономические и социально-экономические приложения, а именно модельное описание конкретных количественных взаимосвязей, существующих между анализируемыми показателями.

К числу типовых экономических моделей, конструируемых и изучаемых с помощью эконометрических методов, относятся: *производственные функции*, выражающие взаимосвязи между затратами и результатами производственной деятельности экономических систем различных уровней; модели функционирования национальной экономики; типологизация объектов и поведения агентов (стран, регионов, фирм, потребителей); целевые функции потребительского предпочтения и функции спроса; модели распределительных отношений в обществе; модели рынка и экономического равновесия; модели интернационализации национальных экономик; модели межстранового и межрегионального анализа и др.

При всём разнообразии спектра решаемых с помощью Э. задач их, тем не менее, было бы удобно расклассифицировать по трём направлениям: по конечным прикладным целям, по уровню иерархии и по профилю анализируемой экономической системы.

По направлению конечных прикладных целей выделим две осн.: прогноз экономических и социально-экономических показателей (переменных), характеризующих состояние и развитие анализируемой системы; имитация различных возможных сценариев социально-экономического развития анализируемой системы, когда статистически выявленные взаимосвязи между характеристиками произ-ва, потребления, социальной и финансовой политики и т.п. используются для прослеживания того, как планируемые (возможные) изменения тех или иных поддающихся управлению параметров произв-ва или распределения скажутся на значениях интересующих нас «выходных» характеристик (в специальной литературе исследования подобного рода называют также сценарным или *ситуационным анализом*.

По уровню иерархии анализируемой экономической системы выделяются макроуровень (т.е. страны в целом), мезоуровень (регионы, отрасли, корпорации) и микроуровень (семьи, пртия, фирмы).

В некоторых случаях должен быть определён профиль эконометрического моделирования: исследование может быть сконцентрировано на проблемах рынка, инвестиционной, финансовой или социальной политики, ценообразования, распределительных отношений, спроса и потребления, или на определённом комплексе проблем. Однако чем претенциознее по широте охвата анализируемых проблем эконометрическое исследование, тем меньше шансов провести его достаточно эффективно.

Метод эконометрики в общей формулировке: постулируется, что анализируемые переменные (экономические показатели) $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ являются случайными величинами, совместный *закон распределения вероятностей* (ЗРВ) которых не известен исследователю, но принадлежит некоторому семейству функций, В процессе функционирования анализируемой экономической системы генерируются наблюдаемые значения $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}$ $(i = 1, 2, \ldots, n)$ интересующих исследователя переменных. Идентификация модели (анализируемой системы) заключается в выборе из упомянутого семейства конкретного ЗРВ наиболее хорошо (в определённом смысле) согласующегося с имеющимися в распоряжении исследователя сгенерированными системой данными. Различные спецификации (конкретизации, основанные на дополнительных исходных допущениях) этой общей постановки проблемы и приводят к широкому спектру методов и моделей эконометрического анализа*: регрессии*, *временным рядам*, системам одновременных уравнений и другим методам, используемым при решении задач экономического прогноза, ситуационного анализа, оценивания важных экономических характеристик.

Все *модели эконометрические*, независимо от того, относятся они ко всему хоз-ву или к его элементам (т.е. к макроэкономике, отрасли, фирме или рынку), имеют некоторые общие особенности. Во-первых, они основаны на предположении, что поведение экономических переменных определяется с помощью совместных и одновременных операций с некоторым числом экономических соотношений. Во-вторых, принимается гипотеза, в силу которой модель, допуская упрощение сложной действительности, тем не менее улавливает гл. характеристики изучаемого объекта. В-третьих, создатель модели полагает, что на основе достигнутого с её помощью понимания реальной системы удастся предсказать её будущее движение и, возможно, управлять им в целях улучшения экономического благосостояния. Напр., предположим, что экономическая теория позволяет сформулировать следующие положения: потребление есть возрастающая функция от имеющегося в наличии дохода, но возрастающая, видимо, медленнее, чем рост дохода; объём инвестиций есть возрастающая функция национального дохода и убывающая функция некоторых характеристик гос. регулирования (напр., нормы процента); национальный доход есть сумма потребительских, инвестиционных и гос. закупок товаров и услуг.

Первая задача эконометриста – перевести эти положения на математический язык. Здесь открывается многообразие возможных решений, удовлетворяющих сформулированным априорным требованиям теории. Какие соотношения

выбрать между переменными – линейные или нелинейные? Если остановиться на нелинейных, то какими они должны быть – логарифмическими, полиномиальными или какими-либо ещё? Даже после определения формы конкретного соотношения, остаётся ещё нерешённой проблема выбора для различных уравнений запаздываний по времени. Будут ли, напр., инвестиции текущего периода реагировать только на национальный доход, произведённый в последнем периоде, или же на них скажется динамика нескольких предыдущих периодов? Обычный выход из этих трудностей состоит в выборе при первоначальном анализе наиболее простой из возможных форм этих соотношений. Тогда появляется возможность записать на основе указанных выше положений линейную модель, относительно анализируемых переменных и аддитивную относительно случайных составляющих:

$$y_t^{(1)} = \alpha_0 + \alpha_1(y_t^{(3)} - x_t^{(1)}) + \varepsilon_t^{(1)}, \qquad (1)$$

$$y_t^{(2)} = \beta_1 y_{t-1}^{(3)} + \beta_2 \cdot x_t^{(2)} + \varepsilon_t^{(2)}, \qquad (2)$$

$$y_t^{(3)} = y_t^{(1)} + y_t^{(2)} + x_t^{(3)}, \qquad (3)$$

где априорные ограничения выражены неравенствами

$$0 < \alpha_1 < 1; \beta_1 > 0; \beta_2 < 0.$$

Эти три соотношения вместе с ограничениями образуют модель. В ней $y_t^{(1)}$ обозначает потребление, $y_t^{(2)}$ – инвестиции, $y_t^{(3)}$ – национальный доход, $x_t^{(1)}$ – подоходный налог, $x_t^{(2)}$ – норму процента как инструмент гос. регулирования, $x_t^{(3)}$ – гос. закупки товаров и услуг, измеренные в «момент времени» *t*.

Присутствие в уравнениях (1) и (2) «остаточных» случайных составляющих $\varepsilon_t^{(1)}$ и $\varepsilon_t^{(2)}$ обусловлено необходимостью учесть влияние соответственно на $y_t^{(1)}$ и $y_t^{(2)}$ ряда неучтённых факторов. Действительно, нереалистично ожидать, что величина потребления $(y_t^{(1)})$ будет однозначно определяться уровнями национального дохода $(y_t^{(3)})$ и подоходного налога $(x_t^{(1)})$; аналогично величина инвестиций $(y_t^{(2)})$ зависит, очевидно, не только от достигнутого в предыдущий год уровня национального дохода $(y_{t-1}^{(3)})$ и от величины нормы процента $(x_t^{(2)})$, но и от ряда не учтённых в уравнении (2) факторов.

Модель (1)–(3) представляет собой пример СОУ. В данном примере потребление $(y_t^{(1)})$, инвестиции $(y_t^{(2)})$ и национальный доход $(y_t^{(3)})$ в текущий момент времени *t* – *эндогенные переменные*; подоходный налог $(x_t^{(1)})$, норма процента как инструмент гос. регулирования $(x_t^{(2)})$ и гос. закупки товаров и услуг $(x_t^{(3)})$ – *экзогенные переменные*, которые вместе с национальным доходом в предшествующий момент времени $(y_{t-1}^{(3)})$ образуют множество предопределённых переменных.

Полученная модель содержит два уравнения, объясняющих поведение потребителей и инвесторов, и одно тождество. Мы сформулировали её для дискретных периодов времени и выбрали запаздывание (лаг) в один период для отражения воздействия национального дохода на инвестиции.

Математико-статистический инструментарий Э. базируется, в основном, на избранных разделах многомерного статистического анализа и временных рядов анализа, развитых в направлении обобщений ряда традиционных для этих разделов постановок задач. Эти обобщения (подчас весьма далеко идущие) инициированы специфическими особенностями именно экономических приложений.

В понятие *регрессионного анализа* в Э. вкладывается широкий смысл. Оно включает в себя, в частности, *модель классическую линейную множественной регрессии* (МКЛМР) и связанный с ней *метод наименьших квадратов* (МНК), *обобщённую линейную модель множественной регрессии* (ОЛММР) и связанный с ней *обобщённый метод наименьших квадратов* (ОМНК), регрессию со стохастическими объясняющими переменными и связанный с ней *метод инструментальных переменных*. В рамках этого же раздела рассматриваются задачи построения регрессионной модели по неоднородным исходным данным (в связи с этим вводится понятие *фиктивных переменных* либо, если граница между однородными подвыборками исходных данных не определена, предлагается предварительно проводить их кластер-анализ),

а также по цензурированным данным или исходным данным, собранным в условиях, препятствующих формированию репрезентативной случайной выборки (в связи с этим рассматриваются различные модели, учитывающие смещения статистических выводов, вызванные ограничениями на отбор элементов выборки) – тобит-модель (т.н. “sample selection model”).

Цензурирование или урезание результатов выборочного обследования естественным образом возникает при исследовании «длительности жизни» какого-либо процесса или элемента, времени нахождения системы (элемента) в определённом состоянии: время жизни индивида, период безотказной работы прибора, время поиска работы безработным, длительность забастовки и т.п. Модели, описывающие механизм подобных явлений, называют моделями длительности жизни. Центральный объект исследования в подобных моделях – т.н. интенсивность отказов или коэффициент смертности λ_t, имеющий следующий смысл: если к моменту времени t процесс ещё не завершился (индивид не умер), то вероятность его окончания (смерти) в течение следующего малого промежутка времени Δ_t есть $\lambda_t \cdot \Delta t$. В эконометрических исследованиях, как правило,' пытаются описать, как интенсивность отказов λ_t, зависит от ряда экзогенных (объясняющих) переменных $x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(p)}$ (напр., в демографии исследуют зависимость коэффициента смертности от ряда социально-экономических характеристик индивида). В этом смысле эконометрические модели длительности жизни можно условно также отнести к разделу «Регрессионный анализ».

К этому же разделу относятся и регрессионные модели, в которых зависимая переменная имеет неколичественную природу, т.н. модели бинарного и множественного выбора (в т.ч. логит- и пробит-модели). Граничное положение (между разделами «Регрессионный анализ» и «Анализ временных рядов») занимают регрессионные *модели распределённых лагов*: постановка задачи здесь регрессионная, а исходные данные представлены в виде *временных рядов*.

Существенную роль в инструментарии Э. играет анализ временных рядов: *модели авторегрессии* порядка p ($\mathrm{AP}(p)$), скользящего среднего порядка q ($\mathrm{CC}(q)$), *авторегрессии – скользящего среднего* АРСС (p,q), *авторегрессии – проинтегрированного скользящего среднего* АРПСС (p,q,k), наконец, различные версии их многомерных обобщений (напр., векторные модели авторегрессии ВАР (*p*), векторные модели авторегрессии – скользящего среднего ВАРСС (*p,q*) и др.

В ряде прикладных эконометрических работ, в частности, при анализе и моделировании макроэкономических данных, характеризующих процессы инфляции и внешней торг., механизм формирования нормы процента и т. п., была выявлена некоторая общая закономерность в поведении случайных остатков (ошибок прогноза) ε исследуемых моделей: их малые и большие значения группировались целыми кластерами, или сериями. Причём это не приводило к нарушению их стационарности и, в частности, их *гомоскедастичности* для относительно больших временных интервалов, т.е. гипотеза $D\varepsilon_t = \mathrm{const}$ не противоречила имеющимся экспериментальным данным. Однако в рамках моделей АРСС удовлетворительно объяснить этот феномен не удавалось, требовалась определённая модификация известных моделей.

Такая модификация была предложена впервые Р. Энглом в 1982. Он рассматривал остатки ε_t, как условно гетероскедастичные, связанные друг с другом простейшей авторегрессионной зависимостью, а именно:

$$\begin{cases} [\varepsilon_t \mid \varepsilon_{t-1}] \in N(0;\sigma_t^2), \\ \text{где } \sigma_t^2 = D(\varepsilon_t \mid \varepsilon_{t-1}) = \theta_0 + \theta_1 \varepsilon_{t-1}^2, \end{cases} \qquad (4)$$

или, что то же самое:

$$\varepsilon_t = \delta_t[\theta_0 + \theta_1\, \varepsilon_{t-1}^2],$$

где последовательность δ_t, $t = 1, 2, \ldots$, образует стандартизованный нормальный *белый шум* (т.е. δ_{t_1} и δ_{t_2} независимы при $t_1 \neq t_2$ и $\delta_t \in N(0;1)$), а параметры θ_0 и θ_1 должны удовлетворять ограничениям, обеспечивающим безусловную *гомоскедастичность* ε_t (такими ограничениями являются требова-

ния $|\theta_1|<1, \theta_0>0$). При этом под $[\varepsilon_t \mid \varepsilon_{t-1}]$ подразумевается то, что речь идёт о случайной величине, рассматриваемой в предположении, что её значение в предшествующий момент времени зафиксировано (задано). Соответственно, её поведение будет описываться условным законом *распределения вероятностей*.

В соответствии с установившейся терминологией, модель (4) называется авторегрессионной условно гетероскедастичной (АРУГ). В англоязычной литературе такие модели называют *модель ARCH* (AutoRegressive Conditional Heteroscedasticity).

Использование такой модели для описания поведения остатков моделей регрессии и временных рядов в упомянутых выше типовых ситуациях оказывается более адекватным действительности и позволяет строить более эффективные оценки параметров рассматриваемых моделей, чем обычные или даже обобщённые МНК-оценки (напр., описание алгоритма построения нелинейных оценок макс. правдоподобия для параметров линейной модели множественной регрессии с авторегрессионными и условно гетероскедастичными остатками; получающиеся оценки оказываются более эффективными, чем даже наиболее эффективные, в классе линейных оценок, МНК-оценки.

Естественное обобщение моделей типа (4) было предложено Р. Энглом и Д. Крафтом:

$$\begin{cases} [\varepsilon_t \mid \varepsilon_{t-1}] \in N(0;\sigma_t^2), \\ \text{где } \sigma_t^2 = \theta_0 + \theta_1\varepsilon_{t-1}^2 + \ldots + \theta_q\varepsilon_{t-q}^2, \end{cases} \qquad (5)$$

а параметры $\theta_0, \theta_1, \ldots, \theta_q$ связаны некоторыми ограничениями, обеспечивающими безусловную гомоскедастичность остатков ε_t.

Модели (5) называются моделями АРУГ порядка q (АРУГ(q)). Содержательно переход к $q>1$ в моделях (5) означает, что процесс формирования значений остатков ε_t имеет «более длинную память» о величинах предшествующих остатков $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$. АРУГ ($q$)-модель (5) может рассматриваться как некая специальная форма СС (q)-модели, что и используется при её анализе.

Дальнейшее обобщение моделей этого типа было сделано Т. Боллерслевом в 1986. Он предложил описывать поведение остатков ε_t с помощью обобщённой авторегрессионной условно гетероскедастичной модели (ОАРУГ-модели) , или, в англоязычном варианте – GARCH-model), которая записывается в виде:

$$\begin{cases} [\varepsilon_t \mid \psi(t)]\ \varepsilon N(0;\sigma_t^2), \\ \text{где условная дисперсия } \sigma_t^2 = D(\varepsilon_t \mid \psi(t)) \text{ имеет вид} \\ \sigma_t^2 = \alpha_1\sigma_{t-1}^2 + \alpha_2\sigma_{t-2}^2 + \ldots + \alpha_p\sigma_{t-p}^2 + \theta_0 + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q}. \end{cases} \qquad (6)$$

В соотношениях (6) под $\psi(t)$ подразумевается вся информация о процессе ε_t, которой мы располагаем к моменту времени t (т.е. все значения ε_τ и σ_τ^2 для $\tau<t$), а параметры α_k и θ_j $(k=1,2,\ldots,p;\ j=0,1,\ldots,q)$ связаны ограничениями, обеспечивающими безусловную гомоскедастичность остатков ε_t. Модель ОАРУГ (p,q), задаваемая соотношениям (6), может интерпретироваться как специальная форма АРСС (p,q)-модели. На ряде примеров показано, что использование ОАРУГ (p,q)-модели позволяет добиваться более экономной параметризации в описании поведения остатков ε_t, чем в рамках АРУГ (q)-моделей (т.е. модели ОАРУГ (p,q) при малых значениях p и q оказываются более точными, чем АРУГ((q)-модели при больших значениях q).

Другие важные понятия, используемые при анализе временных рядов – понятия интегрируемости ряда (определённого порядка) и контеграции временных рядов. Одними из первых эти понятия рассмотрели Р. Энгл и К. Грэнжер в связи с задачей построения модели регрессии по нестационарным временным рядам. Временно́й ряд x_t называется интегрируемым порядка k, если он становится впервые стационарным после k-кратного применения к нему разностного оператора Δ. В регрессионном анализе обычно одновременно рассматривается несколько временных рядов. Очевидно, если x_1 – интегрируемый временно́й ряд порядка k_1 и

y_t – интегрируемый временно́й ряд порядка k_2, причём $k_2 > k_1$, то при любом значении параметра θ (в том числе при $\theta = \theta_{\text{мнк}}$, где $\hat{\theta}_{\text{мнк}}$ – МНК-оценка коэффициента регрессии в модели парной регрессии y по x) случайный остаток $\varepsilon_t = y_t - \theta x_t$ будет интегрируемым временным рядом порядка k_2. Если же $k_1 = k_2 = k$, то константа θ может быть подобрана так, что ε_t будет стационарным (или интегрируемым порядка 0) с нулевым средним. При этом вектор (1; $-\theta$) (или любой другой, отличающийся от этого сомножителем) называется коинтегрирующим. При регрессионном анализе временных рядов x_t и y_t их *коинтеграция* (согласование порядков их интегрируемости) производится обычно по схеме: 1. рассматривается модель $y_t = \theta x_t + \varepsilon_t$ и строится МНК-оценка $\hat{\theta}_{\text{мнк}}$ для параметра θ; 2. ряд $\hat{\varepsilon}_t = y_t - \hat{\theta}_{\text{мнк}} x_t$, анализируется на стационарность в рамках одной из моделей АРСС(p,q), напр., в рамках АР(1)-модели проверяется гипотеза $|\alpha| < 1$ в представлении $\hat{\varepsilon}_t = \alpha \hat{\varepsilon}_{t-1} + \delta_t$; 3. если результат отрицательный, то возвращаются к спецификации исходной модели, пробуя в качестве зависимой и объясняющей переменных различные варианты $\Delta^{k_1} y_t$ и $\Delta^{k_2} x_t$.

Неприменимость (в общем случае) обычного МНК как средства получения состоятельных оценок для неизвестных параметров *системы одновременных уравнений* (СОУ) инициировала разработку ряда специальных методов идентификации СОУ: косвенного МНК, *двух-* и *трёхшагового методов наименьших квадратов* (2МНК и 3МНК), *метода макс. правдоподобия* с ограниченной и с полной информацией, *метода инструментальных переменных* и т.п. Поэтому правомерно выделить проблематику построения и анализа СОУ в качестве одного из трёх осн. разделов Э.

В общих чертах образ действий при идентификации СОУ описывается: 1. методы статистического оценивания параметров СОУ подразделяются на два класса: методы, предназначенные для оценки параметров одного отдельно взятого уравнения системы (МНК, косвенный МНК, 2МНК, метод макс. правдоподобия с ограниченной информацией), методы, предназначенные для одновременного оценивания параметров всех уравнений системы с учётом их взаимосвязей (3МНК, метод макс. правдоподобия с полной информацией); 2. если уравнения *структурной формы модели* могут быть расположены в таком порядке, что i-е уравнение (i=1,2,…,m) может содержать в качестве объясняющих *эндогенных переменных* только переменные $y^{(1)}, y^{(2)}, \ldots, y^{(i-1)}$ (или часть из них), а случайное возмущение $\varepsilon_t^{(i)}$ этого уравнения не коррелирует со всеми этими эндогенными переменными, то такая система называется рекурсивной, и последовательное применение к каждому уравнению такой системы обычного МНК даёт *оценки состоятельные* её структурных параметров. Класс рекурсивных систем – простейший с точки зрения решения задачи оценивания структурных параметров СОУ; 3. если исследователя интересуют только параметры приведённой формы и задача прогноза эндогенных переменных, то он может ограничиться применением обычного метода наименьших квадратов к каждому отдельному уравнению приведённой формы (с последующей оценкой, если это необходимо, идентифицируемых параметров структурной формы). Такой образ действий называют косвенным методом наименьших квадратов, или методом наименьших квадратов без ограничений, а оценки, полученные с его помощью, будут состоятельными; 4. в ситуациях, когда среди уравнений системы имеются неидентифицируемые, так же как и в случаях, когда оценивание и анализ параметров структурной формы представляют для исследователя самостоятельный интерес, обычно применяют двухшаговый метод наименьших квадратов (2МНК). Этот метод предназначен для оценивания параметров отдельного уравнения структурной формы, а его последовательное применение к каждому из уравнений структурной формы СОУ позволяет получить состоятельные оценки всех структурных параметров (хотя 2МНК и не учитывает возможные взаимосвязи между уравнениями системы); 5. сущность двух шагов 2МНК заключается: на 1-м шаге для каждой эндогенной переменной, играющей роль объясняющей в анализируемом уравнении структурной формы, с помощью обычного МНК строится регрессия

на все предопределённые переменные X. На 2-м шаге эта эндогенная переменная заменяется в рассматриваемом уравнении её регрессионным выражением через X, после чего в правой части этого уравнения остаются только предопределённые переменные и к нему применяется обычный МНК. В моделях с большим числом предопределённых переменных в целях снижения размерности рекомендуется на 1-м шаге строить регрессию предикторной эндогенной переменной не на все предопределённые переменные, а лишь на небольшое число их *гл. компонент*; 6. если структурные случайные возмущения $\varepsilon_t^{(i)}$ различных уравнений системы взаимно коррелированы, то для оценивания структурных параметров рекомендуется применять другие методы, напр., трёхшаговый метод наименьших квадратов (3МНК). Этот метод предназначен для одновременного оценивания структурных параметров всех уравнений системы и даёт их состоятельные оценки, по эффективности превосходящие оценки (тоже состоятельные) Д2НК; 7. 3МНК использует полученные на первых двух шагах 2МНК оценки структурных параметров для вычисления оценки ковариационной матрицы возмущений различных уравнений структурной формы. Затем на 3-м шаге оценки структурных параметров системы пересчитываются с помощью обобщённого МНК в рамках соответствующей схемы ОЛММР, в которой в качестве *ковариационной матрицы* остатков используется полученная ранее оценка ковариационной матрицы возмущений; 8. при определённой ситуации могут оказаться полезными и другие методы статистического оценивания параметров СОУ. Для оценивания параметров одного отдельно взятого уравнения, напр., метод макс. правдоподобия с ограниченной информацией (требующий дополнительного априорного предположения о нормальном характере распределения структурных возмущений модели), для одновременной оценки всех структурных параметров системы может использоваться метод макс. правдоподобия с полной информацией; 9. одна из гл. конечных прикладных целей построения и анализа эконометрических моделей в виде СОУ – точечный и интервальный прогноз эндогенных переменных по заданным значениям предопределённых переменных и связанная с этим задача проведения многовариантных сценарных расчётов, показывающих, как будут «себя вести» эндогенные переменные при различных сочетаниях значений предопределённых переменных. «Точечное» решение этих задач основано на подсчёте значений эндогенных переменных с помощью статистически оценённой приведённой формы СОУ. Для получения «интервальных» вариантов решения необходимо уметь оценивать ковариационную матрицу ошибок точечного прогноза, что является задачей аналитически достаточно сложной.

Структуризация перечисленных разделов Э. основана на специфике типовых постановок задач, решаемых в рамках каждого из этих разделов прикладных. Однако, говоря о содержании эконометрики, следует упомянуть и о развиваемом в рамках этой дисциплины методологическом базисе, компоненты которого могут использоваться при решении всех типов задач, перечисленных выше. К осн. составляющим этого методологического базиса относятся: метод макс. правдоподобия; обобщённый метод моментов; теория больших выборок, или асимптотические результаты *теории вероятностей*; методы анализа *панельных данных*, т.е. многомерных исходных данных, регистрируемых на совокупности одних и тех же объектов в течение ряда тактов времени; *непараметрические* и полупараметрические *методы статистики*; статистические методы классификации: дискриминантный и кластер-анализы; статистические методы снижения размерности: гл. компоненты, факторный анализ и пр.; теория имитационно-компьютерного эксперимента: *метод Монте-Карло*, *бутстреп-моделирование*, перекрёстный компьютерный анализ дееспособности модели (cross-validation method) и пр.

Поскольку все эти направления исследований разрабатываются также и в рамках дисциплины «*Математическая статистика*», подчас трудно определить, какие из работ и научных результатов данной проблематики следует отнести к Э., а какие – к математической статисти-

ке. Отличительная особенность эконометрических работ – такая модификация классических постановок задач, которая инициируется спецификой именно экономических приложений.

ЭКСТРЕМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ КЛАССИФИКАЦИИ

формулировка данной задачи, при которой необходимо из определённого множества всех возможных разбиений исходных признаков найти наилучшее разбиение в смысле определённого количественного критерия.

В качестве оптимизируемого критерия используются функционалы качества разбиения, определенные на множестве всех возможных разбиений. Выбор таких функционалов обычно определяется целью и задачей исследования и опирается скорее на эмпирические и профессионально-интуитивные соображения, чем на какую-либо строгую формализованную систему. Э.п.з.к. различаются в зависимости от того, задано или нет число классов разбиения. В задачах первого типа используются, функционалы качества разбиения, определенные на множестве разбиений $S_1, S_2, ..., S_k$ исходных наблюдений $X_1, X_2, ..., X_n$ с учётом метрики *d* в *p*-мерном пространстве признаков. К наиболее распространенным относятся: сумма внутриклассовых дисперсий

$$Q_1(S) = \sum_{l=1}^{k} \sum_{X_i \in S_l} d^2(X_i, \overline{X}(l));$$

сумма попарных внутриклассовых расстояний

$$Q_1(S) = \sum_{l=1}^{k} \sum_{\substack{X_i \in S_l \\ X_j \in S_l}} d^2(X_i, X_j);$$

обобщённая внутриклассовая дисперсия

$$Q_3(S) = det(\sum_{l=1}^{k} n_l \hat{\Sigma}_l).$$

При неизвестном числе классов разбиения функционалы качества разбиения выбирают чаще всего в виде алгебраической комбинации двух функционалов. Наиболее общим подходом является использование схемы, предложенной А.Н. Колмогоровым, в которой один из функционалов $Z_\tau(S)$ представляет собой меру концентрации точек, а второй $I_\tau(S)$ – меру внутриклассового рассеивания:

$$Z_\tau(S) = \left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\nu(X_i)}{n} \right)^\tau \right|^{\frac{1}{\tau}},$$

$$I_\tau(S) = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\nu(X_i)} \sum_{X_l \in S(X_i)} d(X_i, X_l) \right|^{\frac{1}{\tau}},$$

$\nu(X_i)$ – число элементов в кластере, содержащем X_i, τ – параметр, определяемый эвристически.

В таких случаях наиболее часто используют Э.п.з.к. в двух вариантах: комбинирование функционалов качества: требуется найти такое разбиение S^*, для которого некоторая алгебраическая комбинация функционалов $Z_\tau(S)$ и $I_\tau(S)$ достигала бы экстремума; двойственная формулировка: требуется найти разбиение S^*, которое, обладая концентрацией $Z_\tau(S)$, не меньшей заданного значения Z_0, давало бы наименьшее внутриклассовое рассеяние $I_\tau(S)$, и при заданном пороговом значении I_0 найти разбиение S^* с внутриклассовым рассеянием $I_\tau(S^*) \le I_0$ и наибольшей концентрацией $Z_\tau(S)$.

См. также *Классификация многомерных наблюдений*.

ЭНДОГЕННЫЕ ПЕРЕМЕННЫЕ

переменные, которые определяются внутри модели; могут также называться зависимыми переменными, функциями отклика. Все переменные в *эконометрической модели* делятся на две категории: э. (зависимые) п., значения которых хотят находить с помощью модели, и предопределённые переменные, значения которых известны и должны быть использованы для определения значений эндогенных переменных. В свою очередь предопределённые переменные подразделяются на два типа: *экзогенные* (независимые) *переменные*, значения которых определяются (задаются) вне модели и лаговые (прошлые) значения Э.п., определяемые имеющейся статистикой за предыдущие моменты времени. Различают структурную и приведённую форму эконо-

метрической модели. В уравнениях приведённой формы в левой части стоят Э.п., а в правой части предопределённые переменные: экзогенные переменные и лаговые значения Э.п. В структурной форме уравнения могут содержать текущие значения Э.п. как в левой, так и в правой части уравнений, что затрудняет оценивание параметров модели. Поэтому для оценивания коэффициентов обычно переходят к приведённой форме. Однако затем требуется обратный переход к структурной форме, удобной для содержательной интерпретации зависимостей. Однако это не всегда возможно. Возникает проблема *идентифицируемости* эконометрической модели. Известны необходимые и достаточные условия идентифицируемости.

ЭТАПЫ СТАТИСТИЧЕСКОГО ИССЛЕДОВАНИЯ ЗАВИСИМОСТЕЙ

Важнейшие методами статистического исследования зависимостей – графический, *корреляционный* и *регрессионный анализ*. Принято выделять осн. Э.с.и.з.: формулировка цели исследования; определение круга показателей, связь между которыми изучается; отбор статистических данных, характеризующих динамику этих показателей; выполнение графического анализа данных, отобранных в выборку по каждому показателю в отдельности, обращая особое внимание на нетипичные наблюдения, которым желательно найти содержательное объяснение; проведение графического анализа интересующих зависимостей, строя попарно график зависимости одного показателя от другого, определение математических функций (кривых) наиболее подходящих для отражения этих зависимостей; выполнение *корреляционного анализа* взаимосвязей показателей, оценка *корреляционной матрицы*, установка сильных и слабых связей между переменными; переход к *регрессионному анализу*, начиная с выдвижения гипотезы о виде математического уравнения,

$$x_{t-\tau}x_t = a_1x_{t-\tau}x_{t-1} + a_2x_{t-\tau}x_{t-2} + \ldots + a_px_{t-\tau}x_{t-p} + x_{t-\tau}\varepsilon_t \quad (2),$$

переходя к *математическим ожиданиям* обеих частей равенства, получаем уравнение для ковариаций:

$$c_\tau = a_1c_{\tau-1} + a_2c_{\tau-2} + \ldots + a_pc_{\tau-p} \quad (3).$$

выражающего влияние различных показателей (возможно, преобразованных) на интересующий исследователя показатель, опираясь на экономическую теорию, содержательную концепцию, результаты графического и корреляционного анализа; выбор метода оценивания уравнения: оценка построенного уравнения, рассчёт критериев качества регрессии; проверка качества регрессии, значимости коэффициентов, автокорреляции остатков и т.п.; перестройка регрессии, если показатели качества оказались неудовлетворительными, и возвращение к этапу оценивания регрессии или даже к более раннему этапу. В случае возникновения проблемы *мультиколлинеарности* принять меры к её устранению, к понижению размерности объясняющих переменных, используя *метод гл. компонент* или *факторный анализ*; выполнение содержательного анализа полученной модели при условии построения регрессии с хорошими статистическими показателями, интерпретация коэффициентов и принятие окончательного решения о целесообразности использования построенной модели для анализа, управления или *прогнозирования*.

Ю

ЮЛА-УОКЕРА УРАВНЕНИЯ

система линейных уравнений, связывающая коэффициенты авторегрессионного уравнения с коэффициентами автокорреляции.

Пусть дан стационарный *временной ряд*, описываемый авторегрессионной моделью

$$x_t = a_1x_{t-1} + a_2x_{t-2} + \ldots + a_px_{t-p} + \varepsilon_t ,$$

где ε_t – *белый шум*. *Автокорреляционная функция* этого процесса вычисляется с помощью рекуррентного соотношения по первым p её значениям $r_1, r_2, \ldots, r_p$. Это рекуррентное соотношение выводится путём умножения всех членов соотношения (1) на $x_{t-\tau}$ ($\tau>p$), получим:

Отметим, что математическое ожидание последнего слагаемого в (2) равно нулю, т.к. при $\tau>0$ $x_{t-\tau}$ не зависит от будущих ε_t. Поделив все члены уравнения (2) на дисперсию c_0,

находим рекуррентное соотношение для коэффициентов автокорреляции, позволяющее вычислить любой член автокорреляционной функции процесса x_t по её первым p значениям:

$$r_\tau = a_1 r_{\tau-1} + a_2 r_{\tau-2} + ... + a_p r_{\tau-p} \quad (4).$$

Теперь последовательно подставив в (4) значения τ = 1, 2, …, p, получим систему линейных уравнений относительно коэффициентов регрессии a_1, a_2,…,a_p:

$$\left.\begin{array}{l} r_1 = a_1 + a_2 r_1 + ... + a_p r_{p-1} \\ r_2 = a_1 r_1 + a_2 + ... + a_p r_{p-2} \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ r_p = a_1 r_{p-1} + a_2 r_{p-2} + ... + a_p \end{array}\right\} \quad (5).$$

Эти уравнения и называют Ю.-У.у.

Решение этой системы уравнений удобно получать в матричном виде. Введём матричные обозначения:

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ ... \\ a_p \end{pmatrix};\ \mathbf{r} = \begin{pmatrix} r_1 \\ r_2 \\ ... \\ r_p \end{pmatrix};$$

$$\mathbf{R} = \begin{pmatrix} 1 & r_1 & r_2 & ... & r_{p-1} \\ r_1 & 1 & r_1 & ... & r_{p-2} \\ . & . & . & . & . \\ r_{p-1} & r_{p-2} & r_{p-3} & ... & 1 \end{pmatrix}.$$

Тогда система (5) может быть переписана как Ra=r , а её решением будет $a=R^{-1}r$.

А

АНАЛИЗ СПЕКТРАЛЬНЫЙ

аналитический метод выявления периодической составляющей (сезонной или циклической) *временного ряда*, заключающийся в разложении исходного временного ряда на периодические функции синусов и косинусов с различной частотой колебаний и амплитуды. А.с. также называют гармоническим анализом или анализом Фурье. Осн. смысл преобразования Фурье состоит в том, что исходная непериодическая функция произвольной формы, которую невозможно описать аналитически и которая в общем случае трудна для обработки и анализа, представляется в виде совокупности синусов и косинусов с различной частотой и амплитудой. Иными словами, сложная функция представляется в виде совокупности более простых. Каждая синусоида (или косинусоида) называется спектральной составляющей или гармоникой. Один из возможных способов сделать это – решить задачу линейной множественной регрессии, где зависимая переменная – наблюдаемые значения временного ряда, а независимые переменные-регрессоры – функции синусов и косинусов всех возможных частот. Такая модель линейной регрессии может быть представлена в виде:

$$\bar{y}(t) = a_0 + \sum (a_k \cdot \cos(k \cdot t) + b_k \cdot \sin(k \cdot t)),$$

где k=1, … , m – порядок гармоники Фурье.

Коэффициенты при косинусах и коэффициенты при синусах – это коэффициенты регрессии, показывающие степень, с которой соответствующие функции коррелируют с данными. Сами синусы и косинусы на различных частотах не коррелированны друг с другом или ортогональны. На практике, как правило, рассчитывают не более 4 гармоник, т.е. берут k=1,2,3,4. Итак, А.с. определяет корреляцию функций синусов и косинусов различной частоты с наблюдаемыми данными. Если найденная корреляция (коэффициент регрессии при определенном синусе или косинусе значим) велика, то можно сделать вывод, что существует строгая периодичность на соответствующей частоте в данных. Параметры уравнения регрессии, содержащей гармоники Фурье, определяются методом наименьших квадратов (МНК), т.е. при условии минимизации суммы квадратов отклонений эмпирических значений временного ряда от теоретических, полученных по гармоникам Фурье. Оценки параметров модели находится через решение системы нормальных уравнений по формулам:

$$\begin{cases} a_0 = \frac{1}{n}\sum y \\ a_k = \frac{2}{n}\sum y \cdot \cos(k \cdot t) \\ b_k = \frac{2}{n}\sum y \cdot \sin(k \cdot t) \end{cases}$$

Для успешного применения А.с. должна соблюдаться следующая предпосылка. Исходный временной ряд должен иметь достаточную протяжённость, содержащий несколько волн периодических колебаний.

АППРОКСИМАЦИЯ ФУНКЦИЙ

(от англ. – approximation) – математический метод, в основе которого лежит замена одних математических объектов другими, в том или ином смысле близкими к исходным, но более простыми. Аппроксимация позволяет исследовать числовые характеристики и качественные свойства объекта, сводя задачу к изучению более простых или более удобных объектов (напр., таких, характеристики которых легко вычисляются, или свойства которых уже известны). Выбор вида математической функции $f(x)$ может быть осуществлен тремя методами: графическим; аналитическим, т.е. исходя из теории изучаемой взаимосвязи; экспериментальным.

При изучении зависимости между двумя признаками графический метод подбора вида уравнения регрессии достаточно нагляден и базируется на *поле корреляции*. На поле корреляции просто подбирается функция, график которой проходит через наибольшее количе-

ство точек или как можно ближе к ним. Применение аналитического метода основано на изучении материальной природы связи исследуемых признаков, теоретических соображений и опыта подобных предыдущих исследований. Экспериментальный метод предусматривает перебор нескольких функций, т.е. для каждой функции строится уравнение регрессии и оценивается их качество. Сравнивая полученные показатели, выбирают лучшую модель. Наиболее часто этот метод используется при обработке информации на компьютере. Осн. показателем при переборе функций является остаточная дисперсия $D_{ост}$. Необходимо отметить, что сравнивать можно только остаточные дисперсии по одним и тем же данным, но рассчитанным по различным уравнениям регрессии. Если для нескольких функций $D_{ост}$ одинаковы, то предпочтение отдается наиболее простому, более удобному для понимания виду уравнения регрессии. Результаты многих исследований показывают, что число наблюдений должно в три-пять раз превышать число рассчитываемых параметров модели.

В

ВАР–МОДЕЛЬ (VAR–МОДЕЛЬ)

векторная модель *авторегрессии;* обычно применяется для систем *прогнозирования* взаимосвязанных временных рядов и для анализа динамического влияния случайных возмущений на систему переменных. Подход к построению В.-м. обходит потребность в структурном моделировании, рассматривая каждую эндогенную переменную в системе как функцию от *лагированных значений* всех эндогенных переменных. Можно выделить три различных формы В.-м.: приведённая, рекурсивная и структурная. Все три являются динамическими линейными моделями, которые связывают текущие и прошлые значения вектора Y_t *n*-мерного временного ряда. Приведённая форма и рекурсивные В.-м.– статистические модели, которые не используют никакие экономические соображения за исключением выбора переменных. Эти В.-м. используются для описания данных и прогноза. Структурная В.-м. включает ограничения, полученные из макроэкономической теории, и эта В.-м. используется для структурного вывода и анализа политики. Приведённая форма В.-м. выражает Y_t в виде распределённого лага прошлых значений плюс серийно некоррелированный член ошибки, т.е. обобщает одномерную авторегрессию на случай векторов. Математически приведённая форма модели В.-м. – система n уравнений, которые можно записать в матричной форме:

$$Y_t = \alpha + A_1 Y_{t-1} + ... + A_p Y_{t-p} + \varepsilon_t , ,$$

где где α – это $n \times 1$ вектор констант, A_1, A_2, ..., A_p – это $n \times n$ матрицы коэффициентов, а ε_t, – это $n \times l$ вектор серийно некоррелированных ошибок, о которых предполагается, что они имеют среднее ноль и матрицу ковариаций Σ_ε. Если элементы Yt являются эндогенными, то это приводит к тому, что ошибки в приведённой форме модели являются коррелированными между уравнениями. Модель векторной авторегрессии для двух рядов допускает включение в правые части уравнений большего количества запаздываний этих переменных. Наибольший порядок запаздываний, включаемых в правую часть, называется порядком векторной авторегрессии. Если этот порядок равен p , то для такой модели используют обозначение VAR(p). На уровне матричных уравнений рекурсивная и структурная В.-м. выглядят одинаково. Эти две модели В.-м. учитывают в явном виде одновременные взаимодействия между элементами Yt, что сводится к добавлению члена к правой части уравнения (1.). Обе формы, рекурсивная и структурная В.-м. представляются в общем виде:

$$Y_t = \beta + B_0 Y_t + B_1 Y_{t-1} + ... + B_p Y_{t-p} + \eta_t , \quad (2)$$

де β – вектор констант, B_0,..., B_p – матрицы, а η_t – ошибки/

ВЕРОЯТНОСТЬ ОШИБОЧНОЙ КЛАССИФИКАЦИИ

характеристика качества метода классификации. Если множества, используемые в качестве обучающих выборок, близко расположены друг к другу, то возрастает вероятность ошибочной классификации новых объектов, особенно в тех

случаях, когда классифицируемый объект сильно удалён от центров обоих множеств.

Предпочтительны такие методы классификации, которые минимизируют потери (или вероятность) неправильной классификации объектов. Введем $c(j \mid i)$ – «функцию потерь», которая определяет стоимость потерь от отнесения объекта i-го класса к классу с номером j. Если в процессе классификации такие ошибки встречаются $m(j \mid i)$ раз, то потери, связанные с отнесением объектов i-го класса к классу j равны произведению $m(j \mid i)c(j \mid i)$. Чтобы подсчитать общие потери Cn при такой процедуре классификации, надо просуммировать величину произведения $m(j \mid i)c(j \mid i)$ по всем i= 1, 2, … , k и j= 1, 2, … , k, т.е.

$$Cn = \sum_{i=1}^{k}\sum_{j=1}^{k} c(j \mid \mathrm{i})\mathrm{m}(\mathrm{j} \mid \mathrm{i}),$$

где κ – число классов (1). Для того, чтобы потери не зависели от числа n классифицируемых объектов, переходят к удельной характеристике потерь, разделив обе части на n, а затем к пределу по $\mathrm{n} \to \infty$:

$$\mathrm{C} = \lim_{n\to\infty}\left(\frac{1}{n}C_n\right) = \lim_{n\to\infty}\sum_{i=1}^{k}\sum_{j=1}^{k} c(j \mid \mathrm{I})\,\frac{m(j \mid i)}{n_i(n)}\,\frac{n_i(n)}{n} = \sum_{i=1}^{k}\pi_i\sum_{j=1}^{k} c(j \mid i)P(j \mid i),$$

где $\mathrm{n_i(n)}$ – число объектов в i-м классе (2). Предел в (2) понимается в смысле сходимости по вероятности частот $m\ (j \mid i)/n_i(n)$ и $n_i(n)/n$ соответственно к вероятностям $P(j \mid i)$ – отнести объект класса i к классу j, и π_i – извлечения объекта класса i из общей совокупности анализируемых объектов; величину π_i называют также априорной вероятностью (или удельным весом) класса i. Величина

$$C^i = \sum_{j=1}^{k} c(j \mid i)P(j \mid i)$$

определяет средние потери от неправильной классификации объектов i-го класса. Средние удельные потери от неправильной классификации всех анализируемых объектов будут:

$$C = \sum_{i=1}^{k} \pi_i C^{(i)}.$$

В достаточно широком классе ситуаций полагают, что потери $c(j \mid i)$ одинаковы для любой пары i и j, т.е. $c(j \mid i) = c_0 = const$ при $j \neq i;\ i, j = 1,2,...,k$. В этом случае стремление к минимизации средних удельных потерь C будет эквивалентно стремлению максимизации вероятности правильной классификации объектов, равной

$$\sum_{i=1}^{k} \pi_i P(i \mid i).$$

Часто при построении процедур классификации говорят не о потерях, а о вероятностях неправильной классификации:

$$(1 - \sum_{i=1}^{k} \pi_i P(i \mid i)).$$

ВРЕМЕННОЙ РЯД (РЯД ДИНАМИКИ, ДИНАМИЧЕСКИЙ РЯД)

упорядоченная во времени последовательность наблюдений. В англоязычной литературе для В.р. используется термин «time series». Если время измеряется непрерывно, то В.р. называется непрерывным, если же время фиксируется дискретно, то речь идет о дискретных временных рядах. Дискретные В.р. получаются различными способами: выборочным способом из непрерывных временных рядов через равные или произвольные промежутки времени; накоплением значений переменной в течение некоторых периодов времени, при этом рассматриваемые интервалы могут быть как равноотстоящими, так и неравномерными. Анализ последовательности значений статистических показателей (признаков), упорядоченных в хронологическом порядке, занимает видное место в статистической практике. Отдельные наблюдения В.р. называются уровнями этого ряда. Каждый В.р. содержит два элемента: значения времени; соответствующие им значения уровней ряда.

В качестве показателя времени во В.р. могут указываться либо определённые моменты времени (даты), либо отдельные периоды (сутки, мес., кв., полугодия, годы и т.д.). В зависимости от характера временного параметра ряды

делятся на моментные и интервальные. В моментных В.р. уровни характеризуют значения показателя по состоянию на определённые моменты времени. Напр., моментными являются В.р. цен на определённые виды товаров, ряды курсов акций, уровни которых фиксируются для конкретных чисел. Примерами моментных В.р. служат также ряды численности нас. или стоимости осн. фондов, т.к. значения уровней этих В.р. определяются ежегодно на одно и то же число. В интервальных рядах уровни характеризуют значение показателя за определённые интервалы (периоды) времени. Примерами служат ряды годовой (месячной, квартальной) динамики произ-ва продукции в натуральном или стоимостном выражении. Уровни В.р представляют собой абсолютные, относительные и средние величины. Если уровни представляют собой непосредственно не наблюдаемые значения, а производные – средние или относительные, то такие В.р. называются производными. Уровни этих рядов получаются с помощью некоторых вычислений на основе абсолютных показателей. Пример производного ряда ежемесячной динамики – В.р. среднесуточного произ-ва. Уровни этого временного ряда для каждого месяца получаются делением всей произведённой за месяц продукции на количество рабочих дней в месяце.

Важная особенность интервальных рядов динамики абсолютных величин – возможность суммирования их уровней. В результате этой процедуры получаются накопленные итоги, имеющие осмысленное содержание благодаря отсутствию повторного счёта. Напр., суммируя премиальный фонд работников пр-тия за первые два квартала текущего года можно получить премиальный фонд за первое полугодие, а сумма соответствующих значений по полугодиям позволит определить общий премиальный фонд за год.

Суммирование уровней моментного ряда динамики не практикуется, т.к. полученные накопленные итоги лишены всякого смысла. Напр., уровни моментного ряда «остатки вкладов нас. в банках на нач. месяца» содержат элементы повторного счёта. Второй уровень, относящийся к началу фев., частично содержит вклады нас., учтённые первым уровнем, зафиксированным на начало янв., и т.д. Т. О., моментные ряды динамики в отличие от интервальных не обладают свойством *аддитивности*. При исследовании моментного ряда динамики определённый смысл имеет расчёт разностей уровней, характеризующих изменение показателя за некоторый отрезок времени. На практике часто требуется проанализировать динамику показателя не только за данный отрезок времени, но и с учётом ряда предшествующих периодов. Для этого строится ряд динамики с нарастающими итогами, уровни которого дают обобщающий результат развития показателя с начала отчётного периода (квартала, полугодия, года и т.д.). Уровни ряда могут принимать детерминированные или случайные значения. В первом случае уровни В.р. принимают соответствующие значения с вероятностью единица. Примером ряда с детерминированными значениями уровней служит ряд последовательных данных о количестве дней в месяцах. Естественно, анализу, а в дальнейшем и *прогнозированию*, подвергаются ряды со случайными значениями уровней. Такие В.р. изучаются в эконометрике как случайные процессы, при этом В.р. рассматривается как частная реализация теоретического случайного процесса. Возможные значения В.р. в текущий момент времени t могут описываться с помощью случайной величины и связанного с ней *распределения вероятностей.*

Случайные или стохастические процессы в свою очередь могут подразделяться на стационарные и нестационарные, определяя тем самым важные черты генерируемых ими В.р.

Особое внимание следует обратить на понятия стационарности в узком и широком смысле.

В.р. иногда рассматривается как *выборка*. Однако следует учитывать, что В.р. имеют характерные отличия от пространственных выборок. Во-первых, в отличие от пространственных данных уровни временного ряда, как правило, не являются статистически независимыми. Во-вторых, члены В.р. не являются одинаково распределёнными. Очевидно, что эти особенности должны быть учтены в исследовательской работе. Успешность статистического анализа раз-

вития процессов во времени во многом зависит от правильного построения рядов динамики. Большое значение для дальнейшего исследования процесса имеет выбор интервалов между соседними уровнями ряда. Удобнее всего иметь дело с равноотстоящими друг от друга уровнями ряда. Одно из важнейших условий, необходимых для правильного отражения В.р. реального процесса развития – сопоставимость уровней ряда. Появление несопоставимых уровней может быть вызвано разными причинами: изменением методики расчёта показателя, изменением классификации, терминологии и т.д. Напр., уровни В.р., характеризующие количество малых пр-тий, могут оказаться несопоставимыми из-за изменения самого понятия «малое пр-тие». Подразумевается, что это понятие должно быть одинаковым для всего исследуемого периода. Важна сопоставимость, связанная с единицами измерения, кругом охватываемых объектов, методологией расчёта и др. Как правило, термин «В.р.» подразумевает, что этот ряд является одномерным. Если же практический интерес представляет рассмотрение совместной динамики набора В.р., то такой набор называют многомерным В.р.

К

КОЛЕБАНИЯ СЕЗОННЫЕ

регулярно повторяющиеся в динамике изменения показателя с периодом колебаний, не превышающим одного года. Чаще всего причина возникновения К.с. – природно-климатические условия. Примером могут служить колебания цен на с.-х. продукцию (напр., картофель), характеризующиеся снижением цен в период после уборки урожая и последующим повышением цен, связанным с необходимостью хранения продукции. Т.о., в колебаниях цен прослеживается годовая периодичность. Иногда причины К.с. имеют социальный характер, напр., увеличение закупок в предпраздничный период, увеличение платежей в кон. кв. и т.д. К.с. характеризуются длительностью периода колебаний (отрезок времени между соседними точками макс. и мин.), амплитудой (разностью между макс. и мин. значениями показателя) и размещением макс. и мин. во времени. Повторяющиеся изменения показателя, с периодом колебаний, превышающим один год, относят к циклическим. Примерами могут служить циклы деловой активности, исследованные Кондратьевым, демографические, инвестиционные и другие циклы. По многолетним наблюдениям установлена цикличность солнечной активности (с периодом колебаний примерно в 11 лет). В статистической практике при моделировании социально-экономических процессов К.с. отражают либо в виде аддитивной составляющей, прибавляемой к трендовым (или средним) значениям, либо в виде мультипликативной составляющей, на значения которой умножаются соответствующие трендовые (или средние) значения. В настоящее время при анализе и прогнозировании сезонных процессов используются подходы, связанные с применением индексов сезонности в сочетании с кривыми роста и скользящими средними, с использованием спектрального анализа, адаптивных моделей, основанных на экспоненциальном сглаживании, сезонного варианта модели ARIMA, совр. процедур сезонной корректировки и др.

Отличительная особенность аддитивного характера сезонности заключается в том, что амплитуда сезонных колебаний остается примерно постоянной, неизменной во времени. Иногда на стадии графического анализа можно определить характер сезонных колебаний: аддитивный или мультипликативный. Напр., на рис.1, где показана ежемесячная динамика инвестиций в осн. капитал в РФ, отчётливо прослеживаются устойчивые К.с., «наслаивающиеся» на возрастающий *тренд.*

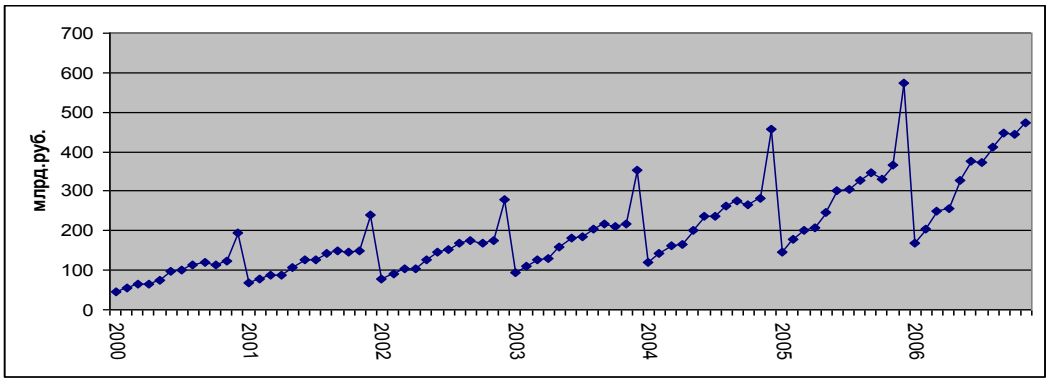


Рис. 1. **Инвестиции в основной капитал в фактически действовавших ценах (с 01.2000 по 11.2006)**

Видны ежегодно повторяющиеся пики активности, приходящиеся на последний месяц года. Причём амплитуда К.с. возрастает с ростом объёмов инвестиций, что приводит к выводу о мультипликативном характере сезонности.

КОЛЕБАНИЯ ЦИКЛИЧЕСКИЕ

факторы, формирующие изменения динамики *временного ряда*, обусловленные действием долговременных циклов экономической, демографической или астрофизической природы (волны Кондратьева, демографические «ямы», циклы солнечной активности и т.п.). К.ц. связаны с общей динамикой конъюнктуры рынка, а также с фазой бизнес-цикла, в которой находится экономика страны. Существует несколько подходов к анализу структуры временных рядов, содержащих К.ц. Причём моделирование К.ц. в целом осуществляется аналогично моделированию *колебаний сезонных*.

Простейший подход – расчёт значений сезонной компоненты методом скользящей средней и построение аддитивной или мультипликативной модели временного ряда. Общий вид аддитивной модели следующий: Y = T + S + E. Эта модель предполагает, что каждый уровень временного ряда может быть представлен как сумма трендовой (T), сезонной (S) и случайной (E) компонент. Общий вид мультипликативной модели выглядит так: $Y = T \cdot S \cdot E$. Эта модель предполагает, что каждый уровень временного ряда может быть представлен как произведение трендовой (T), сезонной (S) и случайной (E) компонент. Выбор одной из двух моделей осуществляется на основе анализа структуры сезонных колебаний. Если амплитуда колебаний приблизительно постоянна, строят аддитивную модель временного ряда, в которой значения сезонной компоненты предполагаются постоянными для различных циклов. Если амплитуда сезонных колебаний возрастает или уменьшается, строят мультипликативную модель временного ряда, которая ставит уровни ряда в зависимость от значений сезонной компоненты. Построение аддитивной и мультипликативной моделей сводится к расчёту значений T, S и E для каждого уровня ряда. Существует еще один метод моделирования временного ряда, содержащего циклические колебания – построение модели регрессии с включением фактора времени и фиктивных (дихотомических) переменных. Количество фиктивных переменных в такой модели должно быть на единицу меньше числа моментов (периодов) времени внутри одного цикла колебаний. Напр., при моделировании поквартальных данных модель должна включать четыре независимые переменные – фактор времени и три фиктивные переменные. Каждая фиктивная переменная отражает сезонную (циклическую) компоненту временного ряда для какого-либо одного периода. Она равна единице для данного периода и нулю для всех остальных периодов. Пусть имеется временной ряд, содержащий К.ц. периодичностью k. Модель регрессии с фиктивными переменными для этого ряда будет иметь вид:

$$y_t = a + b \cdot t + c_1 x_1 + \ldots + c_j x_j + \ldots + c_{k-1} x_{k-1} + \varepsilon_t,$$

где $x_j = \begin{cases} 1 \text{ для каждого наблюдения внутри } j-\text{го цикла}, \\ 0 \text{ во всех остальных случаях.} \end{cases}$

Л

ЛОГИСТИЧЕСКАЯ КРИВАЯ (КРИВАЯ ПЕРЛА-РИДА)

график функции, которая аналитически задается в виде:

$$y = \frac{k}{1 + a\ell^{-bx}},$$

где $k > 0, \quad a > 0, \quad b > 0$ - параметры.

Если $x \to -\infty$, то $y \to 0$, если $x \to +\infty$ то $y \to k$. График функции симметричен относительно точки перегиба с координатами:

$$x = lnb, \quad y = \frac{k}{2}.$$

Л.к. относится к классу т.н. *S*-образных кривых, моделирующих поведение определенных процессов. Для каждого такого процесса характерно разложение его на этапы, которым соответствует сначала рост значения объясняемой переменный с ускорением, затем – рост её значения с замедлением и, наконец, достижение переменной некоторого уровня насыщения. Такая специфика процесса обусловлена тем, что он, как правило, протекает под влиянием определенного ограничивающего воздействия, проявляющегося в характеристиках процесса с некоторого момента времени. Первая Л.к. была предложена для описания популяционной динамики при модификации закона Мальтуса, где роль ограничивающего воздействия играет нарастание дефицита жизненно важного ресурса. С помощью Л.к. описываются многие экономические, социальные, технологические, биологические и другие процессы, в которых есть место ресурсу, которого «на всех не хватит». Наглядным примером служит моделирование развития новой отрасли, когда сначала в условиях высоких издержек произ-ва выпуск продукции растёт медленно, затем при переходе к массовому произ-ву он ускоряется, а далее вместе с ростом конкуренции и насыщением рынка товаром замедляет свой рост и стабилизируется на определённом уровне. На рис. 1 приведена Л.к. с параметрами: $k = 1, \quad a = 1, \quad b = 1.$

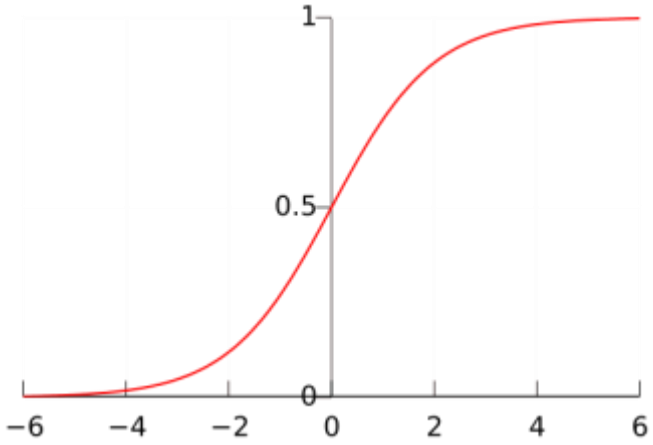


Рис. 1

См. также *Кривая Гомперца.*

Н

НАДЁЖНОСТЬ ПРОГНОЗА (УРОВЕНЬ ДОВЕРИЯ, ДОСТОВЕРНОСТЬ, ДОВЕРИТЕЛЬНАЯ ВЕРОЯТНОСТЬ)

оценка вероятности осуществления прогноза. При построении статистических прогнозов

важное значение имеет оценка его точности. Любой прогноз носит вероятностный характер, и оценка его точности определяется достоверностью. Следует различать понятия *точности прогноза* и Н.п. Точность прогноза оценивается по величине погрешности (ошибки) прогноза – величине, характеризующей расхождение между фактическим и прогнозным значением показателя. Н.п. определяется вероятностью наступления прогнозируемого события, т.е. реализацией соответствующей прогностической оценки. Чем она выше, тем выше Н.п. Вероятность реализации может быть оценена субъективно с помощью экспертных оценок (экспертное прогнозирование) или может быть связана с доверительными интервалами прогноза на основе статистической модели. В первом случае, надёжность, как правило, количественно не оценивается и определяется на основе интуитивно-логических методов. Различные методы прогнозирования характеризуются различными затратами времени на реализацию прогноза. Однако Н.п. во всех случаях с увеличением *периода упреждения прогноза* (дальности прогнозирования) уменьшается. Наибольшей надёжностью характеризуются кратко- и среднесрочные прогнозы. Особую сложность представляют долгосрочные прогнозы, т.к. практически всегда отсутствуют объективные условия для оценки их надежности и точности. В случае построения интервальной оценки прогнозируемого показателя, о точности прогноза можно говорить лишь как об интервале ожидаемых результатов, т.е. можно утверждать с определённой степенью достоверности, что количественные размеры прогнозируемых явлений будут изменяться в заданных пределах. Интервальный прогноз охватывает совокупность значений прогнозируемой величины в определённом строго заданном интервале. Совокупность методов и процедур, направленных на оценку достоверности прогноза называется верификацией прогноза.

Методы верификации дают возможность построить достаточно адекватные надежные статистические прогнозы, правильно отражающие тенденции в развитии прогнозируемого явления.

П

ПАРАМЕТР АДАПТАЦИИ

величина коэффициента, характеризующего силу реакции *модели адаптивной* на текущую ошибку прогноза, т.е. на изменение в динамике исследуемого *временного ряда* x_t . Значение этого параметра обычно лежит между 0 и 1. Выбор наилучшего П.а. осуществляется методом проб различных значений на ретроспективном (прошлом) статистическом материале и выборе значения, приводящего к наименьшей сумме квадратов ошибок прогнозов. Для экспоненциального сглаживания:

$$S_t = \beta S_{t-1} + \alpha x_t = S_{t-1} + \alpha(x_t - S_{t-1}) = S_{t-1} + \alpha e_t ,$$

где S_t – экспоненциальная средняя в момент t, которая принимается в простейшей адаптивной модели за прогноз будущего члена ряда x_{t+1}; α – постоянная сглаживания и является П.а., α=const, $0<\alpha<1$; β=1-α; e_t – ошибка прогнозирования на один шаг: $e_t=x_t-S_{t-1}$. Теоретически вопрос о выборе оптимального значения α для простейшей адаптивной модели рассмотрел Д.Мат (Muth J.F.). Адаптивные модели могут содержать переменные П.а., регулирование которых во времени определяется тем или иным встроенным в модель алгоритмом. В модели может быть несколько П.а. Г. Тейл и С. Вейдж исследовали теоретически вопрос оптимальности двух П. а. в модели со стохастическим трендом.

См. также *Модель Брауна*, *Модель Брауна обобщённая*, *Адаптивные методы прогнозирования*.

ПАРАМЕТРЫ СДВИГА И МАСШТАБА

параметры плотности вероятности распределения.

С каждой случайной величиной X можно связать множество случайных величин Y, заданных формулой Y=aX+b при различных a>0 и b. Это множество называют масштабно-

сдвиговым семейством, порождённым случайной величиной X. Функции распределения FY(x) составляют масштабно сдвиговое семейство распределений, порожденное функцией распределения F(x). Вместо Y=aX+b часто используют запись:

$$Y = \frac{X - c}{b},$$

где

$$d = \frac{1}{a} > 0$$

и

$$c = -\frac{b}{a},$$

где c – параметр сдвига, число d – параметр масштаба. Формула показывает, что X – результат измерения некоторой величины – переходит в У – результат измерения той же величины, если начало измерения перенести в точку c, а затем использовать новую единицу измерения, в d раз большую старой.

Для масштабно-сдвигового семейства распределение X называют стандартным. В вероятностно-статистических методах принятия решений и других прикладных исследованиях используют стандартное нормальное распределение, стандартное распределение Вейбулла-Гнеденко, стандартное гамма-распределение и др.

ПЕРИОДОГРАММА

случайная функция, с помощью которой может быть получена оценка спектральной плотности (*анализ спектральный*) стационарного *временного ряда*. Периодограмма (2-го порядка) имеет вид:

$$I_T(\lambda) = \frac{1}{2\pi T}\left|\sum_{i=1}^{T} e^{-it\lambda} x_t\right|^2,$$

где $x_1, \ldots, x_T$ – члены стационарного временного ряда. При больших T существует равенство:

$$p(\lambda) = MI_t(\lambda) - \frac{1}{2\pi T}\frac{\sin^2\frac{\lambda T}{2}}{\sin\frac{\lambda}{2}} M(x_t) + 0\left(\frac{1}{T}\right). \quad (*)$$

В то же время сама по себе П. не является *оценкой состоятельной* спектральной плотности. Для получения состоятельной оценки $p(\lambda)$ следует рассматривать именно усреднение П. по близким к λ частотам.

Для использования равенства (*) необходимо располагать достаточно длинным временным рядом (T велико). Это ограничивает возможность приложений П., как и всего спектрального анализа временных рядов, к экономическим моделям. При наличии достаточно большого количества наблюдений соответствующий аппарат оказывается удобным.

ПЕРИОД УПРЕЖДЕНИЯ ПРОГНОЗА

отрезок времени от момента, для которого имеются последние статистические данные об изучаемом объекте, до момента, к которому относится прогноз. Традиционно используется следующая классификация экономических прогнозов в зависимости от их периода упреждения: оперативные (П.у.п. до 1 месяца); краткосрочные (П.у.п. от 1 месяца до года); среднесрочные (П.у.п. более 1 года, но не превышает 5 лет); долгосрочные (П.у.п. более 5 лет).

ПОЛИНОМИАЛЬНАЯ РЕГРЕССИЯ

регрессия вида $Y = f(X, \beta)$, где $f(X, \beta)$ – полином степени m от $X = (X^1, \ldots X^{(k)})$; (k – число регрессоров)

$$f(X, \beta) = \sum_{m_1 \ldots m_r} \beta_{m_1 \ldots m_r} \left[x^{(1)}\right]^{m_1} \ldots \left[x^{(k)}\right]^{m_k}.$$

П.р. является нелинейной по объясняющим переменным, но линейной по параметрам. При m =1 П.р. – обычная линейная регрессия.

Проблема выбора степени полинома (как и самого выбора П.р.) не имеет алгоритмического решения. В отсутствии теоретических предпосылок она решается на основании анализа экспериментальных данных. Наиболее простой и

часто используемый – общий приближённый критерий, основанный на группированных данных. Пусть f (X, β) – выбранный полином, b – полученные при этом *методом наименьших квадратов* (МНК) оценки параметров β (). Пусть данные сгруппированы в s гиперпараллелепипедов Γ_i (i=1,…, s), причем s>l, где l – число оцениваемых параметров. Определяется статистика

$$v^2 = \frac{(n-s)\sum_{i=1}^{s} y_i \left[\bar{y}_i - f\left(X_i^0, b\right)\right]^2}{(s-l)\sum_{i=1}^{s}\sum_{j=1}^{\nu_i}\left(y_{ij} - \bar{y}_i\right)^2},$$

где X_i^0 – середина Γ_i , ν_i – количество наблюдений в Γ_i , $y_{ij} = y(x_j)$ при $x_j \in \Gamma_i$, i –усреднение y_{ij} по j . Тогда при условии, что гипотеза о правильности выбора f(X , β) в качестве функции регрессии верна, статистика Z имеет F(s–l, n–s) распределение, и, следовательно, гипотеза может быть проверена с помощью *критерия Фишера*.

П.р. используется для различных типов данных в *регрессионном анализе.*

ПРЕДИКТОРЫ В МОДЕЛИ РЕГРЕССИИ

величины (регрессоры, *объясняющие переменные*), влияющие на значения *случайной величины* (объясняемой переменной).

Модель регрессии имеет вид: Y=f(X, β)+ε, где $X= (X_1,\ldots X_k)$ регрессоры, Y – объясняемая переменная, ε – случайная ошибка, f(X, β) – функция регрессии (чаще всего линейная), β – параметры. Предполагается, что Mε=0, т.е. f(X, β)=M(Y|X). П. в м.р. бывают как детерминированными, так и случайными (стохастические регрессоры).

ПРЕОБРАЗОВАНИЕ БОКСА-КОКСА

процедура подбора линеаризующего преобразования (*линеаризация*) нелинейной модели регрессии.

Пусть рассматривается нелинейная модель *Y* = *f*(*X*, β), причём регрессоры и объясняемая переменная принимают только положительные значения. Гипотеза Бокса-Кокса формулируется: существует вещественное (положительное или отрицательное) число λ такое, что после преобразования:

$$\tilde{y}_i(\lambda) = \frac{y_i^{\lambda} - 1}{\lambda}\ ,\ \tilde{x}_i^{(j)}(\lambda) = \frac{\left(x_i^{(j)}\right)^{\lambda} - 1}{\lambda},$$

i=1, …, n , j=1, …, k (k – число регрессоров) одна из двух моделей

$\tilde{Y}(\lambda) = \tilde{X}(\lambda)\beta + \varepsilon$ или $\tilde{Y}(\lambda) = X\beta + \varepsilon$ (*)

будет удовлетворять всем требованиям нормальной *модели классической линейной множественной регрессии.*

Параметр λ неизвестен; его оценка может быть получена *методом макс. правдоподобия*. При этом предполагается, что в модели (*) ошибки ε распределены нормально. Тогда функция правдоподобия имеет вид:

$$L\left(y_1 \ldots y_n \middle| X, \lambda, \beta, \sigma^2\right) = \frac{J(\lambda)}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2\sigma^2}\left(\tilde{Y} - X\tilde{\beta}\right)'\left(\tilde{Y} - X\tilde{\beta}\right)},$$

где

$$J(\lambda) = \left(\prod_{i=1}^{n} y_i\right)^{\lambda-1},\ \tilde{\beta} = \beta(\lambda) = (X'X)^{-1}X'\tilde{Y}(\lambda),\ \sigma^2(\lambda) = \frac{1}{n}\left(\tilde{Y}(\lambda) - X\tilde{\beta}(\lambda)\right)'\left(\tilde{Y}(\lambda) - X\tilde{\beta}(\lambda)\right).$$

Как правило, максимизация функции правдоподобия производится с помощью пошаговой аппроксимизационной процедуры.

ПРОГНОЗ ИНТЕРВАЛЬНЫЙ

значений объясняемой переменной Y – интервал значений $[\tilde{y}_{\min}(x_{n+1}),\ \tilde{y}_{\max}(x_{n+1})]$, который с выбранной *доверительной вероятностью* P накрывает неизвестное значение

$y(x_{n+1})$, где $x_{n+1} - x^1{}_{n+1},\ldots,x^k{}_{n+1}$, k – число регрессоров, включая константу.

П.и. строится на основе оптимального *прогноза точечного* $\tilde{y}_{n+1}$. Если рассматривается модель Y=Xβ+ε с нормально распределёнными остатками ε, то $\tilde{y}_{n+1} - y(X_{n+1}) \sim N(0, \sigma^2_{\text{прогн}})$;

$\sigma^2_{\text{прогн}} = C_0' \Omega C_0 + M(\varepsilon^2_{n+1}) - 2C_0 \sigma^{n+1}$, где Ω – матрица ковариаций ошибок регрессии $\sigma_{n+1} = M(\varepsilon\varepsilon_{n+1})$, а C_0' $Y = \tilde{Y}_{n+1}$ – оптимальный точечный прогноз.

Т.о., П.и. имеет вид:

$$(\hat{y}_{n+1} - u_{\frac{1+p}{2}}\sigma_{\text{прогн}},\ \hat{y}_{n+1} + u_{\frac{1+p}{2}}\sigma_{\text{прогн}}),$$

где u_a – *квантиль* порядка a стандартного нормального распределения.

Для классической модели П.и. принимает простой вид:

$$X_{n+1}b \pm u_{\frac{1+p}{2}}\sigma\sqrt{X'_{n+1}(X'X)^{-1}X_{n+1}+1},$$

где b – МНК-оценка (*метод наименьших квадратов*) параметра β классической модели.

ПРОГНОЗ ТОЧЕЧНЫЙ

объясняемой переменной в модели регрессии пространственной выборки или *временного ряда* – значение оценки y_{n+1} в том случае, если имеется n наблюдений объясняемой переменной и n +1 – объясняющих. Задача построения П.т. формулируется: имеется набор наблюдений x_{ij}, y_i (i=1,…n; j=1,…k; n – число наблюдений, k – число регрессоров). Имеются также значения объясняющих переменных

$(x_{n+1,1},\ldots,x_{n+1,k}) = X'_{n+1}$.

Используя оцененную модель регрессии, требуется построить наилучший (в смысле среднего квадрата ошибки) линейный относительно y_1, …, y_n и несмещённый прогноз для неизвестного значения y_{n+1}, т.е. подобрать коэффициенты c_1, …, c_n так, чтобы величина

$M(y_{n+1} - \sum c_k y_k)^2$

достигала бы минимума при условии

$M(y_{n+1} - \sum c_k y_k) = 0$.

Решение соответствующей оптимизационной задачи имеет вид:

$$C_0 = C_{opt} = \Omega^{-1}[E_n - X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}]\sigma^{n+1} + \Omega^{-1}X(X'\Omega^{-1}X)^{-1}X_{n+1},$$

где Ω – *матрица ковариаций* ошибок регрессии ε; $\sigma_{n+1} = M(\varepsilon\ \varepsilon_{n+1})$. Т.о., наилучшим прогнозом оказывается величина $\tilde{Y}_{n+1} = C_0 Y$.

Для классической модели $\sigma_{n+1}=0$ и наилучшим прогнозом оказывается модельное значение $\tilde{Y}_{n+1} = X_{n+1}b$, где b – МНК-оценка параметра β (*метод наименьших квадратов).*

ПРОГНОЗИРОВАНИЕ

процесс разработки прогнозов (от греч. prognosis – предвидение, предсказание). Под прогнозом понимается научно обоснованное описание возможных состояний объектов в будущем, а также альтернативных путей и сроков достижения этого состояния. Иногда в литературе выделяют поисковые прогнозы, отвечающие на вопрос о том, что вероятнее всего ожидать в будущем, а также нормативные прогнозы, связанные с ответом на вопрос о том, как нужно изменить условия, чтобы достичь заданного конечного состояния прогнозируемого объекта.

Т.о., под П. понимается научное выявление вероятных путей и результатов будущего развития явлений и процессов, основанное на системе установленных причинно-следственных связей и закономерностей, П. направлено на выявление и исследование альтернативных траекторий развития, зависящих от комплекса внутренних и внешних (относительно исследуемой системы) условий. Практический интерес представляет П. таких процессов, управление которыми в момент выработки прогноза либо невозможно, либо возможно частично или требует учёта влияния факторов, которое не может быть определено полностью. П. взаимосвязано с ретроспективным анализом, позволяющим исследовать тенденции и закономерности развития явлений, выявить причинно-следственные связи. Если результат развития процесса однозначен, то он не представляет интереса для прогностического исследования. Если же результат развития процесса имеет множество альтернатив, то прогнозные оценки

могут принести ощутимую пользу. Роль прогнозов в принятии управленческих решений – достаточно значима. Результаты П. позволяют получить сигнальную, предупреждающую информацию для органов управления, способствующую принятию научно обоснованных решений по устранению выявленных диспропорций, регулированию течения сложных процессов и явлений и др. Получение прогнозов может предшествовать процессу разработки планов, составляя для них основу. В то же время П. может использоваться для оценивания последствий принятых решений, проверки реалистичности разработанных планов, для оценивания хода выполнения намеченных планов. В зависимости от объектов П. принято классифицировать прогнозы на экономические, социальные, демографические, научно-технические и др. Однако такая классификация носит условный характер, так как между этими прогнозами существует множество прямых и обратных связей.

Прогнозы экономических явлений и процессов разрабатываются в виде качественных характеристик, включающих, напр., общее описание тенденции развития, предполагаемого характера изменений или просто утверждения о возможности наступления каких-либо событий, а также в виде количественных оценок будущего развития. Количественные оценки могут включать точечные и интервальные прогнозы, вероятностные оценки достижения этих значений. Экономические прогнозы разделяются в зависимости от масштабности объекта, так как они могут охватывать различные уровни: микроуровень (напр., прогнозы развития отдельных пр-тий), мезоуровень (напр., прогнозы развития регионов, корпораций), а также макроуровень и глобальный уровень.

В зависимости от *периода упреждения прогноза* выделяют оперативные, краткосрочные, среднесрочные, долгосрочные экономические прогнозы. К осн. факторам, приводящим к существенным ошибкам полученных прогнозных оценок относятся: создание прогнозной модели на основе неверных теоретических положений, недостатки исходных данных, неверное оценивание имеющихся параметров выбранной модели, отсутствие подходящих эмпирических критериев проверки, скачкообразные изменения исследуемого процесса и др. Правильность исходных теоретических предпосылок, методологической основы определяет результаты П., их «успешность». С развитием математического аппарата П., с повсеместной распространённостью компьютеров, оснащённых соответствующим программным обеспечением, совершенствованием информационных технологий П. уделяется все больше внимания. При этом в П. не может быть чисто формальных подходов, так как получение количественных оценок должно опираться на качественный анализ, на глубокое знание объекта исследования, содержательный анализ изучаемых явлений.

ПРОВЕРКА ВРЕМЕННОГО РЯДА НА СЛУЧАЙНОСТЬ КОЛЕБАНИЙ

проверка гипотезы об отсутствии неслучайной составляющей (о постоянстве среднего значения, включая утверждение о взаимной стохастической независимости $x_1, \ldots, x_T$ (*статистическая независимость случайных величин*).

Наиболее простой критерий проверки гипотезы – тест серий, основанный на *медиане*. Члены ряда располагаются в порядке возрастания (в виде *вариационного ряда*). Выборочная медиана определяется в виде:

$$x_{med} = \begin{cases} x_{\frac{T+1}{2}}, & \text{если } T \text{ нечетно,} \\ \frac{1}{2}\left(x_{\frac{T}{2}} + x_{\frac{T}{2}+1} \right), & \text{если } T \text{ четно.} \end{cases}$$

Каждому члену x_t вариационного ряда приписывается знак +, если $x_t > x_{med}$ и знак – , если $x_t < x_{med}$ (члены, равные x_{med}, не учитываются). Серией называется последовательность знаков одного типа (+ или –). Пусть $\nu(n)$ – общее число серий, $r(n)$ – длина наиболее протяженной серии. Гипотеза отвергается, если $\nu(n)$ недостаточно велико, а $r(n)$ наоборот велико. Чаще всего в качестве критерия выбираются неравенства:

$$v(n) > \frac{1}{2}\left(n + 2 - 1{,}96\sqrt{n-1}\right);$$

$$r(n) < 1{,}43\ln(n+1).$$

Если хотя бы одно из этих неравенств оказывается нарушенным, гипотеза отвергается на уровне значимости α, причём 0,05<α<0,1.

Используются также критерии «восходящих» и «нисходящих» серий, а также критерий квадратов последовательных разностей (критерий Аббе).

Р

РЕАЛИЗАЦИЯ ВРЕМЕННОГО РЯДА

статистическая совокупность наблюдений *случайной величины*, которые хронологически упорядочены во времени**.** *Временным* (динамическим) *рядом* называется выборка наблюдений, в которой важны не только сами наблюдаемые значения случайных величин, но и порядок их следования друг за другом. На практике упорядоченность обусловлена тем, что исходные данные для моделирования представляют собой серию наблюдений одной и той же случайной величины в последовательные моменты времени. При этом предполагается, что тип распределения наблюдаемой случайной величины постоянен (напр., распределён по нормальному закону), но параметры распределения изменяются в зависимости от времени. Т.о., выборка наблюдаемых значений $y_1, y_2, \ldots, y_n$ рассматривается как одна из реализаций случайной величины *Y*, то временной ряд $y_1, y_2, \ldots, y_n$ рассматривается как одна из реализаций случайного процесса *Y(t)* (Р.в.р.), где t – фактор времени. Следует отметить принципиальные отличия временного ряда y_t ($t=1,2, \ldots, n$) от последовательности наблюдений $y_1, y_2, \ldots, y_n$. Модели временных рядов, как правило, оказываются сложнее моделей пространственной выборки. Во-первых, в отличие от элементов случайной выборки члены временного ряда в большинстве случаев не являются статистически независимыми. Во-вторых, члены временного ряда не являются одинаково распределёнными.

С

СКОЛЬЗЯЩЕГО СРЕДНЕГО МЕТОД

алгоритмический метод выделения неслучайной составляющей *временного ряда*. В основе метода сглаживания (элиминирования) случайных флуктуаций в поведении анализируемого временного ряда лежит следующая идея. Если «индивидуальный» разброс значений временного ряда y(t) около своего среднего (сглаженного) значения a характеризуется дисперсией σ2 , то разброс среднего из N уровней временного ряда $\frac{y(1)+y(2)+\ldots+y(N)}{N}$ около того же значения a будет характеризоваться гораздо меньшей величиной дисперсии, равной $\frac{\sigma^2}{N}$. А уменьшение меры случайного разброса и означает как раз сглаживание соответствующей траектории временного ряда. Суть метода – замена фактических уровней временного ряда расчетными уровнями, которые в меньшей степени подвержены колебаниям. Это способствует более четкому проявлению тенденции развития явления. Скользящие средние позволяют сгладить как случайные, так и периодические колебания, выявить имеющуюся тенденцию в развитии процесса. Процедуры скользящих средних основываются на известной теореме Вейерштрасса, согласно которой любая гладкая функция при самых общих допущениях может быть локально, т.е. в ограниченном интервале изменения её аргумента t, представлена алгебраическим полиномом подходящей степени. На практике для сглаживания фактических значений выбирают некоторую нечётную «длину усреднения» N = 2m + 1, измеренную в числе подряд идущих уровней анализируемого временного ряда. Конкретный выбор m зависит от специфики исходных данных. Как правило, выбирают m < n/3, где n – число уровней временного ряда. Затем сглаженные значения $\bar{y}(t)$ временного ряда y(t) вычисляют по значениям x(t-m), x(t-m + 1), …, x(t), x(t + 1), …, x(t + m) по формуле

$$\bar{y}(t) = \sum_{k=-m}^{m} w_k \cdot y(t+k),$$

где t = m+1, m+2, …, n-m; wk (k=–m, –m+1, …, m) – некоторые положительные «весовые» ко-

эффициенты («веса»). В сумме все веса дают 1. Поскольку, изменяя t от m+1 до n–m, мы как бы «скользим» по оси времени, то метод и получил название «С.с.м.». Алгоритмы С.с.м. отличают в зависимости от выбора параметра m и весов. При линейном характере локальной аппроксимации траектории временного ряда в качестве его сглаженного значения в точке t используют простую скользящую среднюю. Т.о., выравнивание (сглаживание) происходит по полиному первого порядка, когда графическое изображения ряда динамики напоминает прямую. Если для анализируемого временного ряда характерно нелинейное развитие, то используют взвешенную скользящую среднюю. Простая скользящая средняя учитывает все уровни ряда, входящие в активный участок сглаживания, с равными весами wk , а взвешенная средняя приписывает каждому уровню вес, зависящий от удаления данного уровня до уровня, стоящего в середине активного участка сглаживания. Для устранения сезонных колебаний на практике используют и чётные скользящие средние с длиной интервала сглаживания, равной 4 или 12.

СПЛАЙН (СПЛАЙН-ФУНКЦИЯ)

(от англ. spline – планка, рейка) – агрегатная функция, совпадающая с функциями более простой природы на каждом элементе разбиения своей области определения. Полиномиальным С. степени n называется составленная из кусков многочленов степени не выше n кусочно-полиномиальная функция, такая, что она и все её производные до порядка *n-1* включительно непрерывны.

Сплайн-функция – кусочная функция, отдельные куски которой соединены друг с другом гладким образом. В качестве таких кусков обычно выбираются полиномы различных степеней, а условие гладкости формулируется в терминах непрерывности самого С. и его производных. Макс. степень использованных полиномов называется степенью С. Напр., непрерывная ломанная есть С. 1-й степени. На практике используют линейные, кубические, билинейные С. Функции, подобные тем, что сейчас называют С., были известны математикам давно, начиная как минимум с Эйлера, но их интенсивное изучение началось, фактически, только в середине 20 в. Исаак Шенберг в 1946 впервые употребил этот термин в качестве обозначения класса полиномиальных С. До 1960-х гг. С. были в основном инструментом теоретических исследований, они часто появлялись в качестве решений различных экстремальных и вариационных задач, особенно в теории приближений. С. имеют многочисленные применения, как в математической теории, так и в разнообразных вычислительных приложениях. В частности, С. двух переменных интенсивно используются для задания поверхностей в различных системах компьютерного моделирования. Сплайн-функции широко применяются при построении моделей структурных изменений. Напр., часто в эконометрическом моделировании приходится сталкиваться с ситуацией, когда эндогенная переменная *y* рассматривается как линейная функция от *x* над некоторой частью области определения *x*, но такая линейность не может иметь места на всей области определения величины *x*. Такую нелинейную зависимость можно аппроксимировать некоторым полиномом. Другой подход – представить *y* в виде кусочно-линейной функции от *x*. Такой подход при условии непрерывности функции приводит к применению линейных сплайн-функций (см. рис.1). Поскольку эндогенная переменная y есть кусочно-линейная функция от x, то точки, в которых линейная функция меняет свой вид, считаются точками структурных изменений (узлами). Одно из преимуществ такого подхода к построению моделей состоит в том, что внутри каждого отрезка зависимость между y и x имеет очень простой вид, и, следовательно, без труда поддается анализу. Для моделирования тенденции изменения явления в рамках куска функции могут применяться полиномы и более высоких степеней, напр., третьей (см. рис. 2).

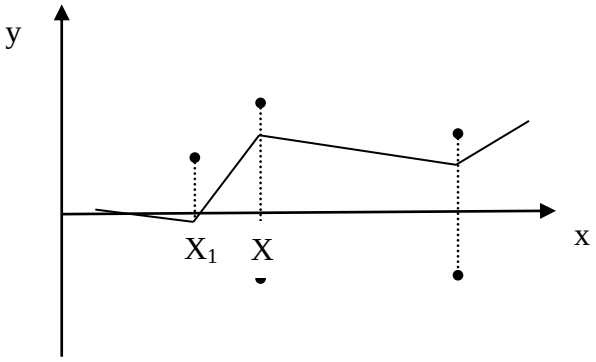


Рис. 1. **Линейный сплайн с тремя узлами**

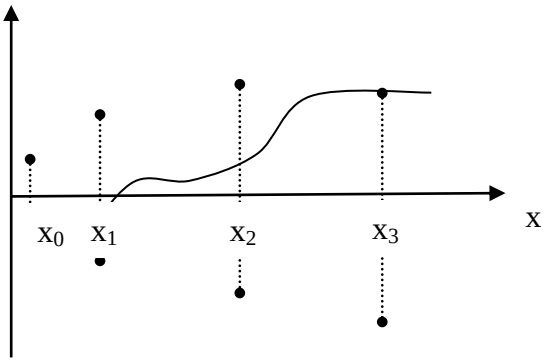


Рис. 2. **Кубический сплайн с четырьмя узлами**

СТАТИСТИКА БОКСА-ПИРСА

статистика, используемая для анализа *автокорреляции* в остатках при диагностической проверке моделей. У адекватной модели остатки должны быть похожи на *белый шум*, т.е. их выборочные автокорреляции не должны существенно отличаться от нуля. С.Б.-П. применяется для проверки значимости множества коэффициентов автокорреляции как группы, а не для проверки значимости каждого коэффициента автокорреляции отдельно. Проверка опирается на Q – С.Б.-П., позволяющую проверить равенство нулю сразу τ первых значений *автокорреляционной функции* остатков. Q – С.Б.-П. определяется как

$$Q = n\sum_{k=1}^{\tau} r_k^2 ,$$

где n – длина временного ряда остатков; r_k – выборочный коэффициент автокорреляции при лаге k (k=1,2,…, τ).

При *нулевой гипотезе* об отсутствии автокорреляции статистика Q имеет асимптотическое распределение χ^2_τ. Если полученное значение Q больше соответствующего критического значения $\chi^2_{кр}$, то нулевая гипотеза отвергается. При проверке нулевой гипотезы об отсутствии автокорреляции в остатках модели АРСС(p, q) (*модели авторегрессии со скользящими средними в остатках*), или в англоязычном варианте *ARMA*(p, q), рассматривается распределение χ^2 с $\nu = \tau - p - q$ степенями свободы, где p и q определяют соответственно порядок авторегрессионной составляющей и порядок скользящих средних модели. В некоторые современные эконометрические пакеты включена модификация этого подхода, опирающаяся на *статистику Льюнга-Бокса*.

СТАТИСТИКА ЛЬЮНГА-БОКСА

модификация *статистики Бокса-Пирса*, используемая для анализа *автокорреляции* в остатках при диагностической проверке моделей. Соответствующая статистика, применяемая для проверки значимости множества τ первых коэффициентов автокорреляции как группы, определяется выражением:

$$\tilde{Q} = n(n+2)\sum_{k=1}^{\tau} \frac{r_k^2}{n-k},$$

где n – длина *временного ряда* остатков; r_k – выборочный коэффициент автокорреляции при лаге k (k=1,2,…, τ).

С.Л.-Б. ($\tilde{Q}$), также как и статистика Бокса-Пирса Q, имеет асимптотическое распределение χ^2_τ. Предложенная Льюнгом и Боксом модифицированная статистика ($\tilde{Q}$) придаёт меньший вес «далёким» коэффициентам автокорреляции (с большим *лагом*), при этом её распределение ближе к χ^2 для конечных выборок. Это делает более предпочтительным на практике подход, связанный с применением модифицированной статистики. Проверка гипотезы об отсутствии автокорреляции (о равенстве нулю τ первых значений автокорреляционной функции остатков) аналогична процедуре, реализуемой при использовании статистики Бокса-Пирса.

СТАТИСТИЧЕСКИЕ МЕТОДЫ ПРОГНОЗИРОВАНИЯ

статистические методы, используемые в процессе разработки прогнозов. Применение С.м.п. предполагает наличие определённой степени инерционности в прогнозируемых социально-экономических процессах. Инерционность проявляется как в сохранении в осн. чертах зависимостей прогнозируемой переменной от одного или нескольких факторных признаков, так и в сохранении в некоторой степени характера динамики. Сначала исследователь на основе обобщения имеющейся информации и представлений о существующих закономерностях определяет вид экономико-статистической модели, а затем на основе имеющихся наблюдений оценивает коэффициенты (параметры) модели, используемой в дальнейшем для построения прогнозов. При этом сначала в базовый набор может быть включено несколько моделей, а в дальнейшем на основе анализа их статистических характеристик осуществлен выбор «лучшей» модели. Т.о., прогнозирование социально-экономических процессов с помощью С.м.п., как правило, включает следующие этапы: постановка задачи и сбор необходимой информации; первичная обработка исходных данных; определение базового набора возможных моделей прогнозирования; оценивание параметров моделей на основе имеющихся наблюдений; исследование «качества» полученных моделей (оценивание точности моделей, их адекватности реальному процессу) и окончательный выбор модели; построение прогноза (точечного, интервального); содержательный анализ полученных результатов. Многие С.м.п. одномерных *временных рядов* опираются на представление уровней ряда в виде сочетания трендовой, сезонной, циклической, случайной составляющих (очевидно, что отдельные из перечисленных составляющих могут отсутствовать). При моделировании и прогнозировании тенденции временного ряда широко используются модели кривых роста (линейная, параболическая, экспоненциальная, логарифмическая и др. модели), часто в сочетании со скользящими средними. При прогнозировании тренд-сезонных процессов используются трендовые модели в сочетании с индексами сезонности, с фиктивными переменными, с гармоническим анализом, адаптивные модели, основанные на *экспоненциальном сглаживании*, модель ARIMA и др. Также при прогнозировании социально-экономических явлений и процессов широко используются регрессионные и другие эконометрические модели и подходы. К важным факторам, послужившим импульсом к стремительному развитию этих методов, можно отнести совершенствование компьютерной техники и информационных технологий, распространение специализированных пакетов прикладных программ, развитие финансовых рынков и инструментов и др. Однако нельзя забывать, что получаемые результаты, прогнозы опираются на предположения о выполнении условий, гипотез, учтённых при разработке соответствующих моделей.

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ СЛУЧАЙНЫХ ПРОЦЕССОВ

вид компьютерного моделирования, позволяющий получать последовательности выборочных значений случайной величины с заданным законом распределением. Метод *статистического моделирования* – особый численный метод, который как бы имитирует элементарные явления, составляющие исследуемый процесс. С.м.с.п. представляет особый интерес для моделирования экономических систем, т.к. позволяет учитывать влияние факторов, воздействие которых на экономические системы носит случайный характер. Учёт действия случайных факторов приводит к необходимости отыскания вероятностных характеристик случайных величин, законы распределения которых, как правило, не известны. На практике наибольшее распространение получил метод моделирования случайных процессов, называемый *методом Монте-Карло.*

Т

ТЕСТ БРЕУША-ПАГАНА НА ГЕТЕРОСКЕДАСТИЧНОСТЬ ОСТАТКОВ

применяется в тех случаях, когда априорно предполагается, что дисперсия остатков регрессионной модели зависит от значений k регрессоров, и эта зависимость описывается аналитически как функция с небольшим числом параметров $\sigma_i^2 = \sigma^2 h\left(\tilde{x}_i^T \nu\right)$, где $\tilde{x}_i$ – вектор значений регрессоров для i-го наблюдения, оказывающих влияние на дисперсию остатков, ν-вектор параметров. После построения стандартной регрессии и получения вектора остатков проводится оценка вектора параметров ν. Для выявления *гетероскедастичности* предполагаемого вида проверяется гипотеза об её отсутствии, т.е. о равенстве всех компонентов вектора коэффициентов ν нулю при альтернативе в виде неравенства нулю хотя бы одного из них. В случае отклонения гипотезы на выбранном уровне значимости α утверждается факт наличия гетероскедастичности с вероятностью ошибки α. Самый простой вариант данного теста предполагает расчёт коэффициента детерминации вспомогательной регрессионной модели $\hat{\sigma}_i^2 = \sigma^2 h\left(\tilde{x}_i^T \nu\right)$ и его умножение на число наблюдений. Статистика, получаемая при отсутствии гетероскедастичности, имеет *распределение Пирсона* χ^2 с k степенями свободы. В случае выявления гетероскедастичности возможно использование результатов моделирования зависимости остатков от значений регрессоров для её устранения и перехода к *модели классической линейной множественной регрессии*.

ТЕСТЫ ДИКИ–ФУЛЛЕРА

(от англ. – Dickey-Fuller test**)** – тесты, используемые при анализе стационарности процессов и часто описываемые как процедуры проверки гипотез о наличии единичных корней (от англ. – unit root – единичный корень). Авторами сначала был рассмотрн простейший вариант линейного авторегрессионного процесса, описываемый *моделью авторегрессии* 1-го порядка – AR(1): $y_t = \alpha y_{t-1} + \varepsilon_t$, где α – числовой коэффициент, ε_t – последовательность случайных величин, образующих *белый шум*. Условие стационарности для AR (1) определяется требованием $|\alpha| < 1$ или, что тоже самое, корень Z_0 уравнения $1 - \alpha z = 0$, являющегося частным случаем характеристического уравнения для общего линейного процесса авторегрессии, должен быть по абсолютной величине больше 1. На процесс AR(1) внешне похож процесс «случайного блуждания» (random walk): $y_t = y_{t-1} + \varepsilon_t$. Однако свойства этого процесса существенно отличаются от стационарного процесса AR (1) (при $|\alpha| < 1$). Случайное блуждание нестационарно. Процесс с $|\alpha| > 1$ тем более является нестационарным, причём подразумевает взрывные ряды, что маловероятно в реальных финансово-экономических задачах.

Так как наличие единичного корня у характеристического уравнения существенно влияет на свойства процесса, то большое практическое значение приобретает вопрос тестирования единичного корня по имеющимся наблюдениям. Используя процедуру дифференцирования (взятие первой разности), можно привести уравнение $y_t = \alpha y_{t-1} + \varepsilon_t$ к виду: $\Delta y_t = \beta y_{t-1} + \varepsilon_t$, где $\Delta y_t = y_t - y_{t-1}$; $\beta = \alpha - 1$. При этом не рекомендуется использовать традиционный *t*-критерий Стьюдента для проверки значимости β. Использование стандартной процедуры проверки, опирающейся на *t*-статистику, приведет к тому, что *нулевая гипотеза* существования единичного корня будет отвергаться (ошибочно) слишком часто. В этом случае рекомендуется использовать распределение t- статистики, описанные Дики и Фуллером, которые рассмотрели модели модификации: 1. $y_t = \alpha\, y_{t-1} + \varepsilon_t$ (модель без константы); 2. $y_t = \alpha_0 + \alpha_1 y_{t-1} + \varepsilon_t$ (модель с константой); 3. $y_t = \alpha_0 + \alpha_2 t + \alpha_1 y_{t-1} + \varepsilon_t$ (модель с константой, с детерминированным трендом).

Все три модели соответственно могут быть преобразованы:

1. $\Delta y_t = \beta\, y_{t-1} + \varepsilon_t$;

2. $\Delta y_t = \alpha_0 + \beta y_{t-1} + \varepsilon_t$;

3. $\Delta y_t = \alpha_0 + \alpha_2 t + \beta\, y_{t-1} + \varepsilon_t$.

Нулевая гипотеза H_0: $\beta = 0$ (при альтернативной H_1 : $\beta < 0$) будет отвергнута, если наблюдаемое значение критерия $t_н$ меньше критического значения $t_{кр}$, взятого из табл. Дики и Фуллера. Значение $t_н$ получается делением оценки коэффициента β на её стандартную ошибку после применения *метода наименьших квадратов*. Значение $t_{кр}$ зависит от фиксированного уровня значимости, размера выборки и вида модели (1, 2, 3). Проблемы при проверке стационарности возникают при наличии *автокорреляции* остатков. Авторами был предложен расширенный критерий (Augmented Dickey – Fuller test, ADF-тест). Он является обобщением обычного DF – теста для моделей, в правую часть которых добавляются в виде слагаемых лаговые значения приращений из левой части. В этом случае также могут быть рассмотрены модификации моделей с добавлением константы и линейного тренда. Напр., модель с константой может быть записана в виде: $\Delta y_t = \alpha_0 + \beta y_{t-1} + \varphi_1 \Delta y_{t-1} + \varphi_2 \Delta y_{t-2} + \ldots + \varphi_p \Delta y_{t-p} + \varepsilon_t$.

Процедура тестирования аналогична предыдущим. Критические значения статистики для ADF-теста остаются такими же, как для обычного DF-теста. Применение ADF-теста позволяет осуществлять проверку единичного корня в AR-моделях более высокого порядка. Добавление приращений в модель производится для того, чтобы устранить возможную автокорреляцию ошибок, так как критические значения статистики Дики-Фуллера справедливы только лишь в случае, если ошибки являются *белым шумом*. Впоследствии Т.Д.-Ф. получили дальнейшее развитие и обобщение; они включены во многие современные эконометрические пакеты.

ТОЧНОСТЬ ПРОГНОЗА

оценка доверительного интервала прогноза для заданной вероятности его осуществления. Эмпирической мерой Т.п. служит величина его ошибки, которая определяется как разность между прогнозными ($\hat{y}_t^*$) и фактическими (y_t) значениями исследуемого показателя. Показатели точности статистических прогнозов условно можно разделить на три группы: аналитические, сравнительные и качественные.

Аналитические показатели Т.п. позволяют количественно определить величину ошибки прогноза: абсолютная ошибка прогноза (Δ^*) определяется как разность между эмпирическими и прогнозными значениями признака и вычисляется по формуле: $\Delta^* = y_t - \hat{y}_t^*$, где: y_t – фактическое значение признака; $\hat{y}_t^*$ – прогнозное значение признака; относительная ошибка прогноза ($d^*_{отн}$) определяется как отношение абсолютной ошибки прогноза (Δ^*) к фактическому значению признака (y_t):

$$d^*_{отн} = \frac{\Delta^*}{y_t} \cdot 100\% = \frac{y_t - \hat{y}_t^*}{y_t} \cdot 100\%,$$

либо к прогнозному значению признака ($\hat{y}_t^*$) – тогда в знаменателе указывают ($\hat{y}_t^*$). Абсолютная и относительная ошибки прогноза являются оценкой точности единичного прогноза. Это снижает их значимость в оценке точности всей модели.

Средним показателем Т.п. является средняя абсолютная ошибка прогноза ($\overline{\Delta^*}$), которая определяется как средняя арифметическая простая из абсолютных ошибок прогноза:

$$\overline{\Delta^*} = \frac{\sum_{t=1}^{n} \left|\Delta^*\right|}{n} = \frac{\sum_{t=1}^{n} \left|y_t - \hat{y}_t^*\right|}{n},$$

где n – длина временного ряда. Средняя абсолютная ошибка прогноза показывает обобщённую характеристику степени отклонения фактических и прогнозных значений признака и имеет ту же размерность, что и размерность изучаемого признака.

Средняя квадратическая ошибка прогноза также используется для оценки Т.п.:

$$\sigma_{ош} = \sqrt{\frac{\sum_{t=1}^{n} \left(y_t - \hat{y}_t^*\right)^2}{n}}.$$

Размерность средней квадратической ошибки прогноза также соответствует размерности изучаемого признака. Между средней абсолютной и средней квадратической ошибками прогноза существует следующее примерное соотношение: $\sigma_{ош} = 1{,}25 \cdot \overline{\Delta^*}$. Недостатками средней абсолютной и средней квадратической ошибок про-

гноза является их существенная зависимость от масштаба измерения уровней изучаемых социально-экономических явлений. Поэтому на практике в качестве характеристики Т.п. определяют среднюю относительную ошибку аппроксимации, которая выражается в процентах:

$$\bar{\varepsilon}^* = \frac{1}{n} \cdot \sum_{t=1}^{n} \frac{\left| y_t - \hat{y}_t^* \right|}{y_t} \cdot 100\% .$$

Данный показатель является относительным показателем Т.п. и не отражает размерность изучаемых признаков, выражается в процентах и на практике используется для сравнения Т.п. полученных по различным моделям и объектам. Интерпретация Т.п.: менее 10% – высокая, 10–20% – хорошая, 20–50% – удовлетворительная, более 50% – не удовлетворительная.

Также для оценки точности используются коэффициент корреляции между прогнозными и фактическими значениями признака и коэффициенты несоответствия.

ТРЕНД

изменение, определяющее общее направление развития *временного ряда* или изменение его «среднего» уровня. Для моделирования Т., для описания осн. тенденции временного ряда используют различные детерминированные функции времени, коэффициенты которых оценивают по выборочным данным. Примером таких функций служат полиномиальные, экспоненциальные, логарифмические модели и др., коэффициенты которых оцениваются с помощью *метода наименьших квадратов.* Широко на практике для выявления Т. применяются такие процедуры сглаживания (выравнивания) временных рядов, как скользящие средние, *экспоненциальное сглаживание.* В совр. статистической литературе различают детерминированный и стохастический Т.

Уровни временного ряда можно представить в виде суммы: $y_t = \varphi_t + \varepsilon_t$, где ε_t – случайные отклонения (колебания). Т., характеризующий закон изменения во времени слагаемого φ_t, может быть представлен детерминированной функцией (детерминированный Т.), случайной функцией (стохастический Т.) или их комбинацией. Примером стохастического Т. служит функция:

$$\varphi_t = \varphi_{t-1} + u_t = \varphi_0 + \sum_{i=1}^{t} u_i ,$$

где φ_0 – некоторое начальное значение, u_t – случайная независимая переменная с нулевым средним.

ТРЕНД ЛИНЕЙНЫЙ

тренд вида $\hat{y}_t = a_0 + a_1 t$, где t – время, a_0, a_1 – параметры, определяемые *методом наименьших квадратов*. Этот вид тренда получил широкое распространение в практических приложениях, в частности в экономических исследованиях. Параметр a_1 определяет средний абсолютный прирост для анализируемого *временного ряда*. Напр., по данным о динамике численности занятых в экономике РФ с 1990 по 1996 получено уравнение Т.л. $\hat{y}_t = 77{,}0 - 1{,}6t$. Согласно этой модели в исследуемом периоде численность занятых в экономике РФ в среднем ежегодно сокращалась на 1,6 млн чел.

ТРЕНД СТЕПЕННОЙ

тренд, определяемый зависимостью вида

$$\hat{y}_t = a_0 t^{a_1} \qquad (1),$$

где t – время. Оценки коэффициентов модели a_0, a_1 могут быть определены после логарифмирования с помощью линейной регрессии: $\ln y_t = \ln a_0 + a_1 \ln t$ (2). Параметр a_0 определяется потенцированием после нахождения коэффициентов модели (2). При $a_1 = 1$ модель (1) переходит в линейную модель, при $a_1 = 2$ – в параболическую.

Наиболее часто в социально-экономических исследованиях встречается Т.с. при $0 < a_1 < 1$, $a_0 > 0$ (см. рис.1).

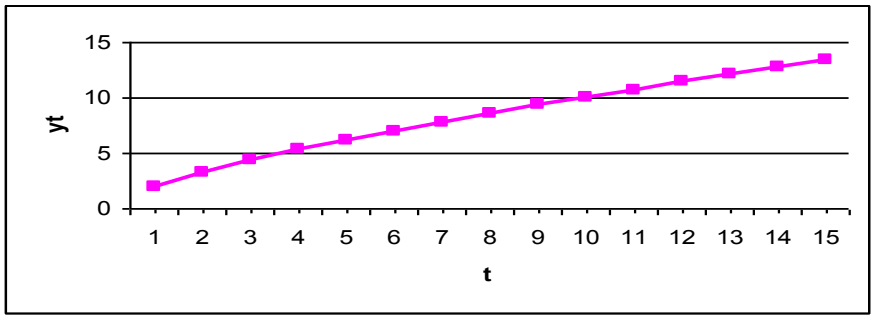


Рис. 1. **Тренд в виде степенной функции**

ТРЕНД ЭКСПОНЕНЦИАЛЬНЫЙ

тренд, определяемый зависимостью вида $\hat{y}_t = a_0 e^{a_1 t}$ (1), где t – время. Оценки коэффициентов модели a_0, a_1 могут быть определены *методом наименьших квадратов* после *линеаризации* с помощью логарифмирования: $\ln y_t = \ln a_0 + a_1 t$ (2). Параметр a_0 определяется потенцированием после нахождения коэффициентов модели (2). Величина e^{a_1} характеризует средний коэффициент роста уровней *временного ряда.*

Экспоненциальная трендовая модель может рассматриваться как частный случай показательной модели $\hat{y}_t = a_0 b^{a_1 t}$.

Э

ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ

распространённый приём выравнивания *временных рядов*, позволяющий придать больший вес более поздним наблюдениям и учесть их большую информационную ценность. Осуществляется с помощью расчёта экспоненциально-взвешенных скользящих средних, или, кратко, экспоненциальных средних. Модель Э.с. ряда описывается следующей рекуррентной формулой: $S_t = \alpha y_t + \beta S_{t-1}$ (1), где S_t – значение экспоненциальной средней в момент t; y_t – текущий уровень временного ряда; α – параметр сглаживания, α=const, 0<α<1; β = 1- α.

Последовательно используя соотношение (1) можно представить экспоненциальную среднюю S_t в виде:

$$S_t = \alpha \sum_{i=0}^{n-1} \beta^i y_{t-i} + \beta^n S_0 \quad (2),$$

где n – длина временного ряда; S_0 – начальное значение экспоненциальной средней.

На практике в качестве начального значения S_0 используется *среднее арифметическое значение* из всех имеющихся уровней временного ряда или из какой-то их части. Очевидно, что при $n \to \infty$ $\beta^n \to 0$, следовательно,

$$S_t = \alpha \sum_{i=0}^{\infty} \beta^i y_{t-i}. \qquad (3)$$

Из (2–3) видно, что величина S_t оказывается взвешенной суммой всех членов ряда, причём веса отдельных уровней ряда экспоненциально убывают по мере их удаления в прошлое. Пусть уровни временного ряда представлены в виде: $y_t = a_1 + \varepsilon_t$, где a_1 = const; ε_t – случайные неавтокоррелированные отклонения с нулевым математическим ожиданием и дисперсией σ^2. Тогда, как показал Р. Браун, *математические ожидания* временного ряда и экспоненциальной средней совпадут, но в то же время дисперсия экспоненциальной средней $D(S_t)$ будет меньше, так как

$$D(S_t) = \frac{\alpha}{2-\alpha}\sigma^2.$$

С уменьшением α дисперсия экспоненциальной средней сокращается, возрастает её отличие от

дисперсии ряда. Т.о., экспоненциальная средняя выполняет роль «фильтра», поглощающего колебания временного ряда. Обобщение процедуры Э.с. привело к появлению целого класса моделей, называемых адаптивными. Адаптивные модели прогнозирования, опирающиеся на процедуру Э.с., широко представлены в современных статистических пакетах, напр., модели линейного роста Ч. Хольта и Р. Брауна, трендсезонные модели Хольта-Уинтерса, Тейла-Вейджа и др.

ЭКСТРАПОЛЯЦИЯ

нахождение значений функции за пределами её области определения с использованием информации о поведении данной функции в точках, принадлежащих области её определения. Э. широко применяется при *прогнозировании временных рядов*. Напр., на Э. построено прогнозирование тенденции изменения показателя для периода упреждения с помощью т.н. трендовых моделей кривых роста (линейной, параболической, экспоненциальной модели и др.). В этом случае сначала по наблюдавшимся уровням временного ряда определяются коэффициенты трендовых моделей, затем с помощью статистических характеристик оценивается «качество» построенных моделей для периода наблюдения (на ретроспективном участке) и определяется окончательная модель для расчёта прогнозных значений. Прогнозирование на основе трендовых моделей кривых роста базируется на Э., продлении в будущее тенденции, наблюдавшейся в прошлом. При этом предполагается, что характер развития показателя обладает свойством инерционности, сложившаяся тенденция не должна претерпевать существенных изменений в течение периода упреждения.

Наряду с Э. используется понятие интерполяции (интерполирование) – нахождение промежуточных значений функции в области её определения. Интерполяция и Э. находят практическое применение в *регрессионном анализе*, при решении широкого спектра задач прогнозирования.

ЭРГОДИЧЕСКОЕ СВОЙСТВО

возможность определения статистических характеристик случайного стационарного процесса (математического ожидания, дисперсии, корреляционной функции) по единственной реализации (временному ряду).

Пусть исследуется случайный стационарный процесс X, и требуется оценить корреляционную связь значений процесса в момент t_1 и t_2, т.е. коэффициент корреляции r_{t_1,t_2}. Тогда в общем случае нужно получить множество реализаций процесса для моментов t_1 и t_2 и вычислить этот коэффициент корреляции. Но в экономических исследованиях это обычно невозможно: как правило, всегда имеется только одна единственная реализация в виде временного ряда и ничего повторить нельзя. Тогда, если процесс стационарный, математическое ожидание $M(X_{t_1}) = M(X_{t_2}) = const$, дисперсия $D(X_{t_1}) = D(X_{t_2}) = const$ и коэффициент корреляции зависит только от разности $\tau = t_2 - t_1$, т.е. $r_{t_1,t_2} = r(\tau)$ для любых моментов времени t_1 и t_2, то говорят, что процесс обладает свойством эргодичности, и его статистические характеристики можно оценить по единственной реализации процесса X в виде временного ряда, представленного выборкой достаточно большого объёма n. Математическое ожидание процесса X приближенно оценивается как среднее по времени

$$\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t ,$$

дисперсия – как средний квадрат отклонений членов ряда от среднего на выборочном периоде

$$D(X) = s_x^2 = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \bar{X})^2$$

(несмещённая оценка), а коэффициент корреляции как

$$r(\tau) = \frac{1}{n-\tau}\frac{\sum_{t=\tau}^{n}(X_t - \bar{X})(X_{t-\tau} - \bar{X})}{s_x^2} .$$

Не всякий стационарный процесс обладает Э.с. Условия эргодичности, напр., не выполняются, если ряд гармоничен. Об эргодичности или не-

эргодичности случайного процесса может непосредственно свидетельствовать вид его корреляционной функции. Стремление корреляционной функции к нулю при $\tau \to \infty$ говорит в пользу эргодичности процесса. Во всяком случае его достаточно для того, чтобы математическое ожидание процесса можно было находить как среднее по времени. Если корреляционная функция стационарного случайного процесса при увеличении τ не убывает, а, начиная с некоторого значения аргумента τ, остается приблизительно постоянной – это обычно есть признак того, что процесс не является эргодическим.

Экономические показатели часто не только не являются эргодическими, но оказываются и нестационарными. Т.о., и математическое ожидание, и дисперсия, и сила корреляционной связи не остаются постоянными во времени, а выборка всего лишь одна. Тогда изменчивость этих статистических характеристик можно оценить с помощью адаптивных подходов.

См. также *Адаптивные методы прогнозирования*, *Параметр адаптации*.

ЭФФЕКТ СЛУЦКОГО-ЮЛА

назван по имени учёных, впервые обративших внимание на тот факт, что *временные ряды*, полученные после применения процедур скользящих средних, могут содержать систематические (периодические) колебания, вызванные лишь усреднением случайных составляющих. Временной ряд, полученный после применения процедуры скользящих средних, также содержит случайную составляющую $\tilde{\varepsilon}_t$, однако её влияние будет сглажено и поэтому выражено не так явно как для исходного ряда. Рассмотрим аддитивную модель исходного временного ряда, в которой случайная компонента ε_t удовлетворяет следующим свойствам:

$M\varepsilon_t=0$;

$$M(\varepsilon_t \cdot \varepsilon_{t\pm\tau}) = \begin{cases} \sigma^2, \text{ при } \tau = 0; \\ 0, \text{ при } \tau \neq 0, \end{cases} \quad (1)$$

Пусть длина интервала сглаживания при использовании процедуры скользящих средних $\ell = 2p+1$, тогда весовые коэффициенты для каждого активного участка могут быть представлены в виде: W_{t-p}, W_{t-p+1}, …, W_{t-1}, W_t, W_{t+1}, …, W_{t+p-1}, W_{t+p}. Эти же весовые коэффициенты будут использованы для определения оценки случайной составляющей в центральной точке активного участка:

$$\tilde{\varepsilon}_t = \sum_{i=t-p}^{t+p} w_i \varepsilon_i .$$

Учитывая (1), можно записать: $M\tilde{\varepsilon}_t = 0$;

$$D\tilde{\varepsilon}_t = \sigma^2 \sum_{i=t-p}^{t+p} w_i^2 \quad \operatorname{cov}(\tilde{\varepsilon}_t, \tilde{\varepsilon}_{t+\tau}) = \sigma^2 \sum_{i=t-p}^{t+p-\tau} w_i w_{i+\tau} \quad (2)$$

Коэффициент автокорреляции порядка τ (для случайных остатков $\tilde{\varepsilon}_t$, отстоящих друг от друга на τ тактов времени) равен:

$$\rho_\tau = \frac{\operatorname{cov}(\tilde{\varepsilon}_t, \tilde{\varepsilon}_{t+\tau})}{\sigma^2 \cdot \sum_{i=t-p}^{t+p} w_i^2} = \frac{\sum_{i=t-p}^{t+p-\tau} w_i \cdot w_{i+\tau}}{\sum_{i=t-p}^{t+p} w_i^2}, \text{ где } \tau = 1,2, \ldots, 2 \cdot p. \quad (3)$$

Из (3) вытекает важное следствие: коэффициенты автокорреляции для производного, сглаженного ряда будут отличны от нуля вплоть до порядка $\tau = 2p$, а коэффициенты более высоких порядков будут равны нулю. С помощью скользящих средних Е. Слуцкий получил подобие цикла, напоминавшее типичное поведение некоторых экономических рядов. Взяв последние цифры облигаций из табл. выигрышного займа, он в своей работе сумел показать, что «сложение случайных причин порождает волнообразные ряды, имеющие тенденцию на про-

тяжении большего или меньшего числа волн из небольшого числа синусоид». имитировать гармонические ряды, сложенные

Подраздел 2.3. Информационные технологии статистического инструментария

Б

БАЗА ДАННЫХ

совокупность связанных данных, организованных по определённым правилам, предусматривающим общие принципы описания, хранения и манипулирования данными. Обычно является представлением информационной модели предметной области. Обращение к ней осуществляется с помощью с помощью специального вида программного обеспечения, называемого системами управления базами данных (СУБД).

БАЗА ЗНАНИЙ

часть экспертной системы, предназначенная для хранения и обработки знаний предметной области, представленных в соответствии с выбранной моделью: логической (чёткой, нечёткой), продукционной, фреймовой, а также рядом сетевых моделей – семантической, байесовской, древовидной, нейронной. Каждой модели отвечает свой язык представления знаний и свои правила вывода решений; на практике обычно используется комбинированное представление знаний с использованием различных моделей.

В

ВИТРИНА ДАННЫХ

специализированное локальное тематическое хранилище, подключенное к централизованному *хранилищу данных* и обслуживающее отдельного пользователя или отдельное направление деятельности орг-ции (см. рис. 1).

Рис. 1

Г

ГЕНЕТИЧЕСКИЙ АЛГОРИТМ

эвристический алгоритм приближённого решения задач оптимизации с использованием механизмов, напоминающих биологическую эволюцию. Отличительная особенность Г.а. – акцент на использование операторов «скрещивания» и «мутации», которые производят рекомбинацию и частичное изменение решений-особей, аналогично скрещиванию и мутации особей популяции в живой природе. Осн. идея механизма эволюции, заложенная в различные конструкции Г.а., заключается в способности

«лучших», более «приспособленных» решений – особей оказывать большее влияние на состав новой популяции за счёт длительного выживания и более многочисленного потомства. В то же время результатом операций скрещивания и мутации может стать генерация принципиально новых решений из тех областей пространства, которые не были представлены точками начальной популяции. Принципиальным преимуществом Г.а. является их естественный параллелизм, когда несколько эволюционных процессов осуществляются параллельно с периодическим обменом генетическим материалом между популяциями.

Д

ДАННЫЕ

информация фактического характера, описывающая реальные объекты (процессы и явления) предметной области, представленная в форме, определяемой метаданными, и предназначенная для дальнейшего использования в информационной системе (ИС).

По степени структурированности Д. разделяются на неструктурированные, структурированные и слабоструктурированные.

К неструктурированным Д. относятся произвольные по форме электронные документы, включающие текст, графику, мультимедиа и т.п.

Структурированные Д. организованы и упорядочены т.о., чтобы обеспечить возможность применения к ним ряда процедур обработки и преобразования, входящих в программное обеспечение ИС. Одна из самых распространённых форм структурированных Д. – табл. Это осн. форма представления сведений об объектах предметной области в базах данных.

Слабоструктурированные данные, как правило, организованы в соответствии с определёнными правилами и форматами, допускающими возможность произвольного представления информации. Это, напр., результаты анкетных опросов, когда анкета включает как закрытые, так и открытые перечни ответов на вопросы, или гипертекстовые документы.

Подавляющее большинство методов обработки, хранения и отображения Д. в ИС работает только со структурированными или слабоструктурированными Д. Поэтому неструктурированные Д. подвергаются специальной автоматической или автоматизированной (с применением ручного труда кодировщиков) структуризации, прочём сам характер Д. в процессе структуризации может существенно измениться.

ДЕРЕВО РЕШЕНИЙ

классификатор, представляющий иерархическую структуру знаний о классах объектов предметной области. Д.р. может быть построено в результате заполнения *базы знаний* экспертной системы путём извлечения знаний экспертов или статистической обработки обучающего множества, содержащего объекты, их характеристики, а также наименование классов, к которым они принадлежат. Д.р. состоит из листьев, указывающих на класс, и узлов, содержащих логические условия ветвления.

Д.р. используется для классификации объектов, не вошедших в обучающее множество. Поиск начинается с корня дерева. В каждом узле проверяется выполнение логического условия для рассматриваемого объекта. Затем осуществляется переход к следующему узлу, для которого логическое условие является истинным, до тех пор, пока не будет обнаружен класс, соответствующий объекту.

ДМ-МЕТОДЫ

(от англ. – Data Mining – «извлечение и обогащение данных») – совокупность методов интеллектуального анализа данных для обнаружения в *хранилище данных* или *витрине данных* ранее неизвестных, нетривиальных, практически полезных и интерпретируемых знаний, необходимых для принятия решений в заданной предметной области. Выделяют четыре класса содержательных задач, решаемых DM-м., т.е. построение правил, по которым каждому объекту (процессу или явлению) предметной области, описанному определённым

набором фактов из хранилища данных, соответствует определённое значение (имя класса): классификация – значение дискретной переменной (классификатора); регрессия – значение непрерывной переменной (регрессора); кластеризация (разбиение на группы) – объекты из одного кластера в определённом смысле более похожи друг на друга, чем объекты из разных кластеров; ассоциация – построение ассоциативного правила, позволяющего описать связь между двумя или более событиями, происшедшими одновременно или в течение определённого промежутка времени.

Для решения перечисленных задач используются различные DM-м., в основе которых, как правило, лежит человеко-машинная процедура структурно-параметрического обучения некоторой модели, преобразующей множество входных переменных во множество выходных в соответствии с заданным алгоритмом и *обучающей выборкой*. Каждая запись обучающей выборки содержит значения входных переменных и соответствующие им правильные (требуемые) значения выходных переменных.

После того как модель обучена (выбраны в определённом смысле наилучшая структура модели и наилучшие значения параметров) и протестирована (проверена на специальной тестовой выборке или экспертами), её можно записать в *базу знаний* и применять на практике. В случае поступления новых объектов в обучающую выборку возможно дообучение модели и, соответственно, обновление базы знаний.

Поскольку процедуры обучения, дообучения и самообучения (автоматического дообучения) свойственны интеллектуальным системам, DM-м. обычно относят к методам искусственного интеллекта.

Разработаны и широко используется на практике большое число DM-м., основанных на моделях и алгоритмах *математической статистики*. В частности, для решения задачи классификации используются различные методы *дискриминантного анализа*. Для решения задачи регрессии – методы множественной линейной, нелинейной, пошаговой и логистической регрессии. Для решения задачи кластеризации – методы расщепления смесей распределений и анализа кластеров высокой плотности. Для поиска ассоциаций – статистические методы поиска ассоциативных правил, построения байесовских сетей доверия. Однако для корректной интерпретации результатов применения статистических методов необходимо, чтобы исходные данные отвечали требованиям *теоретико-вероятностной модели*. Наряду со статистическими методами, широко используются также и эвристические методы интеллектуального анализа данных: *деревья решений*, *нейронные сети*, *генетические алгоритмы*, *нечёткая логика*, *самоорганизующиеся карты Кохонена* и др.

И

ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ (ИИС)

информационные системы, включающие в свой состав методы и технологии экспертных систем и *хранилищ данных*, и предназначенные для решения аналитических задач в определённой предметной области с использованием сведений, входящих в *базы знаний*.

ИНФОРМАЦИОННАЯ МОДЕЛЬ

множество данных, описывающих реальные объекты (процессы и явления) предметной области, их существенные характеристики и отношения между ними, достаточное для удовлетворения информационных потребностей пользователей *информационной системы*.

ИНФОРМАЦИОННАЯ СИСТЕМА (ИС)

компьютерная система, включающая вычислительное и коммуникационное оборудование, программное обеспечение, данные и метаданные, лингвистические средства, обеспечивающая поддержку системным персоналом в актуальном состоянии информационной модели некоторой части реального мира (предметной области) для удовлетворения потребностей пользователей в сборе, обработке, хранении, распространении и отображении информации.

Удовлетворение информационных потребностей пользователей обеспечивается путём выполнения информационных запросов к базам данных (БД) ИС, сформулированных пользователями с применением лингвистических средств ИС. В зависимости от характера запросов и методов их обработки ИС разделяются на транзакционные и интеллектуальные (аналитические).

В транзакционных ИС пользовательский запрос (транзакция) порождает единую завершённую (для пользователя) операцию по обработке данных, входящих в БД информационной системы. Напр., оплата набора продуктов в супермаркете или расчёт стоимости состоявшегося разговора в сети мобильной связи. Каждая транзакция обычно приводит к появлению новых записей в БД или изменению содержания определённых записей, внесённых в БД ранее. Такие информационные системы получили название OLTP (от англ. – On-Line Transaction Processing).

Помимо обработки транзакций, программное обеспечение OLTP-систем позволяет решать также стандартные расчётно-аналитические задачи (напр., расчёт годового баланса). Однако, решение новых (т.е. нестандартных, ранее не запрограммированных) аналитических задач (напр., анализ динамики продаж отдельных видов продукции с учётом изменения кросс-курсов валют) в режиме реального времени в OLTP-системах практически невозможно. В то те время, информация, накопленная в БД OLTP-систем, может оказаться весьма полезной в процессе управления орг-цией.

Для решения аналитических задач используются *интеллектуальные информационные системы* (ИИС), в которых информационные ресурсы обычно хранятся в *базах знаний* и *базах данных* специального вида – хранилищах данных. Решение базовых аналитических задач в режиме реального времени обеспечивают экспертные системы и системы OLAP (On-Line Analytical Processing). Для проведения глубокой аналитической обработки накопленных данных, напр., поиска скрытых («латентных») зависимостей и структур в многомерных массивах временных рядов, вывода из них логических правил, которые действуют в данной предметной области, реализации сценариев типа «что если», а также для проведения статистического анализа многомерных данных применяются *DM-методы*. Знания, полученные от экспертов или извлечённые из баз данных DM-методами, помещаются в базы знаний ИИС для последующего использования.

Относительно самостоятельное направление развития интеллектуальных информационных систем – мультиагентные системы, в которых реализуется принцип автономности отдельных вычислительных процессов (агентов), совместно функционирующих в распределённой системе, где одновременно протекает множество взаимосвязанных процессов накопления, обработки и поиска информации.

Разделение ИС на транзакционные и интеллектуальные – достаточно условно. Характерными примерами гибридных ИС служат транзакционные ИС с естественно-языковым интерфейсом для подготовки запросов или самообучающиеся ИИС с транзакционной формой дообучения решающих правил.

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

совокупность методов, производственных процессов, программных, технических и лингвистических средств, интегрируемых с целью сбора, обработки, хранения, распространения, отображения и использования информации в интересах её пользователей.

Распространение совр. информационных и телекоммуникационных технологий, наряду с интеллектуальной деятельностью и повышением технологического уровня произ-ва, призвано стать осн. источником увеличения добавленной стоимости в совр. экономике страны.

М

МЕТАДАННЫЕ

структурированные или неструктурированные данные, описывающие информационную модель предметной области и свойства *данных*, используемых в *информационной системе*

(ИС), напр., их генезис, состав, структуру, формат представления, содержание, место хранения, владельца, способы и полномочия доступа.

М. хранятся в ИС отдельно от данных в т.н. репозитарии М. Выделяются два уровня М. – технический и предметный. Технический уровень содержит М., необходимые для функционирования ИС (структуры и форматы данных, порядок доступа и т.п.). Предметный уровень фактически – описание информационной модели предметной области, для работы в которой создана ИС – атрибуты (характеристики) объектов и их возможные значения, имена полей в табл. и т.д.

МУЛЬТИАГЕНТНЫЕ (МНОГОАГЕНТНЫЕ) СИСТЕМЫ (МАС)

интеллектуальные информационные системы, состоящие из множества взаимодействующих между собой вычислительных элементов, называемых интеллектуальными агентами (вычислительный процесс, который функционирует в некоторой среде и способен действовать самостоятельно в этой среде с целью выполнения своего функционального назначения. Свойства интеллектуальный агента: 1. реактивность – агент в состоянии своевременно реагировать на изменения среды для выполнения своего функционального назначения; 2. проактивность – агент в состоянии действовать целенаправленно, беря на себя инициативу, для выполнения своего функционального назначения; 3. общественное поведение – агент способен взаимодействовать с другими агентами (или людьми) для выполнения своего функционального назначения.

Осн. характеристики, определяющие МАС.

Различные агенты, составляющие МАС, имеют различную архитектуру. Агентов, которые основаны на различных аппаратных средствах или реализуют различные схемы поведения, называют гетерогенными. Гомогенными называют агентов, имеющих одинаковую структуру и априорно одинаковые возможности.

Среда может быть либо статической, либо динамической. В МАС наличие множества агентов делает среду динамической с точки зрения отдельного агента.

Восприятие. Информация, поступающая агенту в МАС, обычно распределена во времени, пространстве или семантически (допускает различную интерпретацию). Этот факт делает мир частично наблюдаемым для каждого агента, что может по-разному влиять на процесс принятия решений.

Управление в МАС обычно децентрализовано. Это означает, что принятие решений в большей степени осуществляется агентом самостоятельно. В МАС, где агенты имеют одинаковые цели, становится возможным реализовать распределенные вычисления, что в свою очередь потребует разработки необходимого механизма координации.

Уровни знаний каждого агента о текущем состоянии мира могут существенно отличаться. В общем случае, в МАС, каждый должен учитывать знания других агентов в процессе принятия решения. Один из возможных инструментов совместного использования знаний агентами – «доска объявлений».

Коммуникации в МАС рассматривают как двусторонний процесс, в котором все агенты могут потенциально быть отправителями и получателями сообщений. Коммуникация может использоваться в нескольких случаях, напр., для координации среди агентов, действующих совместно, или для переговоров среди агентов, имеющих различные интересы.

Использование МАС имеет следующие преимущества перед централизованными системами: использует вычислительные ресурсы и возможности всей сети связанных агентов; МАС децентрализована и т.о. избавлена от проблем, связанных с отказом в одной точке, что присуще централизованным системам; позволяет организовать взаимодействие между различными классами систем, за счёт создания между ними «прослойки» из агентов; позволяет описать проблему в терминах автономных вза-

имодействующих между собой компонентов, что является естественным методом представления распределения заданий, командного планирования, открытых сред и т.д.; получает, отбирает и управляет информацией из пространственно распределённых источников; обеспечивает информационную поддержку процесса принятия решений в ситуациях, где экспертиза распределена в пространстве и времени.

МАС нашли приложения во многих областях деятельности, которые разделены на две осн. категории: распределённые системы, в которых агенты – узлы обработки данных. Акцент в таких системах ставиться на наличии большого числа агентов; личные программные помощники, в которых агенты играют роль проактивных ассистентов при работе пользователя с некоторым приложением.

Примеры конкретных приложений МАС.

Управление бизнес-процессами компании**.** Системы, автоматизирующие бизнес-процессы компании, создаются обычно для создания и поддержки эффективного документооборота. При реализации такой системы на базе МАС каждый отдел компании представляется отдельным агентом и каждый сотрудник отдела также представляется агентом. Тогда для решения стоящих перед ними задач агенты должны взаимодействовать друг с другом. В этом случае взаимодействие принимает форму переговоров о том, какие услуги может один агент предоставить другому и на каких условиях.

Распределённое восприятие информации – классическое приложение МАС. Осн. идея состоит в том, чтобы создать МАС, в которой агенты являются сетью пространственно распределённых датчиков. Глобальная цель такой МАС состоит в том, чтобы контролировать и отслеживать распространение изменений среды, которые происходят в пределах диапазона действия датчиков (напр., движение автомобиля). Данная задача решается проще, если агенты-датчики в данной сети взаимодействуют друг с другом, обмениваясь, напр., информацией о том, что движущийся объект приближается к границе диапазона и направляется в область действия следующего датчика.

Управление информацией и информационный поиск. МАС могут использоваться для работы с распределёнными, слабо структурированными информационными ресурсами, такими как интернет. Информационным называется агент, имеющий доступ к одному или нескольким информационным источникам и возможность получать и управлять информацией, полученной из этих источников для того, чтобы отвечать на запросы пользователя и других информационных агентов. Тогда типичная задача, решаемая с помощью МАС – поиск информации в различных информационных источниках путём задания необходимых условий поиска информационному агенту.

Электронная коммерция. Самый простой тип МАС для электронной коммерции состоит из агентов, которые сравнивают предложения на рынке с целью найти наиболее выгодные решения для покупателей. Покупатель задаёт агенту ряд параметров, которым должен отвечать искомый товар, а агент в свою очередь осуществляет поиск доступных предложений и сравнивает их согласно данному набору параметров.

Человеко-машинный интерфейс. При взаимодействии человека и компьютера через пользовательский интерфейс, используется парадигма взаимодействия, известная как прямая манипуляция, т.е. компьютерная программа осуществляет только те действия, которые пользователь инициирует самостоятельно (напр., нажатие на пункт меню). Идея использования МАС как интерфейса заключается в том, чтобы компьютерные программы при определённых обстоятельствах могли брать на себя инициативу, а не ждать от пользователя конкретного указания, т.е. работали в двустороннем взаимодействии с пользователем.

Социальное моделирование – одно из приложений МАС как экспериментального инструмента в общественных науках. Идея заключается в том, что агенты могут использоваться для моделирования поведения человеческих обществ. В простейшем случае, индивидуальные агенты представляют отдельных людей, либо орг-ции и подобные объекты. Такой подход – один из немногих способов реалистичного мо-

делирования поведения сети объектов, каждый из которых имеет свои интересы, в контексте правил, норм и общих стратегий.

Н

НЕЙРОННАЯ СЕТЬ

совокупность искусственных нейронов – вычислительных элементов, определённым образом связанных друг с другом и внешней средой. На каждом такте функционирования сети нейрон выполняет вычисление выходного сигнала нейрона – значения функции активации (передаточной функции) нейрона от аргумента – скалярного произведения вектора значений входных сигналов и вектора весовых коэффициентов входных связей нейрона. В процессе функционирования сети осуществляется преобразование вектора входных сигналов, поступающих в сеть из внешней среды, в вектор выходных сигналов, передаваемых из сети во внешнюю среду.

Конкретный вид выполняемого сетью преобразования входных данных обуславливается не только характеристиками нейронов, но и особенностями её архитектуры, а именно топологией межнейронных связей, выбором определённых подмножеств нейронов для обмена информацией с внешней средой, наличием или отсутствием взаимодействия между нейронами, направлением и способами управления и синхронизации передачи информации между нейронами. Важнейшее свойство Н.с. – возможность её обучения, т.е. изменения внутренних числовых и структурных параметров сетевой модели т.о., что на выходе Н.с. генерируется вектор значений, совпадающий в определённом смысле с результатами примеров обучающей выборки. Изменение параметров сетевой модели может выполняться разными способами в соответствии с различными алгоритмами обучения

Наиболее часто Н.с. используются для решения задач: классификации – указания принадлежности объекта, представленного вектором характеристик, одному или нескольким предварительно определённым классам; прогнозирования – предсказания значения целевой функции при известной последовательности её значений в предшествующие моменты времени; оптимизации – нахождения решения, удовлетворяющего системе ограничений и доставляющего экстремум заданной целевой функции; доступа к памяти по содержанию – обеспечения доступа к сегментам памяти вычислительной системы по конкретному содержанию (ассоциативная память); управления – расчёта такого входного воздействия на систему, при котором система следует по желаемой траектории.

НЕЧЁТКАЯ ЛОГИКА

раздел математики, представляющий собой обобщение классической логики и теории множеств; основана на допущении, что принадлежность элемента к множеству может измеряться не логической функцией со значениями 1 («истина») или 0 («ложь»), а некоторой монотонно неубывающей функцией принадлежности, принимающей значения в интервале [0..1]. Такие множества называются нечёткими.

Предмет Н.л. – построение моделей неформальных рассуждений человека и использование их в экспертных системах. В этом случае *база знаний* – нечёткая система – множество нечётких правил, преобразующих входные данные в выходные. Нечёткое правило – условное высказывание вида «если X есть A, то Y есть B», где A и B – нечёткие множества. Можно сказать, что каждое нечёткое правило действует как ассоциативная память, связывающая нечёткий отклик B с нечётким стимулом A. В простейшем случае эти правила устанавливает эксперт. В более сложном случае *нейронная сеть* обучается правилам по данным или по наблюдениям за действиями людей-экспертов.

На каждый пример входных данных в некоторой степени откликаются все правила нечёткой системы. Чем ближе сходство входного примера с частью «если» нечёткого правила, тем больше отклик части «то». В нечёткой системе строится «свертка» всех множеств части «то» нечётких правил, которая и является выходным результатом нечёткой системы. Правило свертки определяется алгеброй Н.л.

П

ПАКЕТ ПРИКЛАДНЫХ ПРОГРАММ MICROSOFT EXCEL

мощное программное средство для работы с табл. данных, позволяющее вычислять значения, упорядочивать, анализировать, графически представлять различные виды данных и обмениваться ими с другими пользователями, что позволяет принимать более обоснованные решения. Данные П.п.п. Microsoft Excel – табл., состоящие из т.н. ячеек; табл. размещаются в рабочих листах. Каждый файл является книгой, которая может содержать несколько рабочих листов. Такое размещение данных представляет преимущества при связывании данных в табл. и облегчает их анализ. Состав функциональных возможностей П.п.п. Microsoft Excel достаточно широк, гибок и применим в самых различных ситуациях. Одно из важнейших свойств программы обработки электронных табл. – возможность использования формул и функций, необходимых для выполнения табличных вычислений. В качестве аргументов в формулах могут использоваться как числа, так и адреса ячеек. Причём формула может содержать ссылки на ячейки, которые расположены на другом рабочем листе или даже в таблице другого файла. Кроме того, П.п.п. Microsoft Excel позволяет создавать нескольких формул и имён, ссылающихся на одинаковые ячейки или диапазоны в нескольких последовательных листах одной книги, а также работать со сложными формулами, содержащими несколько операций. Однажды введённая формула в любое время может быть модифицирована. П.п.п. Microsoft Excel предлагает удобный интерфейс для работы с формулами, что даёт возможность ускорить процесс поиска ошибок при их введении, в частности для отслеживания зависимостей между составляющими формул служат трассировки, позволяющие определить влияющие и зависимые ячейки. Также для помощи пользователю программа содержит встроенный менеджер формул, что помогает найти ячейки, содержащие ошибочные данные или неправильную ссылку в большой табл. Поскольку некоторые формулы и их комбинации встречаются очень часто, П.п.п. Microsoft Excel предлагает более 200 заранее запрограммированных формул, называемых функциями. Список формул включает не только набор стандартных математических функций, но и функции для работы с массивами данных, логические и текстовые функции, а также функции, наиболее часто применяемые в статистическом анализе и при инженерных расчётах. Встроенный в П.п.п. Microsoft Excel «Мастер функций» облегчает работу пользователя по созданию формул, автоматически исправляя наиболее общие ошибки. П.п.п. Microsoft Excel разработана как программа для обработки больших однородных наборов данных. Осн. объект работы программного продукта – список, под которым понимается упорядоченный набор данных, имеющих одинаковую структуру. Для анализа больших списков в П.п.п. Microsoft Excel предусмотрены специальные средства и операции для их обработки. Реализованные в П.п.п. Microsoft Excel опорные табл. (Pivot Table) предназначены для анализа соотношений между данными в списке, что позволяют лучше понять тенденции и закономерности, которым подчиняются табличные данные. Однажды созданную структуру опорной табл. можно легко изменить в интерактивном режиме с помощью «Мастера табл.» путём перемещения названий полей данных из одной части табл. в другую, что расширяет возможности пользователя по анализу исходного списка данных. Кроме опорных табл., в П.п.п. Microsoft Excel имеются и другие методы анализа массивов данных, напр., директива поиска решения уравнений, которая по заданным значениям полей находит другие значения, удовлетворяющие определенным соотношениям.

Еще одно неоспоримое достоинство П.п.п. Microsoft Excel – средства консолидации, которые позволяют собрать на один рабочий лист данные, поступившие от нескольких источников, получить общий итог для всех данных, расположенных на многих рабочих листах. Такое средство есть и в других системах электронных табл., но при этом требуется точное совпадение структуры консолидируемых рабочих листов, что часто нарушается в реальной

жизни. П.п.п. Microsoft Excel может консолидировать не совпадающие по структуре рабочие листы за счёт того, что позволяет помечать необходимые данные на рабочих листах. За счёт включения в программу «Пакета анализа», П.п.п. Microsoft Excel можно использовать и в качестве средства статистической обработки данных. Набор встроенных функций позволяет пользователю проверять различные гипотезы относительно природы формирования данных исходного массива, получать результаты корреляционного, регрессионного, дисперсионного анализа для выявления взаимосвязи между показателями, проводить преобразование и анализ временных рядов. Указанные методы статистической обработки данных реализованы во многих программных продуктах и статистических пакетах анализа данных. П.п.п. Microsoft Excel уступает многим из них по мощности и набору встроенных статистических процедур, однако имеет преимущества по доступности, т.к. этот программный продукт входит в состав Microsoft Office, который имеет наибольшее распространение и используется на большинстве компьютеров. Начиная с версии 5.0 в П.п.п. Microsoft Excel включён специальный язык программирования Visual Basic for Applications (VBA). Обычным пользователям утилита позволяет создавать макросы для автоматизации часто повторяющихся операций, создавать кнопки для быстрого запуска макросов и автоматизировать рабочие листы с помощью имеющихся в арсенале программного продукта форм. Кроме этого, наличие VBA позволяет использовать П.п.п. Microsoft Excel в качестве базы для разработки специализированных прикладных систем, что делает программу привлекательной и для профессионалов.

Помимо разнообразных возможностей числового анализа табл., П.п.п. Microsoft Excel предоставляет широкий выбор средств графического представления данных и результатов анализа. Это расширяет «презентационные качества» программного продукта, так как графическое изображение легче воспринимается визуально в сравнении с табличным представлением. В некоторых случаях доступность графического представления данных позволяет не только провести их анализ на начальном этапе, но и выработать пути дальнейшего исследования. П.п.п. Microsoft Excel совместим с другими приложениями Microsoft Office, а также с подобными себе табл. и базами данных других производителей, что позволяет П.п.п. Microsoft Excel импортировать в свои табл., объекты из других прикладных программ и передавать (экспортировать) свои табл. для встраивания в другие объекты. Программу П.п.п. Microsoft Excel можно настраивать в соответствии с индивидуальными запросами очень широкого круга пользователей. Каждый пользователь программного продукта, определив круг наиболее часто используемых операций, может организовать работу с ними наиболее удобным для себя образом. С другой стороны, те функции программы, которые никогда не используются, можно вообще убрать из конфигурации, чтобы сэкономить ресурсы компьютера и повысить эффективность обработки.

ПАКЕТ ПРИКЛАДНЫХ ПРОГРАММ STATISTICA

всесторонняя система анализа данных, включающей большой набор статистических процедур, тыс. настраиваемых высококачественных графиков презентационного качества, разнообразные методы разведочного анализа и многие другие функции. Помимо общих статистических и графических средств, в системе имеются специализированные модули, напр., для проведения социологических или биомедицинских исследований, решения технических и промышленных задач: карты контроля качества, анализ процессов и планирование эксперимента. Работа со всеми модулями происходит в рамках единого программного пакета, для которого можно выбирать один из нескольких предложенных интерфейсов пользователя. С помощью реализованных в системе STATISTICA мощных языков программирования, снабжённых специальными средствами поддержки, легко создаются законченные пользовательские решения и встраиваются в различные другие приложения или вычислительные среды.

П.п.п. STATISTICA разработан фирмой StatSoft (США), основанной еще в 1984. Первоначально он был представлен в виде простого модуля для самой популярной в то время электронной табл. – Lotus 1-2-3. Как самостоятельный П.п.п. STATISTICA появился в 1991, в нём был реализован т.н. графически-ориентированный подход к анализу данных. Этот пакет обладал рядом существенных преимуществ перед другими статистическими пакетами (за счёт оптимизации удалось добиться повышения скорости обработки более чем в 10 раз по сравнению с другими пакетами, пакет мог работать фактически с неограниченным объёмом данных). В 1992 вышла версия STATISTICA для Macintosh, которая быстро приобрела заслуженную популярность среди пользователей. В 1994 вышла версия STATISTICA для Windows, занявшая лидирующее положение среди статистических пакетов. В 1995 П.п.п. STATISTICA включена в число 100 лучших программных продуктов (WINDOWS Magazine, февр. 1995). В конце 1995 вышла версия STATISTICA 5.0. От предыдущей версии она отличается более удобным пользовательским интерфейсом, полной совместимостью с Windows 95; включает в себя мощные средства работы с данными, богатые графические возможности и большое количество методов и процедур статистического анализа. STATISTICA 5.0 полностью удовлетворяет осн. стандартам среды Windows. Это, прежде всего, стандарты пользовательского интерфейса – MDI, использование технологий DDE – динамического обмена данными из других приложений, OLE – связывания и внедрения объектов, поддержка осн. операций с буфером обмена и др. В отличие от предыдущих версий в неё включен внутренний язык программирования Statistica BASIC, который позволяет пользователю расширять возможности системы. Пользователь может добавить собственную панель инструментов с тем или иным методом статистического анализа. Возможность дописывать (наращивать) систему при помощи встроенного языка программирования (из которого можно вызвать и любую внешнюю библиотеку DLL) является несомненным достоинством пакета.

В 1996–98 появились новые выпуски пакета – STATISTICA 5.1, 5.1-97 и 5.1-98. В систему были добавлены новые специализированные модули, учтены все новые форматы Windows и MS Office, сделаны различные дополнения и улучшения. Благодаря наиболее совр. и удобным аналитическим инструментам компания заняла лидирующие позиции среди производителей статистического программного обеспечения.

В 1999–2000-е гг. StatSoft представила улучшенные и оптимизированные программные продукты, в частности STATISTICA 5.5, включающую такие наиболее полные инструменты анализа, как GLM (общие линейные модели), GSR (общая ступенчатая регрессия), GLZ (обобщённые линейные модели), PLS (частные наименьшие квадраты).

В 2001–03 презентация STATISTICA 6.0 ознаменовала появление нового поколения статистических программных продуктов. STATISTICA 6.0 – высокотехнологичный продукт, основанный на COM архитектуре, обладающий уникальными функциональными и пользовательскими возможностями. STATISTICA 6.0 получила наилучшие отзывы среди клиентов как никакая другая версия за всю практику StatSoft.

У пакета есть специальная версия для обучения основам статистических методов – Student Edition of STATISTICA. Эта версия позволяет анализировать файлы данных, включающих не более 400 наблюдений, и представляет собой урезанный вариант пакета. Осн. версия пакета может дополнительно комплектоваться специализированными модулями: Power Analysis, Neural Networks и др. STATISTICA Neural Networks – универсальная система анализа данных методами нейронных сетей. Пакет содержит самые совр. нейросетевые технологии и средства поддержки пользователя: полный выбор архитектур, автоматический конструктор сетей, генетические алгоритмы и интерфейс прикладного программирования. STATISTICA Industrial Solution – обширный выбор специализированных статистических методов для пром. приложений, включает модули: карты контроля

качества, анализ процессов и планирование эксперимента. STATISTICA Enterprise-wide SPC System (SEWSS) – готовые решения, предназначенные для аналитико-статистической обработки данных и управления процессами в режиме реального времени, позволяющие осуществлять оперативный анализ и контроль процессов и прогнозировать нештатные ситуации.

Технические возможности П.п.п. STATISTICA:

- программа способна обрабатывать огромные массивы данных – табл. до 32000 переменных, основные виды анализа – до 4096 переменных. Возможны двукратная и четырёхкратная точность обработки. Режим экспорта-импорта и поддержка OLE, DDE и ODBC позволяют обмениваться данными со всеми популярными базами данных и электронными таблицами, включая MS Excel и MS Access; в ней имеется специальный модуль – менеджер файлов, который может создавать мегафайлы и манипулировать ими.
- данные анализируются в форматах осн. электронных табл. Встроенные графический и текстовый редакторы позволяют публиковать результаты работы в Internet в формате HTML, а текст и графики, полученные в окне отчёта, можно сохранить в виде HTML-страницы. Открытая архитектура пакета обеспечивает возможность добавления процедур пользователя, в том числе с использованием встроенного языка программирования и языка макрокоманд. Мультимедийный обзор возможностей системы и электронный учебник облегчает начало работы с пакетом.
- программа полностью соответствует стандартам Microsoft и совместим с новыми версиями Windows XP и MS Office. Среди его ключевых особенностей стоит отметить мощный инструмент построения запросов к базам данных Statistica Query.
- повышенная (quadruple) точность математических операций позволяет проводить анализ данных даже с очень малым разбросом величин.
- расчёты и построение графиков выполняются с очень высокой скоростью (за счёт оптимизации программного кода и механизмов управления памятью).
- программа предлагает множество вариантов научных и технических графиков и диаграмм при великолепном качестве и точности отображения информации.

П.п.п. STATISTICA предлагает пользователю широкий выбор методов анализа, среди них осн. модули:

- Quick Basic Statistics (быстрый анализ) – модуль, позволяющий быстро провести анализ наиболее употребительными методами;
- Basic Statistics/Tables (осн. статистические методы и табл.) – описательные методы статистики, табл. частот и корреляций, регрессии и другие базовые статистические методы;
- Nonparametrics/Distribution – внутригрупповые и межгрупповые непараметрические тесты, сравнение различных дискретных и непрерывных теоретических распределений с распределением наблюдаемых величин;
- ANCOVA/MANCOVA – однофакторный и многофакторный дисперсионный и ковариационный анализ;
- Multiple Regression – различные методы множественной линейной и фиксированной нелинейной регрессии (в частности, полиномиальной, экспоненциальной, логарифмической и др.);
- Nonlinear Estimation – методы подгонки к нелинейным зависимостям данных различных функций, в том числе заданных пользователем;
- Time Series/Forecasting – анализ при помощи временных рядов. В модуле «Временные ряды» реализован широкий набор методов описания, построения моделей, декомпозиции и прогнозирования временных рядов, как во временной, так и в частотной области. Процедуры модуля «Временные ряды»: преобразования, построение моделей, графики, автокорреляции; АРПСС и

анализ прерванных временных рядов (рядов с интервенциями); сезонное и несезонное экспоненциальное сглаживание; классическая сезонная декомпозиция (метод Census I); месячная и квартальная сезонная X-11-декомпозиция и корректировка (метод Census II); полиномиальные модели распределённых лагов; спектральный (Фурье) анализ и кросс-спектральный анализ; прогнозирование на основе регрессионных методов.

- Cluster Analysis – различные методы кластерного анализа и классификации;
- Factor Analysis – выделение наиболее существенных факторов сложного объекта методами повторных главных компонент, минимальных остатков, максимального правдоподобия;
- Canonical Analysis – метод канонического анализа корреляции между двумя группами переменных;
- Multidimensional Scaling – многомерное шкалирование;
- SEPATH – многомерный анализ с помощью моделирования причинных связей между переменными линейными структурными уравнениями, в том числе оценка достоверности результатов методом статистического моделирования Монте-Карло;
- Reliability/Item Analysis – анализ надежности сложного объекта на основе результатов диагностики его элементов;
- Discriminant Analysis – дискриминантный анализ, позволяющий на основе определенного критерия отнести объект к некоторому классу;
- Log-linear Analysis – логарифмический линейный анализ сложных многоуровневых таблиц частот;
- Survival Analysis – анализ долговечности (выживания) для задач социологии (особенно необходим страховым компаниям), биологии, инженерных задач (долговечность машин, сооружений) и др.
- Три модуля программы объединены в общий класс «Промышленная статистика».
- Quality Control – широкий набор методов контроля качества;
- Process Analysis – набор методов анализа производственных процессов, в том числе калибровочный анализ повторяющихся партий продукции;
- Experimental Design – модуль планирования эксперимента в промышленных и прикладных областях.
- STATISTICA Neural Networks

Модули программы не являются независимыми друг от друга и часто используют одни и те же процедуры. В процессе работы легко переключиться с одного модуля на другой. Более того, пользуясь встроенным в систему командным языком (Statistica Command Language – SCL), можно запустить программу в т.н. пакетном режиме. В этом случае Statistica шаг за шагом, переключаясь с одного модуля на другой, обработает данные и выводит результаты на печать или в файл. При частом использовании SCL-режима в интерфейс можно добавить специальную кнопку, которая будет автоматически запускать нужную последовательность действий.

П.п.п. STATISTICA предлагает богатые возможности графического представления обрабатываемых данных. Поддерживаются следующие виды графиков и диаграмм: матричные графики, пиктографики, диаграммы рассеяния, гистограммы, тернарные графики, карты линий уровня, круговые диаграммы, категоризованные графики, вероятностные графики, графики поверхностей, трассировочные графики, комбинированные графики, вращение и перспектива, подгонка, сглаживание, сечения. Кроме того, в пакет входят 4 специальных модуля, реализующих дополнительные математические методы обработки информации: визуальные общие линейные модели (Visual General Linear Models [VGLM]) – реализация общей линейной модели (General Linear Model) для анализа откликов одной или нескольких непрерывных переменных как функции одной или нескольких категориальных или непрерывных независимых переменных. Модуль VGLM предлагает наиболее полный набор методов, основанных

на дисперсионном анализе; визуальные обобщённые линейные модели (Visual Generalized Linear Models [VGLZ]) – позволяют пользователю проводить поиск линейных и нелинейных зависимостей между переменной откликов и категориальными или непрерывными предикторами (включая мультиномиальную логит- и пробит регрессию, модели распознавания сигнала и многие другие); визуальная общая пошаговая регрессия (Visual General Stepwise Regression [VGSR]) – предоставляет исчерпывающий набор методов пошаговой регрессии и выбора наилучшего множества предикторов для модели, которые можно применять и к непрерывным и к категориальным переменным; визуальные частные наименьшие квадраты (Visual Partial Least Squares [VPLS]) – использует тот же гибкий интерфейс пользователя, что и модули VGLM, VGSR и VGLZ, и содержит весь спектр алгоритмов для анализа и решения задач с использованием простого и многомерного метода частных наименьших квадратов.

На данный момент П.п.п. STATISTICA имеет более 500 тыс. зарегистрированных пользователей во всем мире и является наиболее динамично развивающимся пакетом на рынке статистического программного обеспечения. Имеются версии системы на немецком, французском, японском, испанском, польском и других языках. Пользователи системы – крупнейшие университеты, исследовательские центры, компании, банки всего мира, гос. учреждения.

ПАКЕТ ПРИКЛАДНЫХ СТАТИСТИЧЕСКИХ ПРОГРАММ (SPSS)

Программное обеспечение SPSS предназначено для использования на всех этапах аналитического процесса исследования: планирования и сбора данных, подготовки данных к анализу, проведения анализа и создания отчётов, представления результатов исследования. В 2009 компания SPSS Inc. (междунар. компания со штаб-квартирой в США) переименовала все свои программные продукты и создала новый бренд PASW (от англ. Predictive Analytics Software – программное обеспечение для прогностической аналитики). Бренд PASW объединяет четыре семейства продуктов.

PASW Statistics – статистический анализ (ранее SPSS Statistics). Программные продукты семейства – полнофункциональная система, предназначенная для решения бизнес- и исследовательских задач при помощи анализа данных. Большой выбор статистических процедур позволяет эффективно выполнять обработку данных различных типов, наглядно представлять итоги анализа в виде табл. и диаграмм, а также распространять и внедрять полученные результаты.

PASW Modeler – моделирование (ранее Clementine). Программные продукты этого семейства обеспечивают принятие обоснованных решений, опираясь на надёжные модели, полученные с помощью *DM-методов*. Программное обеспечение позволяет обнаруживать скрытые закономерности, предсказывать и оценивать возможные результаты альтернативных вариантов действий.

PASW Data Collection – сбор данных (ранее SPSS Dataentry и SPSS Dimensions). Программные продукты этого семейства позволяют проводить опросы и маркетинговые исследования любым удобным способом – по телефону, с помощью портативных компьютеров или в сети Интернет.

PASW Collaboration and Deployment Services – интеграция и внедрение в бизнес-процессы (ранее SPSS Predictive Enterprise Services). Семейство продуктов предназначено для быстрого и надёжного внедрения прогностической аналитики в процессы принятия решений, что позволяет обеспечить необходимый баланс между поставленными целями и практическими результатами использований.

Программные продукты компании SPSS Inc. имеют модульную структуру. Statistics Base – ключевой элемент пакета, одна из функций которого – обеспечение доступа и управление данными. Имеется возможность загружать данные, хранящиеся в *базах данных* различных типов, и подготавливать их к анализу. Про-

граммное обеспечение поставляется с набором драйверов для многих ODBC-совместимых баз данных, включая Oracle®, Microsoft® SQL Server™, Microsoft Access®, IBM DB2® UDB, и Sybase™. При помощи соответствующих драйверов можно организовать доступ к любым другим ODBC-совместимым базам данных. Также возможно импортировать данные из OLEDB источника, минуя ODBC формат. Кроме этого, можно получить доступ к данным в форматах SAS®, Stata®, Microsoft Excel®.

В состав Statistics Base включены осн. статистические процедуры, позволяющие решать задачи подготовки данных для анализа и проведения анализа, а также создание отчётов.

Процедура «Частоты» даёт возможность вычислять статистики и строить диаграммы, полезные для описания номинальных и порядковых типов переменных. В результате выполнения процедуры выводятся такие статистики и графики как частоты, проценты, кумулятивные проценты, медиана и др.

Процедура «Описательные статистики» осуществляет вывод одномерных итоговых статистик для количественных переменных, а также вычисляет стандартизованные значения (z-значения) переменных. В результате выполнения процедуры выводятся такие статистики, как объём выборки, среднее значение, миним. и макс. значения, стандартное отклонение, дисперсия, размах, сумма, стандартная ошибка среднего, асимметрия, эксцесс, стандартные ошибки асимметрии и эксцесса.

Процедура «Исследовать» вычисляет итоговые статистики и выводит диаграммы как для всех наблюдений, так и отдельно для групп наблюдений. Эта процедура удобна в случае идентификации выбросов, проверки предположений и описании различий между группами наблюдений. Процедура «Исследовать» позволяет определить, подходят ли для анализа данных статистические методы, которые предполагается использовать. Результаты выполнения процедуры могут показать, что необходимо провести преобразование данных, если применение выбранного метода требует нормально распределенных данных. Или же может быть принято решение, что надо воспользоваться непараметрическими критериями.

Процедура «Табл. сопряжённости» позволяет сформировать двумерные и многомерные табл., а также вычисляет целый ряд критериев и мер силы связи для двумерных табл. Структура табл. и то, упорядочены категории или нет, определяет, какие меры и критерии использовать: Хи-квадрат Пирсона, хи-квадрат отношение правдоподобия, критерий линейно-линейной связи, точный критерий Фишера, ро Спирмана, коэффициент сопряжённости, фи, V Крамéра и т.д.

В процедуре «Подытожить наблюдения» вычисляются значения статистик для переменных по подгруппам, задаваемым категориями одной или нескольких группирующих переменных. Все уровни группирующей переменной представляются в табл. сопряжённости. Выводятся также итожащие статистики для каждой переменной по всем категориям.

В процедуре «Средние» вычисляются средние значения для подгрупп и связанные с ними одномерные статистики для зависимых переменных внутри категорий одной или нескольких независимых переменных. Дополнительно можно провести однофакторный дисперсионный анализ, найти значения статистики эта (eta), а также выполнить тесты на линейность.

Процедура «OLAP (Online Analytical Processing) Кубы» вычисляет итоги, средние значения и другие одномерные статистики, для количественных подытоживаемых переменных внутри категорий одной или нескольких категориальных группирующих переменных. Для каждой категории каждой группирующей переменной в табл. создаётся отдельный слой.

Процедура «Т-критерий для независимых выборок» позволяет сравнивать средние значения для двух групп наблюдений. Процедура «Однофакторный дисперсионный анализ (ANOVA)» выполняет однофакторный дисперсионный анализ для количественной зависимой переменной по единственной факторной (независимой) переменной. Дисперсионный анализ используется для проверки гипотезы о равенстве нескольких средних значений, соответ-

ствующих различным группам или уровням факторной переменной. Этот метод является расширением двухвыборочного t-критерия.

Процедура «Общая линейная модель (ОЛМ) – одномерный анализ» выполняет регрессионный и дисперсионный анализы для одной зависимой переменной по одному или нескольким факторам и/или переменным. Факторная переменная делит генеральную совокупность на группы. Используя данную процедуру можно проверить нулевую гипотезу о влиянии других переменных на средние различных групп значений единственной зависимой переменной. При этом можно исследовать как взаимодействие между факторами, так и эффекты отдельных факторов, некоторые из которых могут быть случайными.

Для проверки гипотез в процедуре «ОЛМ – одномерный анализ» доступны обычно используемые априорные контрасты. После того как общий тест с использованием F-критерия показал значимость различий, можно использовать апостериорные критерии, чтобы оценить различия между конкретными средними. Помимо проверки гипотез эта процедура дает оценки параметров модели. В процедуре «Парные корреляции» вычисляются коэффициент корреляции Пирсона, ро- Спирмана и тау-b Кендалла, а также уровни значимости для них. Корреляции измеряют связь между переменными или рангами. Процедура «Частные корреляции» вычисляет частные коэффициенты корреляции, которые описывают линейную связь между двумя переменными при устранении влияния одной или нескольких дополнительных переменных. Процедура «Расстояния» вычисляет любую статистику из широкого набора доступных, измеряющих либо сходства, либо различия (расстояния), причем либо между парами переменных, либо между парами наблюдений. Эти меры сходства или расстояния затем используются в других процедурах, таких как «факторный анализ, кластерный анализ или многомерное шкалирование», для того, чтобы помочь анализировать сложные наборы данных. В базовый модуль включены следующие виды регрессии: линейная, порядковая, подгонка кривых, частичная регрессия методом наименьших квадратов. Линейная регрессия оценивает коэффициенты линейного уравнения, содержащего одну или несколько независимых переменных, позволяющие наилучшим образом предсказать значение зависимой переменной. Порядковая регрессия позволяет моделировать зависимость политомической порядковой реакции на набор предикторов. Процедура «Подгонка кривых» позволяет вычислять статистики и строить сопутствующие графики для 11 различных нелинейных регрессионных моделей оценки кривых.

В процедуре «Частичная регрессия» методом наименьших квадратов (МНК) оцениваются модели регрессии частичных наименьших квадратов, известные также как «проекция в латентную структуру». Данная методика прогнозирования – альтернатива регрессии обычным МНК, канонической корреляции или моделированию при помощи структурных уравнений. Это метод особенно полезен для предикторных переменных с высокой корреляцией, или когда число предикторов превышает количество наблюдений. Процедура «Статистики отношений» предоставляет полный список итоговых статистик для описания отношения двух количественных переменных. Процедура «Кривые ROC» полезна для оценки эффективности схем классификации, в которых есть одна переменная с двумя категориями, по которым классифицируются объекты.

В базовый модуль также включены такие процедуры, как «Непараметрические критерии, Дискриминантный анализ, Факторный анализ, Двухэтапный кластерный анализ, иерархический кластерный анализ и Кластерный анализ методом k-средних, Анализ методом ближайшего соседа, Анализ пригодности, Анализ множественных ответов, Многомерное шкалирование».

Процедура «Точные критерии» предоставляет два дополнительных метода для вычисления уровней значимости для статистик, доступных с помощью процедур «Табл. сопряжённости» и «Непараметрические критерии». Эти методы, точный и Монте-Карло, предоставляют средства для получения корректных результатов в случае, когда данные не удовлетворяют усло-

виям стандартного асимптотического подхода. Дополнительные модули встраиваются в Statistics Base и позволяют расширять аналитические возможности программного обеспечения.

Модуль Advanced Statistics – анализ сложных взаимосвязей при помощи мощных инструментов построения моделей, среди них: обобщённые линейные модели, смешанные линейные модели, общие линейные модели с фиксированными, случайными и смешанными факторами; повторные измерения в моделях дисперсионного и ковариационного анализа; оценка компонентов дисперсии; политомические универсальные логит модели; общие логлинейные модели для анализа многовходовых табл. сопряженности; иерархические логлинейные модели для анализа многовходовых таблиц сопряженности; логлинейные и логит модели в процедуре общий логлинейный анализ.

Модуль Categories предназначен для прогнозирования категориальных откликов и исследования категориальных данных, в т.ч. при помощи карт восприятия. Модуль применяется для решения задач анализа больших, громоздких двухвходовых и многовходовых табл.; для работы с порядковыми и номинальными переменными с использованием процедур, аналогичных обычным процедурам регрессии, анализа главных компонент и канонической корреляции, а также для визуализация и исследования категориальных данных. Модуль Complex Samples предназначен для создания сложных планов случайной выборки (расслоенного и многоэтапного, а также отбора с вероятностью пропорциональной размеру элементов) и проведения корректного вычисления статистик по данным сложных выборок, в т.ч. сделать статистически более правильные выводы о ген. совокупности, учитывая план сложной выборки при проведении анализа. Модуль Conjoint предназначен для определения предпочтений клиентов, напр., оценки того, как отдельные атрибуты товаров и услуг оказывают влияние на предпочтения покупателей. Custom Tables – модуль для создания наглядных табличных отчётов любой степени сложности с широким набором итожащих статистик. Визуальный конструктор табл. с интерактивным интерфейсом предоставляет удобные возможности форматирования табл. Предусмотрена возможность вывода статистических критериев, позволяющих проверить достоверность выявленных взаимосвязей в данных, в т.ч. критерий независимости хи-квадрат, критерий сравнения долей, критерий сравнения средних значений по столбцам. Дополнительный модуль Data Preparation обеспечивает доступ к более мощным инструментам по сравнению с Statistics Base, позволяющим осуществлять более качественную подготовку данных к анализу. Становится возможным выявлять нетипичные или ошибочные значения в данных, изучать структуру пропущенных значений и исследовать распределение переменных. Модуль Classification Trees предназначен для сегментации и прогнозирования откликов при помощи деревьев решений. Classification Trees автоматически строит дерево, требуется только указать целевую переменную, переменные-предикторы и выбрать алгоритм построения дерева решений. Classification Trees автоматически просеивает данные и находит статистически значимые группы. Для получения макс. достоверных результатов можно обучить модель на подвыборке, затем протестировать надежность модели на оставшихся данных. Насколько хорошо модель описывает данные, можно увидеть, переключаясь с обучающей модели на контрольную модель. Модуль Exact Tests позволяет получить правильные выводы и корректные решения в условиях выборок малых объемов или подробной группировки, т.е. возможно получать корректные уровни значимости при любой структуре данных. Exact Tests применяется в следующих случаях: небольшое количество наблюдений, наличие переменных с большой долей ответов в одной категории, наличие в данных ярко выраженных подгрупп. Этот модуль также целесообразно применять при исследовании редких событий в больших наборах данных (напр., семьи, в которых более 5-и детей). Модуль Forecasting – средство анализа временных рядов, построения моделей и предсказания будущих событий. Процедуры модуля позволяют автоматически определить наилучшую модель

ARIMA или модель экспоненциального сглаживания, параметры и предикторы для различных временных рядов, используя эксперт построения моделей. Можно оценивать собственные модели с целью более точной настройки для каждого временного ряда, сохранить модели и применить их заново. В модуле доступна процедура сезонной декомпозиции временных рядов на гармонические компоненты.

Модуль Missing Value Analysis предназначен для заполнения пропущенных значений в целях повышения информативности данных и построения адекватных моделей, так как пропущенные значения могут серьезно повлиять на результаты анализа**.** Если игнорировать наличие пропусков в данных или полагать, что достаточно исключить из анализа данные с пропущенными значениями, то существенно вырастает риск получения неверных или незначимых результатов. Missing Value Analysis помогает заполнить пропущенные данные и получить более надёжные результаты. При помощи процедур модуля можно проверять данные и получать диагностические отчёты, помогающие обнаруживать закономерности в распределении пропущенных значений. После этого можно исследовать итоговые статистики и заполнять пропущенные значения при помощи статистических алгоритмов. Модуль Neural Networks содержит нелинейные процедуры моделирования, позволяющие обнаруживать более сложные взаимосвязи в данных. Процедуры этого модуля являются дополнением традиционных статистических методов, содержащихся в Statistics Base и дополнительных модулях к нему. Используя методы data mining модуля можно обнаруживать новые зависимости, а затем подтверждать их значимость с помощью традиционных статистических методов. Модуль Neural Networks содержит две разновидности нейронных сетей – на основе Многослойного перцептрона (MLP) и Радиальных базисных функций (RBF). Обе разновидности относятся к методам обучения с учителем на основе алгоритма обратного распространения ошибок. Они используются для прогнозирования и классификации с факторами или ковариатами в качестве предикторов. В процедуре MLP могут использоваться несколько скрытых слоев нейронов, а в процедуре RBF настройка нейронной сети производится в два этапа и, как правило, быстрее чем MLP. В обеих процедурах можно выбирать метод деления набора данных на обучающую, контрольную и проверочную выборки. В процедуре MLP архитектура нейронной сети может определяться автоматически или же можно выбрать ручную настройку и задать количество скрытых слоев и функцию активации в скрытых и выходном слоях нейронов.

Модуль Regression – средство прогнозирования с помощью широкого спектра методов регрессионного анализа, таких как мультиноминальная логистическая регрессия, бинарная логистическая регрессия, нелинейная регрессия без ограничений, нелинейная регрессия с ограничениями, Двухэтапный метод наименьших квадратов и пробит анализ. В Advanced Models имеется мощный набор методов одномерного и многомерного анализа для решения реальных практических задач.

ПАКЕТ ЭКОНОМЕТРИЧЕСКИЙ E-VIEWS

(от англ. – Econometric Views) – пакет статистического и эконометрического анализа в Windows-ориентированной компьютерной среде; впервые представлен фирмой Quantitative Micro Software (QMS) в марте 1994; версия 7.0 выпущена в дек. 2009.

П.э. E-VIEWS широко используется как экономистами – исследователями, так и финансовыми аналитиками, специалистами в области макроэкономического прогнозирования и т.д. Пакет реализует широкий спектр статистических и эконометрических методов и процедур. П.э. E-VIEWS позволяет проводить анализ данных, представленных в форме объект признак (англ. – cross-section), панельных (англ. – panel data) или временных рядов (англ. – time series). Особенно широкие возможности П.э. E-VIEWS открывает при анализе данных, представленных в виде временных рядов. Пакет включает все осн. процедуры первичного анализа данных, в т.ч., построения графиков и диа-

грамм, расчёта описательных статистик, проверки гипотез, выполнение дисперсионного и факторного анализа, вычисление автокорреляционной, частной автокорреляционной и кросскорреляционных функций для временных рядов, методы сезонной декомпозиции X11, X12-ARIMA, критерии проверки гипотезы единичного корня Дики-Фуллера, Филипса-Перрона, KPSS, критерии коинтеграции и др.

В случае одного уравнения для построения оценок могут использоваться обычный и обобщённый метод наименьших квадратов, взвешенный метод наименьших квадратов, двухшаговый метод наименьших квадратов (метод инструментальных переменных), пошаговая линейная регрессия, квантильная регрессия, обобщенный метод моментов. Реализованы методы построения моделей для ограниченных зависимых переменных, в том числе, бинарных, упорядоченных, цензурированных или усеченных.

Методы анализа временных рядов представлены моделями авторегрессии – скользящего среднего ARMA и ARMAX, и моделями авторегрессионной условной гетероскедастичности ARCH(p), GARCH(p,q), EGARCH(p,q), TARCH(p,q), PARCH(p,q).

П.э. E-VIEWS позволяет решать задачу оценивания параметров линейных и нелинейных систем одновременных уравнений с использованием метода наименьших квадратов, двухшагового метода наименьших квадратов, кажущихся несвязанными регрессий (SUR), трехшаговым методом наименьших квадратов, обобщенным методом моментов (GMM) и методом макс. правдоподобия с полной информацией (FIML). Системы уравнений могут включать идентифицирующие ограничения и авторегрессионные члены. Удобные средства работы предусмотрены для моделей векторной авторегрессии – исправления ошибок, включая тесты причинности, построение функции реакции на импульсы и разложения ошибки прогноза. Богатый спектр процедур представлен для панельных данных.

Практически все возможности П.э. E-VIEWS доступны через систему меню. Вместе с тем в пакет встроен командный язык, позволяющий вводить команды в режиме командной строки. Набранные в командной строке команды немедленно исполняются, либо обрабатываются в пакетном режиме. Командный язык позволяет получить доступ ко всем возможностям пакета, доступного через меню, позволяет программировать циклы или условные переходы, поддерживает основные матричные операции, включая сложение, умножение, произведение Кронекера, нахождение собственных чисел и векторов и др.

Простота освоения и использования П.э. E-VIEWS связана с реализацией в пакете концепции объектно-ориентированного программирования. Объекты представляют собой наборы совместно используемых данных и операций над ними. Примеры объектов: рабочий файл (Workfile), ряд данных (Series), уравнение (Equation), вектор коэффициентов (Coefficient Vector), график (Graph), модель (Model), векторная авторегрессия (VAR) и др. Каждый из типов объектов отображается в рабочем файле специальным значком. С каждым объектом ассоциирован собственный набор представлений и процедур, которые могут применяться к данному типу объектов. Напр., объект серия (series) представляет собой информацию о значениях заданной переменной для множества объектов. Процедуры представления включают просмотр значений в виде электронной табл., графика, расчёт описательных статистик и т.д. Объект уравнение (equation) включает информацию относящуюся в взаимосвязи между переменными. Т.к. объект equations включает всю информацию, относящуюся к оцениваемому уравнению, можно свободно переключаться между несколькими уравнениями, просматривать результаты, проверять гипотезы или строить прогнозы.

П.э. E-VIEWS содержит развитую справочную систему, содержащую сведения, как об особенностях команд, так и о реализуемых эконометрических методах.

ПАКЕТ ЭКОНОМЕТРИЧЕСКИЙ STATA

мощный универсальный пакет статистического анализа данных, ориентированный на научных работников, студентов, аспирантов, исследователей в прикладных областях, интенсивно пользующихся эконометрическим аппаратом в своей работе.

Производитель и осн. распространитель программы – компания Stata Corporation, расположенная в г. Колледж Стейшн (College Station, США); первая версия пакета выпущена в 1985; в 2009 стала доступна одиннадцатая версия программы.

Пакет реализован в средах Windows, Macintosh или Unix, в т.ч. Linux. Версии пакета для разных операционных систем являются полностью совместимыми по функциональным возможностям и форматам данных. Пакет особенно хорош для обработки пространственных данных (англ. – cross-section data), панельных (англ. – panel data) и данных типа времени жизни (англ. – survival-time data).

Работа с пакетом основана на командном интерфейсе – каждая необходимая операция описывается пользователем в виде отдельной команды с жёстко заданным синтаксисом. Большинство команд Stata имеют формат: [by variable list:] command [variable list] [if condition] [in range] [using _file name] [[weights]], [options]. Напр., команда by region: regress y x1 x2 if age<55 определяет расчёт для всех категорий переменной region моделей линейной регрессии зависимой переменной y от регрессоров x1 и x2, для наблюдений, удовлетворяющих условию age<55.

Начиная с версии 8.0 П.э. STATA включает графический интерфейс. Меню и диалоги обеспечивают доступ практически ко всем возможностям пакета. Кроме того П.э. STATA имеет развитую систему встроенной подсказки и предоставляет пользователю возможность выполнения заранее подготовленного набора команд. Последовательность команд обычно сохраняется в текстовых файлах, содержащих команды, разделённые символами переноса строки (do- файлах). Данная возможность является полезной, если необходимо отработать набор команд на ограниченной выборке и затем применить ко всему имеющемуся набору данных; регулярно повторять эконометрический анализ при получении новых данных; выполнять заданный набор команд для нового исследования; сохранять протокол работы при исследовательском анализе данных для обеспечения воспроизводимости результатов расчётов. Возможности интерактивного режима работы полностью идентичны возможностям пакетной обработки.

П.э. STATA может одновременно работать лишь с одним массивом данных. Размещение всего набора данных в оперативной памяти обеспечивает высокую скорость выполнения команд. Если массив данных не может быть полностью размещен в оперативной памяти, скорость работы пакета существенно падет, что затрудняет обработку очень больших наборов данных. П.э. STATA располагает развитой системой команд манипуляции данными. Среди них: возможности определения и преобразования переменных, добавления и удаления объектов, присоединения файлов, сортировки, обработки по группам наблюдений, транспонирования данных и т.д. Также встроены специализированные команды для управления специальными типами данных, такими как временные ряды, панельные данные, данные типа времени жизни, множественные соответствия. Отдельный модуль Stata Transfer обеспечивает возможности импорта данных из большинства популярных форматов.

Важнейшее достоинство П.э. STATA – большой спектр реализованных статистических методов. Пакет постоянно расширяется за счёт пополнения архива пользовательских модулей, доступного через Интернет. Написание новых модулей облегчается использованием интегрированного языка операций с матрицами – Mata. В П.э. STATA реализованы несколько сотен статистических процедур, начиная с методов первичной обработки данных, факторного или кластерного анализа, моделей линейной регрессии, так и углубленного эконометрического анализа данных, в т.ч. обобщённых линейных моделей, моделей с бинарными и ограничен-

ными зависимыми переменными, временных рядов, панельных/повторных данных, эпидемиологических данных, стратифицированных выборок, максимизации функций правдоподобия, заданных пользователем, проверки гипотез и работы с оцененными моделями и др.

С

САМООРГАНИЗУЮЩИЕСЯ КАРТЫ КОХОНЕНА

эвристическая процедура нелинейного отображения многомерных данных на участок плоскости прямоугольной или квадратной формы (карту), разбитой на сегменты квадратной или шестиугольной формы, называемых ячейками. Каждая ячейка связана с центром кластера метода k-средних, где k – количество ячеек карты; в общем случае – это большое число, возможно, большее числа наблюдений.

Цель процедуры самоорганизации состоит в том, чтобы одновременно реализовать процедуру кластерного анализа и расположить ячейки на карте в таком порядке, при котором ячейки, соответствующие близким в исходном пространстве центрам кластеров, занимали соседние позиции на карте. Обычно алгоритм самоорганизации описывается в терминах обучения *нейронной сети* специального вида, однако возможны и другие алгоритмы решения этой оптимизационной задачи.

Для построения на карте кластеров задается порог близости. Если расстояние между центрами кластеров, соответствующих каждой паре соседних на карте ячеек больше порога, на карте между ячейками рисуется граница. В результате анализа всех пар соседних ячеек карта разбивается на связные области, каждая из которых, обычно окрашенная в свой цвет, образует отдельный кластер. Осн. топологическое свойство кластера –возможность построения пути между любыми двумя входящими в него ячейками через ячейки кластера. При увеличении значения порога близости формируется иерархия кластеров, аналогичная иерархическим кластеризациям, используемым в кластерном анализе.

Важнейшее преимущество самоорганизующихся карт – наглядность представления и интерпретации результатов кластеризации, возможность нанесения на карту дополнительной аналитической информации, а также отображения траекторий перемещения отдельных наблюдений, если данные меняются во времени. В последнем случае каждое новое наблюдение помешается в ту ячейку карты, для которой минимально расстояние в исходном пространстве между её центром и наблюдением.

Х

ХРАНИЛИЩЕ ДАННЫХ (ХД)

информационная система, ориентированная на поддержку процесса принятия управленческих решений в конкретной предметной области, интегрированная с *DM-методами* и обеспечивающая целостность, достоверность, непротиворечивость, неизменяемость и хронологию данных, а также высокую скорость выполнения аналитических запросов.

Важнейшая особенность предметно-ориентированного ХД – семантическое единство используемой терминологии, обеспечиваемое лингвистическими средствами ХД на основе метаданных предметного уровня. Это даёт пользователю возможность оперировать терминами предметной области для формирования аналитических запросов и не задумываться о механизме извлечения информации из ХД.

Требование непротиворечивости и целостности означает, что должна быть обеспечена возможность загрузки в ХД информации из внутренних и внешних источников, поддерживающие различные форматы данных и созданные в различных приложениях. Загружаемые данные должны быть преобразованы в единый формат, определяемый метаданными технического уровня, проверены на целостность, достоверность и непротиворечивость, и агрегированы до необходимого уровня обобщения программными и лингвистическими средствами ХД в соответствии с метаданными предметного уровня.

Требование неизменяемости предполагает, что данные после загрузки не должны как-либо изменяться.

Требование обеспечения хронологии означает соблюдение порядка следования записей в ХД в соответствии с обязательным атрибутом «Время».

Осн. цели создания ХД в орг-ции: своевременное обеспечение руководителей и аналитиков информацией, необходимой для выработки обоснованных и качественных управленческих решений; создание единой информационной модели предметной области, в которой функционирует орг-ция; создание интегрированного источника данных, предоставляющего удобный доступ к разнородной информации и гарантирующего получение одинаковых ответов на одинаковые запросы из различных аналитических приложений.

При разработке ХД необходимо решить осн. задачи: выбрать структуру хранения данных, обеспечивающую высокую скорость выполнения запросов при ограничениях на объем памяти; выполнить первоначальное заполнение и организовать последующее пополнение ХД; разработать единую методику работы с ХД и удобный пользовательский интерфейс.

Аналитические приложения ХД предоставляют пользователями возможность формировать запросы и получать по ним данные из хранилища. Выделяются три осн. формы аналитических запросов к ХД: «произвольные» аналитическое запросы в режиме реального времени (осн. функция OLAP-системы); регулярные аналитические отчёты стандартной формы, выполняемые с определённой периодичностью; выполнение процедур *DM-методов* с целью выявления скрытых закономерностей, структур и объектов, построения моделей, прогнозов и т.п. на основе информации, хранящейся в ХД.

Наиболее распространённая структура хранения данных в ХД базируется на концепции «многомерных кубов». Базовые понятия этой концепции – понятия измерений и фактов. Измерения – совокупность атрибутов, существенных характеристик объектов (процессов и явлений) предметной области, заданная при формировании информационной модели и описанная в метаданных ХД. Предполагается, что измерения (атрибуты) являются дискретными и принимают конечное число значений (категорий). В число измерений обязательно входит атрибут «Время», который также – дискретен. Начальный отсчёт атрибута «Время» соответствует времени поступления первой записи в ХД, последний – текущему календарному времени. Уровень детализации измерений определяется информационной моделью. Напр., если принят уровень детализации, соответствующий календарным суткам, то вместо атрибута «Время» в число измерений многомерного ХД целесообразно ввести атрибут «Дата». Для количественных характеристик, напр., атрибута «Возраст», порядок дискретизации определяется путём задания числовых интервалов, границы которых и наименования соответствующих категорий (напр., «Старший возраст») сохраняются в метаданных. Некоторые измерения могут формироваться путём агрегирования значений других измерений. Так, атрибут «Область» формируется путём агрегирования соответствующих категорий атрибута «Р-он» по правилам (в данном случае – спискам р-нов, входящих в те или иные области), записанным в метаданных. Из приведённых примеров видно, какую важную роль в функционировании ХД играют метаданные и средства их обработки.

Факты – данные, количественно описывающие события или процессы , протекающие в предметной области. Напр., это количество единиц товара, число покупателей, сумма продаж, факт состоявшейся или несостоявшейся продажи данного вида товара (в кодировке (0,1)) и т.п.

Многомерное хранилище можно рассматривать как многомерный «куб» (точнее «параллепипед»), осями которого служат дискретные измерения, а в ячейках куба записаны факты. Напр., в трёхмерном хранилище сумм продаж с измерениями «Орг-ция» × «Товар» × «Дата» ячейки будут содержать суммарную стоимость продаж определённых видов товаров определённым орг-циям в определённые дни. Заполнение очередного «слоя» выполняется программными средствами пополнения ХД с момента начала временного отсчёта и завершается

по истечении заданного времени. После этого никакие другие изменения в ХД не внесятся. Напр., факт возврата товара будет представлен в хранилище путём указания отрицательной стоимости в соответствующей ячейке другого, более позднего слоя.

Осн. преимущество многомерной модели ХД состоит в том, что агрегированные данные помещаются в многомерную табл. заранее и, как правило, хранятся в общем поле памяти вычислительной системы. Это позволяет решать «произвольные» аналитические задачи в режиме реального времени. При этом постановка «произвольных» аналитических задач ограничивается включёнными в информационную модель характеристиками объектов предметной области, уровнем заданной в метаданных детализации значений характеристик и фактами, которые накапливаются в ячейках табл. При необходимости можно, конечно, увеличить число характеристик и повысить уровень детализации их значений, но это может привести к катастрофическому увеличению объёма требуемой памяти при условии, что большинство ячеек хранилища окажутся пустыми. Поэтому при построении информационной модели предметной области проблеме избыточности уделяется большое внимание.

Ситуация, когда пользователю для анализа необходима вся информация, содержащаяся в ХД, возникает редко. В большинстве случаев пользователь используют профильную информацию, касающуюся только отдельных аспектов общей предметной области. Обычно это связано с выполнением процедур DM-методов. Объём требуемой тематической информации невелик по сравнению с объёмом ХД и поэтому создание, хранение и актуализация (т.е. пополнение новой информацией) такого «локального», пользовательского многомерного хранилища может обеспечиваться программными средствами осн. ХД. Такие пользовательские многомерные массивы получили название *витрин данных*. Преимущества при использовании витрин данных: близость к конечному пользователю; содержание данных тематически ориентировано на пользователя; разграничение прав доступа пользователей к общей информации; повышение достоверности решения прикладных аналитических задач за счёт средств обеспечения достоверности, целостности и непротиворечивости данных осн. хранилища.

Э

ЭКСПЕРТНАЯ СИСТЕМА (ЭС)

информационная система, оперирующая со знаниями, накопленными в результате практической деятельности в определённой предметной области, с целью выработки рекомендаций или решения проблем.

Знания в ЭС представлены на машинно-реализуемом языке представления знаний и хранятся в *базе знаний*. Осн. функции ЭС: извлечение знаний – построение информационной модели предметной области путём передачи опыта решения проблем от источника знаний (эксперта, специальной литературы) к приёмнику знаний (инженеру по знаниям) для содержательного описания осн. понятий предметной области и взаимосвязей между ними в виде графиков, диаграмм, табл. и текстов в соответствии с выбранной моделью представления знаний (это, напр., продукционные модели, семантические сети, фреймы, модели *нечёткой логики* и др.); представление знаний – описание приобретённых знаний на языке представления знаний и запись их в базе знаний; управление базой знаний – интерпретация запросов пользователей ЭС и организация доступа к знаниям, а также модификация и пополнение базы знаний по мере функционирования ЭС; разъяснение предлагаемых решений – детальное представление пользователю всего процесса получения решения из баз знаний для оценки его корректности.

Подраздел 2.4. Актуарная математика и актуарные расчёты

А

АКТУАРИЙ

(от лат. actuarius – скорописец) – специалист в области *актуарных расчётов*, экономики и финансов, имеющий квалификационный аттестат и решающий задачи финансовой безопасности, оценки рисков и разработки математически обоснованных страховых, финансовых, инвестиционных, социальных и пенсионных схем. Наибольшее распространение имеет специализация актуарий страховой, который профессионально подготовлен и обучен математическим аспектам страхования и занимается разработкой методологии расчёта *страховых тарифов* и *страховых резервов* страховщика, оценкой его инвестиционных проектов, обеспечением финансовой устойчивости страховой компании. В инвестиционном бизнесе в задачи А. входит разработка и применение моделей оценки рисковых инструментов, расчёт резервов инвестиционного фонда (в т.ч. обязательных законодательно).

Специальность А. возникла в 18 в. в связи с развитием страхового дела в европейских странах. Начало этой профессии связывают с 1756, когда член Королевского общества Великобритании Джеймс Додсон представил табл. премий страхования жизни, после того как ему самому отказали в таком страховании из-за его возраста. Деятельность А. оказала определённое влияние на становление демографической статистики, особенно на измерение смертности и разработку табл. смертности. Долгие годы гл. задачей А. считалась экспертиза рисков и случайностей в страховании жизни, но вследствие развития страхования, глубокого проникновения всех финансовых структур друг в друга, актуарная профессия стала восприниматься более широко – как эксперта по финансовой безопасности в решении сложных задач в области бизнеса, финансов, страхования, социальной сферы и демографии.

А., т.о., должен быть уникальным специалистом, имеющим высокий уровень знаний на стыке нескольких научных дисциплин – математики, связанной с расчётами рисков (теории вероятностей, математической статистики, теории случайных процессов, актуарной и финансовой математики, эконометрики), экономики, права, информационных систем и т.п. По сути, от навыков и знаний А. зависит устойчивость всей страховой системы в целом.

Круг осн. задач страховых А. в совр. экономике весьма широк: исчисление и группировка рисков в рамках страховой совокупности (классификация рисков для создания однородных групп); оценка и управление рисками (в т.ч. с использованием *андеррайтинга*); оценка частот страховых событий, расчёт статистической вероятности наступления страховых случаев; определение распределения ущерба в случае наступления страхового события, как по отдельным рискам, так и по портфелю в целом; всесторонний анализ финансовой устойчивости и платежеспособности компании, убыточности страхового портфеля компании в разрезе продуктов, филиалов/агентств, линий продаж и т.д.; установление тарифных ставок по каждому виду страхования с учётом долгосрочного и краткосрочного характера их проведения; математическое обоснование расходов на ведение дела, расчёт себестоимости страховых услуг; оценка страховых резервов компании, достаточных для исполнения обязательств по страхованию, сострахованию, перестрахованию, взаимному страхованию с учётом конкретных видов страхования и действующего законодательства; анализ адекватности резервов убытков; определение величины и структуры собственных средств страховщика (включая уставный капитал, резервный капитал, добавочный капитал, нераспределённую прибыль), обеспечивающих выживание (не разорение) компании с определённой надёжностью; анализ возможности повышения устойчивости компании с помощью перестрахования и расчёт платы за перестрахование при различных условиях договора о перестраховании; оптимизация перестраховочной защиты; оценка рисков инвестирования и иных способов размещения собранных страховщиком взносов; оценка положения страховой компании на страховом рынке и, в

зависимости от ситуации, формулирование подтвержденных расчётами рекомендаций по укреплению позиций компании и т.п.

Учитывая важность задач, поставленных перед А., во многих странах деятельность А. регулируется на гос. уровне. Страховое законодательство ряда стран требует наличия сертификата (аттестата) А., которым удостоверяется уровень профессиональных знаний специалиста в данной области и разрешается профессиональный консалтинг и сотрудничество со страховыми компаниями. Сертификат выдается после успешной сдачи квалификационного экзамена кандидатом в национальной ассоциации А. и/или в Лондонском ин-те А. (конвертируемый диплом). Список осн. предметов, утверждённый Междунар. актуарной ассоциацией включает: финансовую математику; теорию вероятностей и математическую статистику; экономику; бухгалтерский учёт; моделирование; статистические методы; актуарную математику; страхование жизни; общее страхование; пенсии; финансирование медицины; управление инвестициями и активами; основы актуарного управления; профессионализм.

Обучение А. считается одним из самых дорогостоящих в мире, но полученная профессия относится за рубежом к одним из самых высокооплачиваемых и престижных. В соответствии с рейтингом, составленным исследовательской компаний CareerCast США, профессия А. в 2010 была признана лучшей и наиболее перспективной профессией среди 200 обследованных.

Существует множество актуарных национальных орг-ций, самыми известные и влиятельные – Ин-тут (1848) и Ф-т (1856) А. Великобритании (Faculty and Institute of Actuaries, Американское общество А. (Society of Actuaries, 1949), Американское общество А. в страховании ином, чем страхование жизни (Casualty Actuarial Society, 1914) и несколько междунар. ассоциаций А.

В 1895 национальные профессиональные общества Бельгии, Франции, Германии, Великобритании и США организовали Междунар. актуарную ассоциацию (International Actuarial Association, IAA), базирующуюся в Брюсселе. В 1998 ассоциация была реформирована введением новой конституции и реструктурирована в 2010. Междунар. актуарная ассоциация – мощная междунар. структура, объединяющая профессиональные актуарные орг-ции (в 2010 – 63 действительных и 22 ассоциированных члена), представляющих более 50 тыс. А. из более чем 100 стран мира. Каждые четыре года IAA проводит свои Конгрессы (очередной, 29-ой Междунар. Конгресс А. был проведён в марте 2010 в Кейптауне, ЮАР). Одна из традиций междунар. конгрессов А. состоит в том, чтобы дать возможность всем национальным обществам информировать актуарное сообщество о своем устройстве, функционировании, актуальных проблемах актуарной науки и практики, представить интересные статистические данные. Конгрессы включают большую научную программу, привлекающую как представителей чистой науки, так и её приложений.

IAA имеет несколько профессиональных секций – по финансовым рискам (AFIR), страхованию жизни (IAALS), здоровья (IAAHS), пенсий (PBSS) и т.д. Первая и наиболее известная из этих секций – ASTIN (Actuarial STudies In Non-life insurance), созданная в 1957 и объединяющая А., занимающихся вопросами рискового страхования (не-жизни), издающая научный бюллетень и проводящая регулярные коллоквиумы.

Россия имеет давние традиции страхования – первое рос. страховое общество было создано ещё в 1765. К концу 19 в. сложился весьма развитый (по тем временам) страховой рынок, в котором профессия А. была весьма востребована.

После социалистической революции 1917 потребность в А. отпала, т.к. страхование стало гос. монополией. Актуарная наука в России не развивалась почти 70 лет, т.е. фактически осталась в дореволюционном состоянии. Только после начала рыночных реформ в 1988 и формирования совр. страхового рынка РФ профессия А. начала сложный путь возрождения и становления. В 1994 было учреждено Общество А. и в МГУ состоялся первый съезд А. России.

Первым президентом созданного Общества А. стал член-корр. РАН *Ширяев А.Н.* Гильдия А., зарегистрированная 22 окт. 2002, – правопреемник Общества А. В качестве его уставной деятельности была определена постановка серьезной совр. системы актуарного образования и формирования соответствующей инфраструктуры актуарной науки и практики. С 2006 Гильдия –ассоциированный членом Междунар. актуарной ассоциации, а 4 нояб. 2008 Гильдия А. стала действительным членом IAA, что говорит о выходе рос. актуарной науки на междунар. уровень. Существует ещё несколько рос. актуарных орг-ций – Независимое актуарное общество (НОА), Независимый актуарный информационно-аналитический центр (АНО НААЦ) и др.

Законодательная база в РФ в отношении А. только формируется. Федеральным законом РФ от 10.12.2003 №172-ФЗ были внесены изменения в закон «Об организации страхового дела в РФ» и введена ст. 4.1, где страховые А. признаются участниками страховых отношений, субъектами страхового дела. В 2006 – 07 вступили в силу положения об обязательной аттестации и проведении квалификационных экзаменов страховых А., и обязанности страховщиков ежегодно предоставлять в Федеральную службу страхового надзора (ФССН) актуарные заключения. Наконец, в янв. 2010 Гос. Думой РФ в первом чтении был принят проект Федерального закона № 445108-4 "Об актуарной деятельности в РФ", который направлен на создание правовой основы для осуществления актуарной деятельности в РФ и введение института саморегулирования А. Принятие законопроекта позволит установить эффективное нормативное регулирование актуарной деятельности и достичь гармонизации рос. законодательства об актуарной деятельности с междунар. законодательной практикой в этой области.

АКТУАРНЫЕ РАСЧЁТЫ

совокупность математических, статистических, финансовых, демографических и экономических методов, используемых при оценке финансовых взаимоотношений сторон в различных видах финансовой деятельности, прежде всего, в страховании. Осн. задача А.р. – определение размеров *страховых тарифов* и *резервов*, обеспечивающих безубыточность страховых операций.

А.р., как и страхование в целом, делятся на два раздела, существенно отличных друг от друга: *А.р. в страховании жизни* (life insurance) и *А.р. в страховании ином, чем страхование жизни (*рисковых видах страхования или страхования не-жизни (non-life insurance)). В Великобритании для страхования жизни принят термин «assurance», а термин «insurance» используется в страховании не-жизни. Вследствие значительных отличий в методологии А.р. жизни и не-жизни, *актуарии* обычно также имеют соответствующие специализации. Регулирование страховой деятельности в зарубежных странах основано на концепции специализации страховых компаний, которая предполагает либо полное разделение страховых компаний, специализирующихся на страховании жизни и на прочих видах страхования, либо, по крайней мере, разделение движения всех средств и формирующихся фондов. Требования к специализации страховых компаний сформулированы в документах Междунар. ассоциации страхового надзора и законе РФ.

Зародившись в 17 веке, А.р. претерпели серьезные изменения. А.р. становятся все более сложной и многогранной сферой науки, сочетающей в себе сложнейшие совр. методы и достижения, круг задач которой постоянно расширяется. Одна из последних задач А.р. – создание системы управления рисками (риск-менеджмента) компании (Enterprise Risk Management, ERM). Управление рисками компании – это единая систематическая оценка и контроль всех рисков, возникающих в работе страховой орг-ции, их взаимосвязей и взаимозависимостей, в их совокупности, включая не только непосредственно страховые риски, но и финансовые, операционные и стратегические риски. Финансовые риски, в отличие от большинства страховых, являются коррелированными, непрерывными и требуют стохастического подхода для адекватной оценки. Для это-

го используются методы управления финансовыми рисками, включая производные инструменты, такие как форварды, фьючерсы, опционы и свопы, причём с точки зрения их взаимосвязи со всеми остальными рисками компании. Это позволяет перераспределять риски и активы компании, вырабатывать перестраховочные стратегии, оптимизировать структуру и размер принимаемых страховых рисков и инвестиционных вложений, моделировать влияние на развитие компании и её прибыль различных событий и стратегических решений.

АКТУАРНЫЕ РАСЧЁТЫ В СТРАХОВАНИИ ЖИЗНИ

актуарные расчёты, обеспечивающие договоры страхования жизни. Исторически начали развиваться первыми и значительно раньше актуарных расчётов в рисковом страховании. Основы теории *актуарных расчетов* как особой отрасли науки были заложены в 17 в. Математика страхования жизни имеет следующие отличительные особенности. Страховая компания осуществляет выплаты по договору страхования на случай смерти, дожития или смешанного страхования только один раз, и размер выплат заранее определён страховой суммой договора, поэтому случайным, как правило, является только момент выплаты – напр., смерть застрахованного. Вследствие этого А.р. в страховании жизни основаны на методах демографической статистики – табл. смертности, функциях дожития и т.п. На основе статистического наблюдения над смертностью нас. вычисляются вероятности дожития и смерти для лиц разных возрастов и строятся табл. смертности или функции дожития, которые характеризуют закономерность изменения под влиянием возраста численности определённой совокупности людей. Эти табл. и функции используются для расчёта *тарифных ставок* по страхованию жизни и пенсий для лиц каждого конкретного возраста. Страхователей, заключающих договоры страхования жизни, классифицируют, как правило, по возрасту, полу и состоянию здоровья, т.к. эти факторы существенно влияют на *страховой тариф*. Кроме того, А.р. широко используют финансовую математику, т.к. договоры страхования жизни в своей основе имеют долгосрочный характер, и страховщик имеет возможность инвестировать средства, собранные в виде *страховых премий*. Помимо расчёта *страховых тарифов*, А.р. в накопительном страховании жизни направлены на решение задач оценки *страховых резервов*, учёта инфляции, возможности участия страхователей в прибыли страховщика (бонусы), индексации *страховых взносов* и страховых выплат, долгосрочного прогнозирования денежных потоков, разработку схем добровольного негосударственного пенсионного обеспечения, размеров подлежащих выплате выкупных, редуцированных страховых сумм, ссуд и пр. Классическая система актуарных обозначений в страховании жизни весьма сложна, представляет собой междунар. символику условных обозначений, которая была принята в 1898 в Лондоне на втором междунар. конгрессе *актуариев* и до сих пор является общепринятой в мировой актуарной литературе. Наиболее известная специализированная актуарная орг-ция, объединяющая актуариев страхования жизни – профессиональная секция по страхованию жизни (IAALS) Междунар. актуарной ассоциации (International Actuarial Association, IAA).

АКТУАРНЫЕ РАСЧЁТЫ В СТРАХОВАНИИ ИНОМ, ЧЕМ СТРАХОВАНИЕ ЖИЗНИ

актуарные расчёты, обеспечивающие договоры страхования иного, чем страхование жизни. А.р. в рисковом страховании (не-жизни) были развиты значительно позднее, чем математическая теория страхования жизни, и имеют существенно более сложный математико-статистический аппарат. Развитие математики и статистики достигло нужного уровня только к нач. 20 в., когда и началось развитие математики рискового страхования.

В рисковых видах страхования, в отличие от страхования жизни, выплаты обычно могут осуществляться неоднократно, и изучение распределения числа страховых выплат в одном договоре страхования является одной из акту-

арных задач. Кроме того, в рисковых видах страхования размер убытков, как правило, является случайной величиной, и страховая компания заранее не знает величину потерь в случае наступления страхового случая. Так что анализ распределения размера страховой выплаты при наступлении одного страхового случая и совокупного годового убытка по портфелю договоров в целом, является одной из важных и сложных задач *актуариев*. Статистические задачи, возникающие при оценке параметров в рисковом страховании, вычислительно значительно сложнее в силу необходимости учёта быстрых изменений экономических условий.

Рисковые договоры страхования заключаются обычно на год, в отличие от договоров страхования жизни, имеющих, как правило, долгосрочный характер. Если в страховании жизни премии можно разделить на рисковую и накопительную компоненту, создающую возможность инвестирования средств, то в рисковых видах страхования накопительную компоненту имеют в гораздо меньшем масштабе и только за счёт разрыва по времени сбора премий и выплат по наступившим убыткам.

Страхователей, заключающих договоры рискового страхования, значительно сложнее разделить на априорно однородные классы, чем в договорах страхования жизни. Кроме того, отбор тарифных факторов, влияющих на величину *страхового тарифа*, осуществляют из гораздо более широкого круга параметров, что представляет собой самостоятельную задачу в актуарных расчётах.

Расчёт *страховых резервов* в рисковом страховании также имеет существенные отличия и значительно сложнее вследствие большей неопределённости и случайности размера будущих платежей страховой компании.

Наиболее влиятельные специализированные актуарные орг-ции, занимающиеся рисковым страхованием – профессиональная секция Междунар. актуарной ассоциации ASTIN (Actuarial STudies In Non-life insurance), созданная в 1957, и Американское общество актуариев в страховании ином, чем страхование жизни (Casualty Actuarial Society, 1914).

АНДЕРРАЙТЕР

см. в ст. Андеррайтинг

АНДЕРРАЙТИНГ

(от англ. underwriting – «подписание под» чем-либо, под какими-либо условиями) – принятие решения о страховании потенциального риска. А. связан с процессом отбора и классификации рисков с точки зрения возможности принятия их на страхование, а также соответствующих ставок премии или отказа от принятия на страхование рисков, не соответствующих квалификационным требованиям. Цель А. при страховании – обеспечение заданных показателей убыточности вида страхования и страхового портфеля в целом посредством селекции рисков и выбора условий страхования и страхового покрытия объектов страхования. Термин зародился в страховой практике Лондонского «Ллойда» (Lloyd’s of London), где каждый, желавший принять на себя часть риска, подписывался своим именем под описанием риска.

Лицо, осуществляющее А. – андеррайтер – высококвалифицированный специалист в области страхования, обладающий специальными знаниями, опытом и репутацией, достаточными для вынесения решения по принятию тех или иных рисков на страхование. В круг обязанностей андеррайтера входит оценка качества риска и определение ставки страховой премии, адекватной принятию всего или части риска, конкретных условий договора страхования, а также заключение о возможности (или невозможности) заключения договора страхования на определенных условиях. Андеррайтер может выполнять функции сюрвейера (оценщика рисков) – представителя страховщика, осуществляющего осмотр и оценку имущества, принимаемого на страхование.

Условия А. обычно выражаются в андеррайтерской политике, при помощи которой страховая компания рассматривает новые объекты страхования и риски и приходит к выводу о приня-

тии или отклонении предложенного дела. Андеррайтерская политика предусматривает, в частности, перечень объектов (рисков) с указанием лимитов убытков по ним, которые страховщик склонен принять, и второй перечень с объектами (рисками), которые страховщик, исходя из своего опыта, не принимает.

В финансовой деятельности наряду со страховым А., выделяют А. на рынке ценных бумаг, под которым понимается деятельность инвестиционных посредников по гарантированному размещению займа или выпуску ценных бумаг на первичном рынке, и банковский А., заключающийся в оценке риска заёмщика и принятии решения о выдаче кредита.

Б

БРУТТО-ПРЕМИЯ

см. в ст. Страховой взнос (страховой платёж)

БРУТТО-СТАВКА

см. в ст. Страховой тариф (тарифная ставка)

К

КОММУТАЦИОННЫЕ ФУНКЦИИ

(коммутационные числа, коммутационные таблицы) – специальные функции, разработанные в *актуарных расчётах* страхования жизни для уменьшения трудоёмкости выполнения вычислений и сокращения записи формул страховых аннуитетов и премий. Обязательные элементы К.ф. – два рода величин: показатели смертности нас. (табл. смертности или функции дожития) и коэффициент дисконтирования (дисконтный множитель) по сложной ставке процентов, отражающий изменение цены денег:

$$v^n = \frac{1}{(1+i)^n},$$

где i – ставка сложных процентов (техническая норма доходности, заложенная в расчеты); n – срок, за который производится дисконтирование.

Первая известная табл. К.ф. опубликована в работе Вильяма Дэйла в 1772, а название "коммутационные табл., было введено в актуарную науку Огастесом де Морганом в 1840. К.ф. делятся на две группы. В основу первых положены числа доживающих l_x, в основу вторых – числа умерших d_x.

Осн. в первой группе являются коммутационные числа D_x и N_x. Величина D_x – дисконтированное число лиц данной совокупности, доживающих до возраста x лет:

$$D_x = l_x \cdot v^x,$$

где l_x – число лиц данной совокупности, доживших до возраста x, определяемое по табл. смертности; v – коэффициент дисконтирования по используемой процентной ставке.

Сумма дисконтированных чисел всех доживающих и переживающих x лет:

$$N_x = \sum_{j=x}^{\omega} D_j .$$

где ω – предельный возраст, до которого составлена табл. смертности.

Ещё одну группу коммутационных чисел составляют реже используемые величины:

$$S_x = \sum_{j=x}^{\omega} N_j = \sum_{j=x}^{\omega} \left(\sum_{j=x}^{\omega} D_j \right).$$

Наиболее важные представители К.ф. второй группы – функции C_x и M_x. Величина C_x – дисконтированное число умерших в возрасте от x до x+1 лет; вычисляется непосредственно: $C_x = d_x \cdot v^{x+1}$, где d_x – число лиц данной совокупности, умерших в возрасте от x до x+1, определяемое по табл. смертности.

Сумма дисконтированных чисел умирающих в возрасте x и выше:

$$M_x = \sum_{j=x}^{\omega} C_j .$$

Реже используемая коммутационная функция R_x – сумма:

$$R_x = \sum_{j=x}^{\omega} M_j = \sum_{j=x}^{\omega} \left(\sum_{j=x}^{\omega} C_j \right).$$

К.ф. не интерпретируют содержательно, их следует воспринимать как чисто технические, вспомогательные величины. К.ф. и их табл. играли большую роль в актуарных расчётах до появления мощных компьютеров, поскольку значительно сокращали объём вычислительной

работы. Причем, табл. этих функций приводились для ограниченного диапазона процентных ставок. Сейчас их роль значительно уменьшилась, поскольку их непосредственное вычисление на компьютере занимает намного меньше времени, чем поиск в табл. Тем не менее, знание К.ф. по-прежнему является обязательным в образовании *актуария.* К тому же по-прежнему актуальна удобная краткая запись формул нетто-премий договоров страхования жизни и рент с помощью К.ф. Напр., единовременная нетто-ставка при заключении пожизненного договора страхования на случай смерти для лица возраста x, при котором выплаты страхового обеспечения производятся в конце года смерти, равна:

$$A_x = \frac{\nu \cdot d_x + \nu^2 \cdot d_{x+1} + ... + \nu^{\omega-x+1} \cdot d_\omega}{l_x},$$

А через К.ф. эта же формула записывается кратко:

$$A_x = \frac{M_x}{D_x}.$$

В непрерывном случае – при использовании функций дожития вместо табл. смертности и выплат страхового возмещения в точные моменты времени – К.ф. имеют аналогичный вид, но с использованием аппарата интегрирования, напр.:

$$\overline{N_x} = \int_x^\infty D_t dt,$$

где $D_t = \nu^t \cdot l_t$ – дисконтированное число лиц, доживающих до точного возраста t.

Н

НАГРУЗКА

часть *брутто-премии* (*страховой премии*), поступающая в полное распоряжение страховщика (в отличие от *нетто-премии*, остающейся в собственности страхователей) и предназначенная на финансирование самого процесса страхования и формирование прибыли.

Состав расходов, относимых к Н., может отличаться в зависимости от вида страхования и определяется *актуариями* конкретной страховой компании.

Осн. компоненты Н.: расходы на ведение дела: организационные расходы, связанные с учреждением страховой компании; аквизиционные расходы – производственные расходы страховой компании, связанные с привлечением новых страхователей и заключением новых договоров (вознаграждение страховых агентов, затраты на рекламу, расходы по изготовлению, оформлению и регистрации страховых полисов, оплата консультаций, например, медицинских при страховании жизни); инкассационные расходы, связанные с обслуживанием денежной наличности; ликвидационные расходы – расходы по ликвидации ущерба, вызванного страховым случаем; административные расходы – расходы по управлению страховой компанией (оплата труда, арендная плата, коммунальные платежи); резервные фонды, обеспечивающие безубыточную деятельность страховой компании в случае наступления неблагоприятных форс-мажорных ситуаций, главным образом - резерв предупредительных мероприятий (РПМ) – часть тарифа, из которой можно было бы сформировать средства для профилактики случайных убытков, процентная доля которой согласовывается с ФССН, а также иных резервов по согласованию с ФССН; фонд прибыли для выплаты дивидендов акционерам.

Н. содержит составляющие, зависящие от размера нетто-премий (прибыль, резервы и т.д.) и не зависящие от него (аренда помещения, заработная плата персонала, затраты на оформление договора и т.д.).

С количественной точки зрения при расчете брутто-премии важна, прежде всего, связь величины различных компонент нагрузки с величиной страховой суммы, длительностью контрактов, периодичностью выплат по ним. Аквизиционные расходы исчисляются, как правило, пропорционально страховым суммам. Инкасационные – пропорционально брутто-премиям. Адм. расходы не связаны непосредственно с размерами страховых сумм или премий – этот вид расходов зависит от общего числа заключенных договоров и их сроков,

уровня занятости в данной страховой компании.

Как правило, Н. составляет фиксированный процент от брутто-премии.

НЕТТО-ПРЕМИЯ

осн. часть *брутто-премии* (*страховой премии*), определяемая исключительно характеристиками взятого на страхование риска. Н.-п. или чистая премия отражает рассчитанную *актуариями* цену страхового риска и предназначена для возмещения убытков при наступлении страховых случаев. Сумма всех нетто-премий, собранных страховой компанией, составляет страховой фонд и идёт на формирование *страховых резервов*, предназначенных для осуществления страховых выплат. Т.о. Н.-п. – часть страховой премии, остающаяся в групповой собственности страхователей (в отличие от второй части брутто-премии – *нагрузки*, поступающей в распоряжение страховщика).

В *актуарных расчетах страхования жизни* Н.-п. рассчитываются, исходя из табл. смертности или функций дожития и нормы доходности.

В *актуарных расчетах страхования, иного чем страхование жизни* (рисковых видах страхования) Н.-п. определяются вероятностно-статистическими методами оценки риска как случайной величины.

Н.-п. обычно представляется суммой двух составляющих: *рисковой премии,* определяемой как среднее значение ущерба в одном договоре страхования и *рисковой надбавки*, предназначенной на обеспечение выплат, превышающих ожидаемое среднее значение убытка.

Р

РИСКОВАЯ НАДБАВКА

часть *нетто-премии*, являющейся в свою очередь осн. составляющей *брутто-премии* (*страховой премии*), необходимая для того, чтобы обеспечить безубыточность страховых операций, если фактическая сумма выплат будет выше математического ожидания ущерба. Необходимость её формирования обусловлена тем, что на практике страховые портфели всегда ограничены по объёму и остаются зависимыми от случайности. Р.н. называют также гарантийной, потому что она поддерживает капитал, обеспечивающий надежность и гарантирует повышение устойчивости страховой компании.

Отношение Р.н. к величине *рисковой премии* (второй, осн. составляющей нетто-премии) называют относительной Р.н., она показывает, какую долю составляет Р.н. в рисковой премии, и может выражаться в процентах.

В *актуарных расчётах* имеется большое число методов расчёта Р.н., как правило, связанных с характеристиками разброса – дисперсией или средним квадратическим отклонением – возможного ущерба по договору страхования убытка. Один из наиболее простых и используемых в актуарных расчётах квантильный принцип (принцип доверительного оценивания) расчёта Р.н. состоит в том, что задается уровень надёжности (неразорения страховой компании), и для заданного уровня надежности рассчитывается правая граница доверительного интервала величины совокупного ущерба во всем страховом портфеле, т.е. задается вероятность того, что суммарные выплаты не превысят определенной величины. Исходя из значения правой границы доверительного интервала, определяют Р.н.

При увеличении объёма портфеля договоров Р.н., приходящаяся на один договор, обычно уменьшается. Этим объясняется тот факт, что крупные страховые компании, собирающие портфели договоров большого объёма, могут предложить своим клиентам более низкие тарифы, повысив свою конкурентоспособность, при этом обеспечивая высокий уровень надёжности (вероятности неразорения).

РИСКОВАЯ ПРЕМИЯ

осн. часть *нетто-премии*, являющейся в свою очередь составляющей *брутто-премии* (*страховой премии*), определяется как средняя ожидаемая величина ущерба. Р.п. согласно закону больших чисел теории вероятностей, фундаментальному закону страхования, вычисляется как математическое ожидание убытка, насту-

пающего в одном договоре страхования, и отражает осн. принцип *актуарных расчётов* – принцип эквивалентности обязательств страховщика и страхователя, который выражается в равенстве математических ожиданий двух величин: суммы всех страховых взносов и суммы всех страховых возмещений.

При увеличении объёма портфеля договоров Р.п., приходящаяся на один договор, не изменяется (в отличие от *рисковой надбавки*, второй составляющей нетто-премии).

С

СТРАХОВАЯ НАГРУЗКА

см. в ст. Страховой взнос (страховой платёж)

СТРАХОВАЯ ПРЕМИЯ

в *актуарных расчётах* брутто-премия, денежная сумма, которую страхователь обязан уплатить страховщику за страховую защиту передаваемого ему объекта страхования от характерных рисков. С.п. определяется в соответствие с законом (для обязательных видов страхования) или договором страхования (для добровольных), исходя из *страхового тарифа* для данного страхового риска (объекта страхования), страховой суммы, срока страхования и некоторых других факторов. С.п. вносится страхователем при вступлении в страховые отношения со страховщиком единовременно или частями в течение некоторого (или всего) срока страхования.

С.п. – рассчитанная методами актуарных расчётов брутто-премия, которая скорректирована и согласована со страхователем, с учётом ситуации на рынке, соотношения спроса и предложения на страховом рынке услуг, конкурентоспособности компании и взаимоотношений с конкретным страхователем, и будет им оплачена. Брутто-премия по своей структуре равна сумме *нетто-премии*, предназначенной для формирования *страховых резервов* и осуществления страховых выплат, и *нагрузки,* обеспечивающей поступление средств для сопровождения процесса страхования и получения прибыли страховой компанией.

Брутто-премия равна произведению *страхового тарифа* (брутто-ставки) на страховую сумму данного риска согласно договору страхования.

Объём собранных С.п. от всех функционирующих страховщиков – один из важнейших показателей состояния страхового рынка.

СТРАХОВОЙ ВЗНОС (СТРАХОВОЙ ПЛАТЁЖ)

единовременно перечисляемая часть *страховой премии*. Если вся сумма выплачивается сразу за 1 раз, С.в. называется единовременной премией. Страховые платежи, которые выплачиваются с рассрочкой, некоторой периодичностью (напр., ежегодно, ежеквартально), называют периодическими премиями. При наличии изменения стоимости денег во времени (с использованием аппарата финансовой математики) и риска недополучения части платежей цена договора возрастает.

СТРАХОВОЙ ТАРИФ (ТАРИФНАЯ СТАВКА)

ставка *страховой премии* с единицы страховой суммы с учётом объекта страхования и характера страхового риска; в *актуарных расчётах* брутто-ставка. Это предполагаемая цена страхового товара, определяемая методами актуарных расчётов, на основании которой назначается плата за страхование. Аналогично структуре брутто-премии, брутто-ставка равна сумме нетто-ставки и нагрузки. Обычно исчисляется в процентах от страховой суммы. Расчёт С.т. по различным видам страхования – важнейшая задача актуарных расчётов, т.к. определяет финансовую устойчивость и конкурентоспособность страховщика.

Объём собранных страховых премий от всех функционирующих страховщиков – один из важнейших показателей состояния страхового рынка.

СТРАХОВЫЕ РЕЗЕРВЫ

создаваемые в обязательном порядке, определяемом законодательством, за счет поступления *страховых премий* денежные фонды, предназначенные для обеспечения исполнения обя-

зательств страховых компаний по договорам страхования, перестрахования, взаимного страхования. С.р. являются выраженной в денежной форме оценкой обязательств страховщика по обеспечению предстоящих страховых выплат и являются, т.о., одной из основных гарантий финансовой устойчивости страховщика. Потребность в их формировании обусловлена вероятностным характером страховых событий и неопределенностью момента наступления и размера ущерба, долгим периодом урегулирования убытков в некоторых видах страхования. Средства С.р. используются исключительно для осуществления страховых выплат при наступлении страховых случаев по договорам страхования, не подлежат изъятию в федеральный бюджет и бюджеты иных уровней бюджетной системы Российской Федерации и могут инвестироваться и размещаться страховщиками на условиях диверсификации, возвратности, прибыльности и ликвидности.

Расчет С.р. осуществляется *актуариями* страховых компаний методами *актуарных расчетов* и регулируется органами страхового регулирования – Министерством финансов РФ и ФССН. Состав, назначение и порядок формирования С.р., образуемых страховщиками для выполнения обязательств по договорам страхования различаются, как и актуарные расчеты в целом, в рисковых видах страхования (страхования, иного чем страхование жизни) (где они часто называются техническими резервами) и в страховании жизни (именуемые часто математическими резервами), определяются нормативно-техническими указаниями, утвержденными Министерством финансов в виде соответствующих документов, регламентирующих действия актуариев при формировании страховых и иных резервов. Размеры С.р. рассчитываются при определении финансового результата от страховой деятельности на отчетную дату. Отчет о С.р. предоставляется в Министерство Финансов РФ в составе годового бухгалтерского отчета.

С.р. по рисковым видам страхования согласно «Правилам формирования страховых резервов по видам страхования иным, чем страхование жизни» от 11 июня 2002 г. №51н, включают: резерв незаработанной премии (РНП); резервы убытков: резерв заявленных, но неурегулированных убытков (РЗУ); резерв произошедших, но незаявленных убытков (РПНУ); стабилизационный резерв (СР) (ранее резерв катастроф и резерв колебаний убыточности для разных видов страхования); иные страховые резервы.

Наибольшую сложность с точки зрения *актуарных расчетов* вызывает формирование резерва произошедших, но незаявленных убытков (РПНУ) – денежной оценки обязательств страховщика на отчетную дату по произошедшим, но незаявленным убыткам, включая расходы по урегулированию убытков, возникших в связи со страховыми случаями, происшедшими в отчетном или предшествующих ему периодах, о факте наступления которых в установленном законом или договором порядке не заявлено страховщику, как величины прогнозируемой и до конца неизвестной. В последние годы надзорные органы все большего числа стран требуют актуарного обоснования резерва позднего убытка. Поэтому современный этап развития актуарной математики характеризуется разработкой большого количества математических методов оценки РПНУ. Методы объединены общей идеей – спроецировать опыт прошлых лет событий на последующие годы событий (когда убыток произошел или должен быть учтен в бухгалтерской отчетности). Методы расчета РПНУ основаны на треугольниках выбывания, называемых в литературе также треугольниками развития, треугольниками убытков и др. Существует два основных вида данных, из которых может быть построен треугольник развития. В одном случае элементы треугольника Y_{ij} – сумма убытков, оплаченных в течение j периодов развития убытков (без каких-либо резервов заявленных убытков), а в другом – сумму убытков, произошедших в течение j периодов, куда включают и все осуществленные выплаты, а также и сформированные к этому моменту резервы заявленных убытков. Соответственно, y_{ij} представляют собой либо сумму выплат, произведенных в j-м году развития по убыткам i-го года события, либо сумму этих же выплат с добавлением сальдо всех изменений резервов заявленных

убытков. Треугольник развития может быть создан и из резервов заявленных убытков, количества убытков, среднего размера урегулированных или заявленных убытков, среднего изменения суммарного убытка и т.д.

В общем случае, статистические методы, лежащие в основе расчета РПНУ, предполагают наличие стабильной структуры урегулирования претензий в прошлом, а также сохранение этой стабильности в будущем. Наиболее известные методы оценки РПНУ - метод Борнхуеттера-Фергюсона (обязательный к применению российскими актуариями согласно «Правилам»), методы цепной лестницы; методы средней стоимости претензии; методы коэффициента убытков; смеси (комбинации методов) и др.

В состав страховых резервов по страхованию жизни согласно «Порядку формирования страховых резервов по страхованию жизни» от 9 апреля 2009 г. № 32н включаются следующие резервы: математический резерв; резерв расходов на обслуживание страховых обязательств; резерв выплат по заявленным, но неурегулированным страховым случаям; резерв выплат по произошедшим, но не заявленным страховым случаям; резерв дополнительных выплат (страховых бонусов); выравнивающий резерв.

Расчет страховых резервов является одной из самых сложных и важных задач *актуариев* страховой компании. От того, насколько правильно рассчитываются страховые резервы, как они учитывают неисполненные или исполненные не полностью обязательства, зависят финансовая устойчивость страховой организации, ее платежеспособность, возможность выполнить принятые перед страхователями обязательства по страховым выплатам.

СЮРВЕЙЕР

см. в ст. Андеррайтинг

Список литературы

Подраздел 2.1. Теория вероятностей и математическая статистика

Абдулгалимов А.М. Статистическое прогнозирование социально-экономических процессов. Махачкала, 1998.

Айвазян С.А., Мхитарян В.С. Прикладная статистика. Основы эконометрики. Т. 1: Теория вероятностей и прикладная статистика. М., 2001.

Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичной обработки данных. 1988.

Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. М., 1977.

Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. М., 1989.

Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. М., 1985.

Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы моделирования и первичная обработка данных. М., 1983.

Андронов А.М., Копытов Е.А., Гринглаз Л.Я. Теория вероятностей и математическая статистика. Спб., 2004.

Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ. М., 1985.

Арлей Н., Бух К. Р., Введение в теорию вероятностей и математическую статистику. М., 1951.

Атон Г. Анализ таблиц сопряжённости. М., 1982.

Афанасьев В.Н., Юзбашев М.М., Гуляева Т.И. Эконометрика, 2005.

Афифи А., Эйзен С., Статистический анализ. Подход с использованием ЭВМ. М., 1962.

Баврин, И.И. Теория вероятностей и математическая статистика: Учебник. М., 2005.

Банников В.А. Векторные модели авторегрессии и коррекции регрессионных остатков. Прикладная эконометрика. 2006.

Березин И.С., Жидков Н.П. Методы вычислений. Т. 1. М., 1966.

Благовещенский Ю.Н. Тайны корреляционных связей в статистике: М., 2009.

Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. Вып. 1. М., 1974.

Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М., 1983.

Борель Э. Вероятность и достоверность. М.: Государственное издательство физико-математической литературы, 1961.

Боровиков В.П., Боровиков И.П. Statistica – статистический анализ и обработка данных в среде Windows. 1997.

Боровков А.А. Математическая статистика. Новосибирск, 1997.

Браунли К.А. Статистическая теорема и методология в науке и технике. М., 1977.

Бусленко Н.П. Математическое моделирование производственных процессов на цифровых вычислительных машинах. М., 1964.

Бусленко Н.П. Метод статистического моделирования. М., 1970.

Бусленко Н.П., Голенко Д.И., Соболь И.М., Срагович В.Г., Шрейдер Ю.А. Метод статистических испытаний (метод Монте-Карло). М., 1962.

Бухштабер В.М., Маслов В. К. Задачи прикладной статистики как экстремальные задачи на нестандартных областях. Алгоритмическое и программное обеспечение прикладного статистического анализа. Ученые записки по статистике. Т. 36. М., 1980.

Бухштабер В.М., Маслов В. К. Факторный анализ и экстремальные задачи на многообразиях Грассмана. Математические методы решения экономических задач. № 7. М., 1977.

Вальд А. Последовательный анализ. М., 1960.

Вальд А. Статистические решающие функции. М., 1967.

Ван дер Варден. Математическая статистика. М., 1960.

Варюхин А.М., Панкина О.Ю., Яковлева А.В. Эконометрика, 2005.

Вентцель А.Д. Курс теории случайных процессов. М., 1996.

Вентцель Е.С. Овчаров Л.А. Теория вероятностей и её инженерные приложения. М., 2007.

Вентцель Е.С., Овчаров А.Л. Прикладные задачи теории вероятностей, М., 1983.

Вербик М. Путеводитель по современной эконометрике. М., 2008.

Вероятность и математическая статистика: Энциклопедия. Под ред. Ю.В. Прохорова. М., 2003.

Виленкин Н.Я. Комбинаторика. М., 1969.

Виленкин Н.Я. Популярная комбинаторика. М., 1975.

Вилкас Э.Й., Майминас Е.З. Решения: теория, информация, моделирование. М., 1981.

Вуколов, Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов Statistica и Excel. М., 2004.

Гантмахер Ф.Р. Теория матриц. М., 1966.

Гельфанд И. М. Лекции по линейной алгебре. М., 1971.

Герасимович А.И. Математическая статистика. Минск, 1983.

Гмурман В.Е. Теория вероятностей и математическая статистика. М., 2003.

Гнеденко Б.В. Курс теории вероятностей. Из-во Физико-математической литературы, М., 1961.

Гончаров В.Л. Теория интерполирования и приближения функций. М., 1954.

Де Гроот М. Оптимальные статистические решения. М., 1974.

Демидович Б.П., Марон И.А., Шувалова Э.З. Численные методы анализа. Приближение функций, дифференциальные и интегральные уравнения./ Под редакцией Б.П. Демидовича. М., 1967.

Джеффри Мур, Ларри Уэддерфорд и др. Экономическое моделирование в Microsoft Excel, М., 2004.

Дзядык В.К. Введение в теорию равномерного приближення функций полиномами. М., 1977.

Донелли-мл. Р., М., 2007.

Дубина А.Г., Орлова С.С., Шубина И.Ю., Хромов А.В. Excel для экономистов и менеджеров. СПб., 2004.

Дубров А.М. Компонентный анализ и эффективность в экономике. М., 2002.

Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: Учеб. М., 1998.

Дынкин Е.Б. Марковские процессы. М., 1963.

Дэйвисон М. Многомерное шкалирование. М., 1988.

Дюран Б., Оделл П. Кластерный анализ. М., 1977.

Елисеева И.И. Эконометрика.

Елисеева И.И., С.В.Курышева, Т.В. Костеева и др. М., 2005.

Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа. Пакет ППСА. М., 1986.

Ермаков С.М., Михайлов Г.А. Статистическое моделирование. М., 1982.

Зайдель А.Н. Элементарные оценки ошибок измерений. Л., 1968.

Закс Л. Статистическое оценивание. М., 1976.

Зельнер А. Байесовские методы в эконометрии. М, 1980.

Иберла К. Факторный анализ, М., 1980.

Ибрагимов И.А., Хасьминский Р.З. Асимптотическая теория оценивания, М., 1979.

Ильин В.А., Позняк Э.Г. Линейная алгебра. М., 1999.

Канторович Г.Г. Анализ временных рядов / Экономический журнал ВШЭ. № 2. 2002.

Карлин С. Основы теории случайных процессов. М., 1971.

Кендалл М., Стьюарт А. Статистические выводы и связи. М., 1973.

Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. М., 1976.

Кендэл М. Временные ряды. М., 1981.

Классификация и кластер. Под ред. Дж. Вэн Райзина. М., 1980.

Коваленко И.Н., Филиппова А.А. Теория вероятностей и математическая статистика. М., 1973.

Козлов М.В., Прохоров А.В. Введение в математическую статистику. М., 1987.

Кокрен У. Методы выборочного исследования. М., 1976.

Колемаев В.А., Староверов О.В., Турундаевский В.В. Теория вероятностей и математическая статистика. М., 1991.

Колмогоров А.Н., Журбенко И.Г., Прохоров А.В. Введение в теорию вероятностей. М., 1982.

Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. М., 1974.

Корнейчук Н.П. Экстремальные задачи теории приближения. М., 1976.

Крамер Г. Математические методы статистики. Ижевск, 2003.

Кремер Н.Ш., Путко Б.А. Эконометрика: Учебник для вузов. М., 2005.

Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебник для вузов, М., 2004.

Лабскер Л.Г., Бабешко Л.О. Теория массового обслуживания в экономической сфере. М., 1998.

Левашов В.Ф., Шишов В.Ф., Шмельков В.Б. Теория вероятностей. Пенза, 2003.

Леман Э. Проверка статистических гипотез. М., 1964.

Линник Ю.В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. М., 1962.

Луговская Л.В. Эконометрика в вопросах и ответах. 2006.

Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования временных рядов. М., 2003.

Математическая энциклопедия. М., Т. 1. 1977; Т. 4. 1984.

Математический энциклопедический словарь. М., 1988.

Минько, А.А. Статистический анализ в MS Excel. Профессиональная работа. М., 2005.

Миркин Б.Г. Анализ качественных признаков и структур. М., 1976.

Мостеллер Ф., Рурке Р., Томас Дж. Вероятность. М., 1969.

Мхитарян В.С. Козлов А.Ю., Шишов В.Ф. Статистические функции MS Excel в экономико-статистических расчётах. М., 2003.

Мхитарян В.С., Архипова М.Ю., Балаш В.А., Балаш О.С., Дуброва Т.А., Сиротин В.П. Эконометрика: Учеб. 2008.

Мхитарян В.С., Трошин Л.И. и др. Теория вероятностей и математическая статистика: Учеб. пособие. М., 2007.

Мышкис А.Д. Элементы теории математических моделей. Учебник. М., 2007.

Носко В.П. Эконометрика. Введение в регрессионный анализ временных рядов. М., 2002.

Носко В.П. Эконометрика для начинающих. М., 2005.

Печинкин А.В., Тескин О.И., Цветкова Г.М., Бочаров П.П., Козлов Н.Е. Теория вероятностей. М., 2004.

Поллард Дж. Справочник по вычислительным методам статистики. М., 1982.

Просветов Г.И. Эконометрика: Задачи и решения. М., 2006.

Прохоров Ю.В. Вероятность и математическая статистика: Энциклопедия. М., 1999.

Пфанцагль И. Теория измерений, М., 1976.

Раушенбах Г.В. Меры близости и сходства. Анализ нечисловой информации в социологических исследованиях. М., 1985.

Рунион Р. Справочник по непараметрической статистике. М., 1982.

Сажин Ю.В. Многомерные статистические методы анализа экономических процессов. Саранск, 2008.

Саймон Джинджер. Анализ данных в Excel: наглядный курс создания отчетов, диаграмм и сводных табл. М., 2004.

Самарский А.А., Михайлов А. П. Математическое моделирование. Идеи. Методы. Примеры. М., 2001.

Самыловский А.И. Многомерные модели прикладного статистического анализа, М.,1988.

Сачков В.Н. Комбинаторные методы дискретной математики. М., 1977.

Сигел Э.Ф. Практическая бизнес-статистика. М., 2002.

Слуцкий Е.Е. Сложение случайных величин как источник циклических процессов // Вопросы конъюнктуры. 1927, вып. 1.

Смирнов Н.В., Дунин-Барковский И. В. Курс теории вероятностей и математической статистики для технических приложений. М., 1965..

Соболь И.М. Метод Монте-Карло. М., 1968.

Сошникова Л.А., Тамашевич В.Н., Уебе Г., Шефер М. Многомерный статистический анализ в экономике. М., 1999.

Справочник по прикладной статистике. Под ред. Э.Ллойда, У.Лидермана. М., 1989.

Справочник по теории вероятностей и математической статистике. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. М., 1985.

Статистический словарь / Гл. ред. Королев М.А. М., 1989.

Суслов В.И., Ибрагимов Н.М., Талышева Л.П., Цыплаков А.А. Эконометрия: учебник. Новосибирск, 2005.

Стивенс С. Математика, измерение и психофизика. Экспериментальная психология. Т.1, М., 1960.

Таха Х. Введение в исследование операций, М., 2001.

Теория статистики: Учеб. М., 1998.

Терехина А.Ю. Анализ данных методами многомерного шкалирования. М, 1986.

Тихомиров В.М. Некоторые вопросы теории приближений. М., 1976.

Толстова Ю.Н. Основы многомерного шкалирования. М., 2006.

Турунцева М.Ю. Анализ временных рядов. М., 2003.

Ульрих Л.А. Электронные таблицы Microsoft Excel. Проблемы и решения. М., 2002.

Фаддеев Д.К., Фаддеева В.Н. Вычислительные методы линейной алгебры. М.-Л., 1963.

Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка и др. Под ред. И.С. Енюкова. М., 1989.

Феллер В. Введение в теорию вероятностей и её применения. М., 1964.

Фихтенгольц Г. М. Курс дифференциального и интегрального исчисления. Т. 2. М., 1966.

Халафян А.А. Statistica 6. Статистический анализ данных. Учебник. М., 2008.

Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. Робастность в статистике. Подход на основе функций влияния. М., 1989.

Ханк Д.Э., Уичерн Д.У., Райтс А.Дж. Бизнес-прогнозирование. М., 2003.

Харман Г. Современный факторный анализ. М., 1972.

Хастингс Н., Пикок Дж. Справочник по статистическим распределениям. М., 1980.

Хинчин А.Я. Работы по математической теории массового обслуживания. М., 1963.

Холл М. Комбинаторика. М., 1970.

Хьюбер П. Робастность в статистике. М., 1984.

Чекотовский, Э. Графический анализ статистических данных в Microsoft Excel 2000. М., 2002.

Чернов Г., Мозес Л. Элементарная теория статистических решений, М., 1962.

Четыркин Е.М. Статистические методы прогнозирования. М., 1979.

Шеффе Г. Дисперсионный анализ. М., 1980.

Эддоус М., Стэнсфилд Р. Методы принятия решений, М., 1997.

Янке Е., Эмде Ф. Таблицы функций с формулами и кривыми. Пер. с нем., 3 изд. М., 1959.

Almon S. Distributed Lag between Capital Appropriations and Expenditures. Econometrica, 1965.

Box G.E.P., Pierce D.A. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. Journal of the American Statistical Association. 1970.

Dickey D.A., Fuller W.A. Distribution of the Estimators for Autoregressive Time Series With a Unit Root // Journal of the American Statistical Association. 1979.

Dixon W. J., Massey F. J., Introduction to statistical analysis, N.-Y. – Toronto – L., 1957.

Fisher R.A. On an absolute criterion for fitting frequency curves. Messenger of Mathematics. 1912.

Koyck, L. M. Distributed Lags and lnvestorent Analysis. Amsterdam, 1954.

Luing G.M., Box G.E.P. On a Measure of Lack of Fit in Time Series Models Biometrica. 1978.

Metropolis N., Ulam S. The Monte Carlo Method. Journal of American statistical association. 1949.

Helmholtz H. Numbering and measuring from an episte-mological viewpoint. Epistemological writings. 1930.

Stevens S.S. On the theory of scales of measurement. 1946.

Stevens S.S. Measurement, statistics, and the schemapiric view. 1968.

Thurstone L.L. Multiple factor analysis. 6th.ed. Chicago, 1961.

White H.L. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity / Econometrica, 1980.

Yule G. Udny. On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers. Philosophical Transactions of the Royal Society of London, Ser. A.

Подраздел 2.2. Многомерный статистический анализ и эконометрика

Айвазян А.С., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы моделирования и первичная обработка данных. М., 1983.

Бара Ж.-Р. Основные понятия математической статистики. М., 1974.

Бокс Дж., Дженкинс Г. Анализ временных рядов, прогноз и управление. М., 1974. Вып. 1.

Боровков А.А. Теория вероятностей. М., 1999.

Бююль Ахим, Цёфель Петр. SPSS: Искусство обработки информации.

Анализ статистических данных и восстановление скрытых закономерностей. Спб., 2005.

Вальд А. Последовательный анализ. М., 1960.

Вербик М. Путеводитель по современной эконометрике. – М.,2008.
Гантмахер Ф.Р. Теория матриц. М., 1966.

Гнеденко Б.В. Курс теории вероятностей. М., 1988.

Бернулли Я. О законе больших чисел. М., 1986.

Боровков А.А. Математическая статистика. М.: Наука, 1984.

Дубров А.М. Компонентный анализ и эффективность в экономике. М., 2002.

Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: Учеб. М., 2003.

Дэвид Г. Метод парных сравнений. М.: Статистика, 1978.

Канторович Л.В. Экономический расчет наилучшего использования ресурсов. М., 1959 г.

Козлов М.В. Введение в математическую статистику. М., 1987.

Кокрен У. «Методы выборочного исследования». М.: «Статистика», 1976.

Колмогоров А.Н. Теория вероятностей . Математика, ее содержание, методы и значение. М., 1956.

Кострикин А.И., Манин Ю.И. Линейная алгебра и геометрия. М., 1980.

Крамер Г. Математические методы статистики. М., 1976.

Кунявский М. С., Отношения непосредственного производства при социализме. Минск, 1972.

Курош А.Г. Курс высшей алгебры, М., 1968.

Литтл Р. Дж.А., Рубин Д.Б. «Статистический анализ данных с пропусками» М., 1990.

Лукаш Е.Н. Эконометрика: метод наименьших квадратов в регрессионном анализе. Глава в учеб. «Моделирование экономических процессов». Под ред. М.В.Грачевой и др. М.: Юнити. 2005.

Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. М.: Финансы и статистика, 2003.

Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. - М., 2003.

Лукашин Ю.П. Адаптивная эконометрика. Нелинейные адаптивные регрессионные модел. Вопросы статистики. № 6, 2006.

Лукашин Ю.П. Адаптивный корреляционный анализ экономических показателей. Вестник статистики,1991.

Лукашин Ю.П. Анализ распределения кассовых остатков: адаптивная гистограмма, проблема оптимизации. Экономика и математические методы. 1997.

Лукашин Ю.П. Линейная регрессия с переменными параметрами. М., 1992.

Лукашин Ю.П. Нетрадиционный корреляционный анализ временных рядов. Экономика и математические методы,1992.

Лукашин Ю.П. Фазовый анализ временных рядов. Экономика и математические методы, 1993.

Лурье А. Л., Экономический анализ моделей планирования социалистического хозяйства. М., 1973.

Математический анализ данных и интерпретация результатов его применения. М., 1988.

Миркин Б.Г. Группировка в социально-экономических исследованиях. М., 1985.

Мхитарян В.С., Архипова М.Ю. и др. Эконометрика: Учеб. Под ред. Мхитаряна В.С. М., 2008.

Наследов А.Д. Математические методы психологического исследования: Анализ и интерпретация данных: Учебное пособие. – СПб., 2004.

Новожилов В.В. Проблемы измерения затрат и результатов при оптимальном планировании. – М., Экономика, 1967 г.

Нэреш К. Малхотра. «Маркетинговые исследования. Практическое руководство». М ., 2002.

Нэреш К. Малхотра. Маркетинговые исследования и эффективный анализ статистических данных. М., 2002.

Нэреш К. Малхотра. Маркетинговые исследования с помощью SPSS. Практическое руководство. М., 2007.

Пациорковский В.В., Пациорковская В.В. SPSS для социологов. Учебное пособие. М.: ИСЭПН РАН, 2005.

Пуарье Д. Эконометрия структурных изменений: С применением сплайн-функций) М., 1981.

Рао С.Р. Линейные статистические методы и их применения. М., 1968.

Прохоров Ю.В., Розанов Ю.А. Теория вероятностей. М., 1973.

Соколов Г.А. Теория вероятностей: Учеб. Г.А. Соколов, Н.А. Чистякова. М., 2005.

Статистическое моделирование и прогнозирование. Под ред. Гранберга А.Г. М., 1990.

Строгалев В. П., Толкачева И. О. Имитационное моделирование. М., 2008.

Татарова Г.Г. Типологический анализ в социологии. М., 1993.

Терстоун Л.Л. Психофизический анализ // Проблемы и методы оценки. М., 1974.

Типология и классификация в социологических исследованиях. М., 1982.

Феллер В. Введение в теорию вероятностей и ее приложения: В 2-х тт. М., 1984.

Хемди А. Таха Имитационное моделирование. Введение в исследование операций М., 2007 г.

Ширяев А.Н. Вероятность. В 2-х кн. М., 2004. Кн. 1.

Ширяев А.Н. Статистический последовательный анализ: Оптимальные правила остановки. М.

Шмойлова Р.А., Минашкин В.Г., Садовникова Н.А., Шувалова Е.Б. Под ред. Р.А. Юнг К. Психологические типы. М., 1995.

Field A. Discovering Statistics Using SPSS, Second Edition. 2005.

Bollerrslev T. Generalized Autoregressive Conditional Heteroscedasticity. Journal of Econometrics, 1986.

Bradley RA, Terry M.E. Rank analysis of incomplete block designs I: The method of paired comparisons. 1952.

Brown R.G. Smoothing forecasting and prediction of discrete time series. N.Y., 1963.

Brown R.G. Statistical forecasting for inventory control. – N.Y., 1959.

Brown R.G., Meyer R.F. The fundamental theorum of exponential smoothing// Oper. Res. - 1961.

Carl-Erik Sarndal, Bengh Swensson, Jan Wretman. «Model Assisted Survey Sampling». Sprenger-Verlag New York, Inc. 1992.

D'Esopo D.A. A note on forecasting by the exponential smoothing operator,1961.

George A. Morgan, Nancy L. Leech, Gene W. Gloeckner, Karen Caplovitz Barrett, SPSS for Introductory Statistics: Use and Interpretation.

Gottwald, Siegfried, A Treatise on Many-Valued Logics. Research Studies Press LTD. (2001) Baldock, Hertfordshire, England.

Griffith A. SPSS for Dummies. - Hoboken: Wiley Publishing, 2007.

Holt C.C. Forecasting trends and seasonals by exponentially weighted moving average. O.N.R. Memorandum, Carnegie Inst. of Technology. 1957.

Holt C.C. Forecasting trends and seasonals by exponentially weighted moving averages//O.N.R. Memorandum, Carnegie Inst. of Technology. 1957. №2.

Leech, N. A., Barett K.C., & Morgan, G.A.(2004). SPSS for intermediate statistics: Use and interpretation.

Lukashin Y.P. An adaptive method of regression analysis. // Statistical Analysis and Forecasting of Economic Structural Change, Peter Hackle, Ed. – IIASA, Springer-Verlag, 1989. Ch.13. - P.209-216.

Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2004). SPSS for introductory statistics: Use and Interpretation. Mahwah, NJ: Lawrence Erlbaum Associates. Mosteller F. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations // Psychometrika, 1951, v.16.

Muth J.F. Optimal properties of exponentially weighted forecasts of time series with transitory components//J. Amer. Statist. Ass. -1960.

Muth J.F. Optimal properties of exponentially weighted forecasts of time series with transitory components. J. Amer. Statist. Ass.1960.

Paul R. Kinnear, Colin D. Gray, SPSS For Windows Made Simple.

Programming and Data Management for SPSS 16.0: A Guide for SPSS and SAS Users Raj B., Ullah A. Econometrics: a varying coefficient approach. L., Croom Helm, 1981.

Theil H., Wage S. Some observations on adaptive forecasting. Management Science., 1964.

Thurstone L.L. A Law of comparative judgement // Psychological review. 1927.

Lukashin Y.P. Analysis of data when constructing an adaptive regression model. Model-Oriented Data Analysis, Proceedings. V.Fedorov, H.Lauter, eds Eisenach, GDR/Lecture Notes in Economics and Mathematical Systems., 1987.

Winters P.R. Forecasting sales by exponentially weighted moving averages.

Management Science.-1960.

Zadeh, L. A., Fuzzy Sets as a Basis for a Theory of Possibility, Fuzzy Sets and Systems, 1977.

Zadeh, L. A., Fuzzy sets. Information and Control, Vol. 8, 1965.

Zadeh, L. A., The concept of a linguistic variable and its application to approximate reasoning. Information Sciences, 1975.

Подраздел 2.3.Информационные технологии статистического инструментария

Андрейчиков А.В., Андрейчикова О.Н. Интеллектуальные информационные системы. М., 2004.

Борисов В.В., Круглов В.В., Федулов А.С. Нечёткие модели и сети. М., 2007.

Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб., 2000.

Дебок Г., Кохонен Т. Анализ финансовых данных с помощью самоорганизующихся карт. М, 2001.

Джексон П. Введение в экспертные системы. М., 2001.

Паклин Н., Орешков В. Бизнес-аналитика: от данных к знаниям. СПб., 2009.

Романов В.П. Интеллектуальные информационные системы в экономике. М., 2003.

Подраздел 2.4. Актуарная математика и актуарные расчёты

Бауэрс Н., Гербер Х., Джонс Н., Несбит С., Хикман Дж. Актуарная математика. М., 2001.

Каас Р., Гувертс М., Дэнэ Ж., Денут М. Современная актуарная теория риска. М., 2007.

Касимов Ю.Ф. Введение в актуарную математику (страхование жизни и пенсионных схем). М., 2001.

Лемер Ж. Автомобильное страхование. Актуарные модели. М., 2003.

Савич С.Е. Элементарная теория страхования жизни и трудоспособности. М., 2003.

Томас М. Математика рискового страхования. М., 2005.

Фалин Г.И. Математические основы теории страхования жизни и пенсионных схем. М., 2007.

Bühlmann H. The actuary: the role and limitations of the profession since the mid-19th century. ASTIN Bulletin № 27 (ч. 2), 1997.

Интернет-сайты:
www.actuaries.org.uk
www.soa.org
www.casact.org

Указатель статей (от А до Я)

А

АНАЛИЗ ТИПОЛОГИЧЕСКИЙ

АНДЕРРАЙТЕР

АНДЕРРАЙТИНГ

АПОСТЕРИОРНОЕ РАСПРЕДЕЛЕНИЕ

АПОСТЕРИОРНЫЙ БАЙЕСОВСКИЙ РИСК

АППРОКСИМАЦИЯ ФУНКЦИЙ

АПРИОРНАЯ СТАТИСТИЧЕСКАЯ (ВЫБОРОЧНАЯ) ИНФОРМАЦИЯ

АПРИОРНОЕ РАСПРЕДЕЛЕНИЕ

АСИМПТОТИЧЕСКАЯ ОТНОСИТЕЛЬНАЯ ЭФФЕКТИВНОСТЬ

АСИМПТОТИЧЕСКИЕ СВОЙСТВА ОЦЕНОК

АСИМПТОТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ

Б

БАЗА ДАННЫХ

БАЗА ЗНАНИЙ

БАЙЕСА ТЕОРЕМА (ФОРМУЛА)

БАЙЕСОВСКАЯ КЛАССИФИКАЦИЯ

БАЙЕСОВСКАЯ ОЦЕНКА

БАЙЕСОВСКИЙ ПОДХОД К ОЦЕНИВАНИЮ

БАЙЕСОВСКИЙ РИСК

БАЙЕСОВСКИЙ РИСК АПОСТЕРИОРНЫЙ

БЕЛЫЙ ШУМ

БЕРЕНСА-ФИШЕРА ПРОБЛЕМА

БЕРНУЛЛИ ИСПЫТАНИЯ

БЕРНУЛЛИ ТЕОРЕМА

БИНАРНЫЕ ПЕРЕМЕННЫЕ (БУЛЕВЫЕ, ДИХОТОМИЧЕСКИЕ)

БЛОЧНАЯ МАТРИЦА

БОКСА-ДЖЕНКИНСА ПОДХОД

БРУТТО-ПРЕМИЯ

БРУТТО-СТАВКА

БУЛЕВА МАТРИЦА ПАРНЫХ СРАВНЕНИЙ

БУТСТРЕП МЕТОД

БУТСТРЕП-МОДЕЛИРОВАНИЕ

В

ВАРИАЦИОННЫЙ РЯД

ВАРИМАКС-МЕТОД

ВАР–МОДЕЛЬ (VAR–МОДЕЛЬ)

ВЕРОЯТНОСТНАЯ БУМАГА

ВЕРОЯТНОСТНАЯ МЕРА

ВЕРОЯТНОСТНАЯ МОДЕЛЬ

ВЕРОЯТНОСТНОЕ ПРОСТРАНСТВО

ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКАЯ МОДЕЛЬ

ВЕРОЯТНОСТЬ

ВЕРОЯТНОСТЬ ОШИБОЧНОЙ КЛАССИФИКАЦИИ

ВЕРОЯТНОСТЬ СЛУЧАЙНОГО СОБЫТИЯ А

ВЕРОЯТНОСТЬ УСЛОВНАЯ

ВЗВЕШЕННОЕ ЕВКЛИДОВО РАССТОЯНИЕ

ВЗВЕШИВАНИЕ ВЫБОРОЧНЫХ ДАННЫХ

ВИЗУАЛИЗАЦИЯ ДАННЫХ

ВИТРИНА ДАННЫХ

ВОЗМОЖНЫЕ ЗНАЧЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

ВОССТАНОВЛЕНИЕ ПРОПУЩЕННЫХ НАБЛЮДЕНИЙ

ВРАЩЕНИЯ ФАКТОРОВ ПРОБЛЕМА

ВРЕМЕННОЙ РЯД (РЯД ДИНАМИКИ, ДИНАМИЧЕСКИЙ РЯД)

ВЫБОРКА

ВЫБОРКА КЛАСТЕРНАЯ

ВЫБОРКА КЛАСТЕРНАЯ СЛУЧАЙНАЯ

ВЫБОРКА МНОГОЭТАПНАЯ СЛУЧАЙНАЯ

ВЫБОРКА НЕСЛУЧАЙНАЯ

ВЫБОРКА ПРОСТАЯ СЛУЧАЙНАЯ

ВЫБОРКА РАССЛОЕННАЯ

ВЫБОРКА РАССЛОЕННАЯ СЛУЧАЙНАЯ

ВЫБОРКА СИСТЕМАТИЧЕСКАЯ

ВЫБОРКА СИСТЕМАТИЧЕСКАЯ СЛУЧАЙНАЯ

ВЫБОРКА ЭЛЕМЕНТОВ

ВЫБОРОЧНАЯ ДИСПЕРСИЯ

ВЫБОРОЧНАЯ ЧАСТОТА СОБЫТИЯ

ВЫБОРОЧНЫЕ ХАРАКТЕРИСТИКИ

Г

ГАММА-РАСПРЕДЕЛЕНИЕ

ГАММА-ФУНКЦИЯ ЭЙЛЕРА

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ

ГЕНЕТИЧЕСКИЙ АЛГОРИТМ

ГЕТЕРОСКЕДАСТИЧНОСТЬ

ГИПОТЕЗА СТАТИСТИЧЕСКАЯ

ГИПОТЕЗА СТАТИСТИЧЕСКАЯ АЛЬТЕРНАТИВНАЯ

ГИПОТЕЗА СТАТИСТИЧЕСКАЯ НУЛЕВАЯ (ОСНОВНАЯ)

ГИПОТЕЗА СТАТИСТИЧЕСКАЯ ПРОСТАЯ

ГИПОТЕЗА СТАТИСТИЧЕСКАЯ СЛОЖНАЯ

ГИСТОГРАММА

ГЛАВНЫЕ КОМПОНЕНТЫ

ГЛАВНЫХ КОМПОНЕНТ АНАЛИЗ

ГОМОСКЕДАСТИЧНОСТЬ

ГРЕБНЕВАЯ РЕГРЕССИЯ

ГРЕКО-ЛАТИНСКИЙ КВАДРАТ

Д

ДАННЫЕ

ДВУХШАГОВЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (2МНК)

ДЕНДРОГРАММА

ДЕРЕВО РЕШЕНИЙ

DM-МЕТОДЫ

ДЕСКРИПТИВНАЯ МОДЕЛЬ

ДЕСКРИПТИВНАЯ СТАТИСТИКА

ДЕТЕРМИНАНТ (ОПРЕДЕЛИТЕЛЬ) МАТРИЦЫ

ДЕТЕРМИНИСТСКАЯ МОДЕЛЬ

ДИСКРЕТНОЕ ВЕРОЯТНОСТНОЕ ПРОСТРАНСТВО

ДИСКРИМИНАНТНАЯ ФУНКЦИЯ

ДИСКРИМИНАНТНАЯ ФУНКЦИЯ ЛИНЕЙНАЯ (ФИШЕРА)

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

ДИСПЕРСИОННЫЙ АНАЛИЗ

ДИСПЕРСИЯ

ДИСПЕРСИЯ ВЫБОРОЧНАЯ

ДИСПЕРСИЯ ГЕНЕРАЛЬНАЯ

ДИСПЕРСИЯ ОСТАТОЧНАЯ

ДИСПЕРСИЯ УСЛОВНАЯ

ДИХОТОМИЧЕСИКИЕ (БИНАРНЫЕ) ПЕРЕМЕННЫЕ

ДОВЕРИТЕЛЬНАЯ ВЕРОЯТНОСТЬ

ДОВЕРИТЕЛЬНАЯ ОБЛАСТЬ

ДОВЕРИТЕЛЬНЫЕ ГРАНИЦЫ

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ

ДОСТОВЕРНОЕ СОБЫТИЕ

Е

ЕВКЛИДОВО РАССТОЯНИЕ

З

ЗАВИСИМЫЕ И НЕЗАВИСИМЫЕ СОБЫТИЯ

ЗАКОН БОЛЬШИХ ЧИСЕЛ

ЗАКОН БОЛЬШИХ ЧИСЕЛ УСИЛЕННЫЙ

ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

ЗНАЧИМОСТЬ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

И

ИДЕНТИФИЦИРУЕМОСТЬ

ИДЕНТИФИЦИРУЕМОСТЬ МОДЕЛИ

ИЕРАРХИЧЕСКИЕ ПРОЦЕДУРЫ КЛАСТЕРНОГО АНАЛИЗА

ИНСТРУМЕНТАЛЬНЫХ ПЕРЕМЕННЫХ МЕТОД

ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ (ИИС)

ИНТЕРВАЛ ГРУППИРОВАНИЯ

ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

ИНТЕРПРЕТАЦИЯ ФАКТОРОВ

ИНФОРМАЦИОННАЯ МАТРИЦА ФИШЕРА

ИНФОРМАЦИОННАЯ МОДЕЛЬ

ИНФОРМАЦИОННАЯ СИСТЕМА (ИС)

ИНФОРМАЦИОННАЯ ХАРАКТЕРИСТИКА СВЯЗИ

ИНФОРМАЦИОННОЕ РАССТОЯНИЕ КАЛЛБЭКА

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

ИНФОРМАЦИОННЫЙ КРИТЕРИЙ АКАИКЕ

ИНФОРМАЦИОННЫЙ КРИТЕРИЙ ШВАРЦА

ИНФОРМАЦИОННЫЙ ЭТАП ИССЛЕДОВАНИЯ

ИСПЫТАНИЯ БЕРНУЛЛИ

ИСХОДНЫЕ СТАТИСТИЧЕСКИЕ ДАННЫЕ

К

КВАНТИЛЬ

КВАНТИЛЬНАЯ ТОЧКА

КЛАСС

КЛАССИФИКАЦИОННЫЕ ПЕРЕМЕННЫЕ

КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ

КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ БЕЗ ОБУЧЕНИЯ

КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ С ОБУЧЕНИЕМ

КЛАСТЕР

КЛАСТЕРНЫЙ АНАЛИЗ

КОВАРИАЦИОННАЯ МАТРИЦА

КОВАРИАЦИОННАЯ ФУНКЦИЯ

КОВАРИАЦИОННЫЙ АНАЛИЗ

КОВАРИАЦИЯ

КОИНТЕГРАЦИЯ

КОЛЕБАНИЯ СЕЗОННЫЕ

КОЛЕБАНИЯ ЦИКЛИЧЕСКИЕ

КОЛИЧЕСТВЕННАЯ ШКАЛА

КОЛИЧЕСТВЕННЫЕ ПЕРЕМЕННЫЕ

КОЛИЧЕСТВО ИНФОРМАЦИИ ФИШЕРА

КОМБИНАТОРИКА

КОММУТАЦИОННЫЕ ФУНКЦИИ

КОМПОЗИЦИЯ РАСПРЕДЕЛЕНИЙ

КОМПОНЕНТНЫЙ АНАЛИЗ

КОРРЕЛИРОВАННЫЕ ВЕЛИЧИНЫ

КОРРЕЛЯЦИОННАЯ МАТРИЦА

КОРРЕЛЯЦИОННОЕ ОТНОШЕНИЕ

КОРРЕЛЯЦИОННОЕ ПОЛЕ

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

КОРРЕЛЯЦИЯ

КОРРЕЛЯЦИЯ

КОРРЕЛЯЦИЯ ЛОЖНАЯ

КОЭФФИЦИЕНТ АСИММЕТРИИ

КОЭФФИЦИЕНТ ВАРИАЦИИ

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

КОЭФФИЦИЕНТ КОНКОРДАЦИИ

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ

КОЭФФИЦИЕНТ СВЯЗИ КАЧЕСТВЕННЫХ ПРИЗНАКОВ

КОЭФФИЦИЕНТ СОПРЯЖЁННОСТИ (ПИРСОНА)

КОЭФФИЦИЕНТ ЭКСЦЕССА

КОЭФФИЦИЕНТ ЭЛАСТИЧНОСТИ

КРИТЕРИЙ ω^2

КРИТЕРИЙ АСИММЕТРИИ И ЭКСЦЕССА

КРИТЕРИЙ БАРТЛЕТТА

КРИТЕРИЙ ВИЛКОКСОНА

КРИТЕРИЙ ДАРБИНА-УОТСОНА

КРИТЕРИЙ ЗНАКОВ

КРИТЕРИЙ КОХРАНА

КРИТЕРИЙ ЛОГАРИФМА ОТНОШЕНИЯ ПРАВДОПОДОБИЯ

КРИТЕРИЙ НЕПАРАМЕТРИЧЕСКИЙ

КРИТЕРИЙ НЕСМЕЩЁННЫЙ

КРИТЕРИЙ ОДНОРОДНОСТИ ДИСПЕРСИЙ

КРИТЕРИЙ ОТНОШЕНИЯ ПРАВДОПОДОБИЯ

КРИТЕРИЙ ПИРСОНА X^2

КРИТЕРИЙ РАНГОВЫЙ

КРИТЕРИЙ СОГЛАСИЯ

КРИТЕРИЙ СОГЛАСИЯ КОЛМОГОРОВА

КРИТЕРИЙ СОГЛАСИЯ КОЛМОГОРОВА-СМИРНОВА

КРИТЕРИЙ СОСТОЯТЕЛЬНЫЙ

КРИТЕРИЙ СТАТИСТИЧЕСКИЙ

КРИТЕРИЙ СТАТИСТИЧЕСКИЙ НАИБОЛЕЕ МОЩНЫЙ

КРИТЕРИЙ СТЬЮДЕНТА (Т-КРИТЕРИЙ)

КРИТЕРИЙ ФИШЕРА (F-КРИТЕРИЙ)

КРИТЕРИЙ ПОСЛЕДОВАТЕЛЬНЫЙ

КРИТИЧЕСКАЯ ОБЛАСТЬ

КРИТИЧЕСКАЯ СТАТИСТИКА (СТАТИСТИКА КРИТЕРИЯ)

Л

ЛАГ

ЛАГИРОВАННАЯ ПЕРЕМЕННАЯ (ЛАГОВАЯ ПЕРЕМЕННАЯ, ЗАПАЗДЫВАЮЩАЯ ПЕРЕМЕННАЯ)

ЛАГОВАЯ СТРУКТУРА

ЛАГОВАЯ СТРУКТУРА ВЕРОЯТНОСТНАЯ

ЛАГОВАЯ СТРУКТУРА ГЕОМЕТРИЧЕСКАЯ (КОЙКА)

ЛАГОВАЯ СТРУКТУРА ПОЛИНОМИАЛЬНАЯ (АЛМОН)

ЛАТЕНТНО-СТРУКТУРНЫЙ АНАЛИЗ

ЛАТИНСКИЙ КВАДРАТ

ЛИНЕАРИЗАЦИЯ

ЛОГАРИФМИЧЕСКАЯ ФУНКЦИЯ ПРАВДОПОДОБИЯ

ЛОГАРИФМИЧЕСКИ-НОРМАЛЬНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ

ЛОГИСТИЧЕСКАЯ КРИВАЯ (КРИВАЯ ПЕРЛА-РИДА)

ЛОГИТ-МОДЕЛЬ БИНАРНОГО ВЫБОРА

М

МАРКОВСКАЯ ЦЕПЬ

МАРКОВСКИЙ ПРОЦЕСС (ПРОЦЕСС МАРКОВА)

МАССИВ ИСХОДНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

МАТЕМАТИКО-СТАТИЧЕСКИЕ МЕТОДЫ

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ (СРЕДНЕЕ ЗНАЧЕНИЕ)

МАТРИЦА

МАТРИЦА «ОБЪЕКТ − СВОЙСТВО»

МАТРИЦА ИДЕМПОТЕНТНАЯ

МАТРИЦА КВАДРАТНАЯ

МАТРИЦА КОВАРИАЦИОННАЯ

МАТРИЦА КОРРЕЛЯЦИОННАЯ

МАТРИЦА НАГРУЗОК

МАТРИЦА НЕВЫРОЖДЕННАЯ

МАТРИЦА НЕОТРИЦАТЕЛЬНО ОПРЕДЕЛЁННАЯ

МАТРИЦА ОБРАТНАЯ

МАТРИЦА ОРТОГОНАЛЬНАЯ

МАТРИЦА ПАРНЫХ СРАВНЕНИЙ

МАТРИЦА ПЕРЕХОДНЫХ ВЕРОЯТНОСТЕЙ

МАТРИЦА ПОЛНОГО РАНГА

МАТРИЦА ПОЛОЖИТЕЛЬНО (НЕОТРИЦАТЕЛЬНО) ОПРЕДЕЛЁННАЯ

МАТРИЦА СИММЕТРИЧЕСКАЯ

МАТРИЦА ТРАНСПОНИРОВАННАЯ

МЕДИАНА

МЕРА БЛИЗОСТИ

МЕРА ТОЧНОСТИ

МЕТАДАННЫЕ

МЕТОД «БЛИЖАЙШЕГО СОСЕДА» (ОДИНОЧНОЙ СВЯЗИ)

МЕТОД «ДАЛЬНЕГО СОСЕДА» (ПОЛНЫХ СВЯЗЕЙ)

МЕТОД К-СРЕДНИХ (МЕТОД МАК-КУИНА)

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

МЕТОД ИНСТРУМЕНТАЛЬНЫХ ПЕРЕМЕННЫХ

МЕТОД КОРРЕЛЯЦИОННЫХ ПЛЕЯД

МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

МЕТОД МОМЕНТОВ

МЕТОД МОНТЕ-КАРЛО

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (МНК)

МЕТОД ЦЕНТРОИДНЫЙ

МЕТОДЫ ОЦЕНКИ НЕПАРАМЕТРИЧЕСКИЕ

МЕТОДЫ РОБАСТНЫЕ

МНОГОМЕ́РНОЕ ШКАЛИРОВАНИЕ

МНОГОМЕРНОЕ НАБЛЮДЕНИЕ

МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ МЕТРИЧЕСКОЕ

МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ НЕМЕТРИЧЕСКОЕ

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ ГАММА НОРМАЛЬНЫЙ

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ ПОЛИНОМИАЛЬНЫЙ

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ СТЪЮДЕНТА (ОБОБЩЁННЫЙ)

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ СТЬЮДЕНТА

МНОГОМЕРНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ УИШАРТА

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

МНОГОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

МОДА

МОДЕЛЬ ARCH

МОДЕЛЬ GARCH

МОДЕЛЬ АВТОРЕГРЕСИИ-ПРОИНТЕГРИРОВАННОГО СКОЛЬЗЯЩЕГО СРЕДНЕГО В ОСТАТКАХ (МОДЕЛЬ БОКСА-ДЖЕНКИНСА)

МОДЕЛЬ АВТОРЕГРЕССИИ – СКОЛЬЗЯЩЕГО СРЕДНЕГО (АРСС)

МОДЕЛЬ АВТОРЕГРЕССИИ СО СКОЛЬЗЯЩИМ СРЕДНИМ В ОСТАТКАХ

МОДЕЛЬ АВТОРЕГРЕССИОННЫХ УСЛОВНО ГЕТЕРОСКЕДАСТИЧНЫХ ОСТАТКОВ

МОДЕЛЬ АВТОРЕГРЕССИОННЫХ УСЛОВНО ГЕТЕРОСКЕДАСТИЧНЫХ ОСТАТКОВ ОБОБЩЁННАЯ

МОДЕЛЬ АДАПТИВНАЯ

МОДЕЛЬ БИНАРНОГО ВЫБОРА

МОДЕЛЬ БРАУНА

МОДЕЛЬ БРАУНА ОБОБЩЁННАЯ

МОДЕЛЬ ГРАВИТАЦИОННАЯ

МОДЕЛЬ ИМИТАЦИОННАЯ

МОДЕЛЬ КЛАССИЧЕСКАЯ ЛИНЕЙНАЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ (МКЛМР)

МОДЕЛЬ КЛАССИЧЕСКАЯ ЛИНЕЙНАЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ С ПЕРЕМЕННОЙ СТРУКТУРОЙ

МОДЕЛЬ КОРРЕКЦИИ ОСТАТКОВ (ОШИБОК)

МОДЕЛЬ МАКРОЭКОНОМИЧЕСКАЯ

МОДЕЛЬ МНОЖЕСТВЕННОГО ВЫБОРА

МОДЕЛЬ РАСПРЕДЕЛЁННЫХ ЛАГОВ

МОДЕЛЬ РЕКУРСИВНАЯ

МОДЕЛЬ СКОЛЬЗЯЩЕГО СРЕДНЕГО

МОДЕЛЬ СМЕСИ РАСПРЕДЕЛЕНИЙ

МОДЕЛЬ ТЕОРЕТИКО-ЭКОНОМИЧЕСКАЯ

МОДЕЛЬ ЭКОНОМЕТРИЧЕСКАЯ

МОДЕЛЬ ЭКОНОМИКО-МАТЕМАТИЧЕСКАЯ

МОДЕЛЬ ЭКОНОМИКО-СТАТИСТИЧЕСКАЯ

МОДЕЛЬ ЭКОНОМИКО-СТАТИСТИЧЕСКАЯ

МОМЕНТЫ ВЫБОРОЧНЫЕ

МОЩНОСТЬ КРИТЕРИЯ

МУАВРА – ЛАПЛАСА ТЕОРЕМА

МУЛЬТИАГЕНТНЫЕ (МНОГОАГЕНТНЫЕ) СИСТЕМЫ (МАС)

МУЛЬТИКОЛЛИНЕАРНОСТЬ

Н

НАГРУЗКА

НАГРУЗКА (В ФАКТОРНОМ АНАЛИЗЕ)

НАДЁЖНОСТЬ ПРОГНОЗА (УРОВЕНЬ ДОВЕРИЯ, ДОСТОВЕРНОСТЬ, ДОВЕРИТЕЛЬНАЯ ВЕРОЯТНОСТЬ)

НАКОПЛЕННАЯ ЧАСТОТА

НАЧАЛЬНЫЕ МОМЕНТЫ ВЫБОРОЧНЫЕ

НАЧАЛЬНЫЙ МОМЕНТ СЛУЧАЙНОЙ ВЕЛИЧИНЫ (МОМЕНТ ПОРЯДКА Q ОТНОСИТЕЛЬНО НАЧАЛА ОТСЧТА)

НЕВОЗМОЖНОЕ СОБЫТИЕ

НЕЗАВИСИМОСТЬ НАБЛЮДЕНИЙ

НЕЗАВИСИМОСТЬ СЛУЧАЙНЫХ ВЕЛИЧИН

НЕЗАВИСИМОСТЬ СОБЫТИЙ

НЕЗАВИСИМЫЕ СОБЫТИЯ

НЕИДЕНТИФИЦИРУЕМОСТЬ

НЕЙРОННАЯ СЕТЬ

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ СТАТИСТИКИ

НЕПАРАМЕТРИЧЕСКИЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

НЕПРЕРЫВНОЕ ВЕРОЯТНОСТНОЕ ПРОСТРАНСТВО

НЕРАВЕНСТВО ЧЕБЫШЕВА

НЕСОВМЕСТНОЕ СОБЫТИЕ

НЕТТО-ПРЕМИЯ

НЕЧЁТКАЯ ЛОГИКА

НЕЧЁТКОЕ МНОЖЕСТВО

НОМИНАЛЬНАЯ ШКАЛА

НОРМАЛИЗУЮЩЕЕ ПРЕОБРАЗОВАНИЕ

НУЛЕВАЯ ГИПОТЕЗА

О

ОБОБЩЁННАЯ ДИСПЕРСИЯ

ОБОБЩЁННАЯ ЛИНЕЙНАЯ МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ (ОЛММР)

ОБОБЩЁННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (ОМНК)

ОБРАЩЕНИЕ БЛОЧНОЙ МАТРИЦЫ

ОБРАЩЕНИЕ МАТРИЦЫ

ОБУЧАЮЩАЯ ВЫБОРКА

ОБЪЕДИНЕНИЕ (СУММА) СОБЫТИЙ

ОБЪЯСНЯЮЩАЯ ПЕРЕМЕННАЯ

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

ОПРЕДЕЛИТЕЛЬ (ДЕТЕРМИНАНТ) БЛОЧНОЙ МАТРИЦЫ

ОПРЕДЕЛИТЕЛЬ (ДЕТЕРМИНАНТ) МАТРИЦЫ

ОПРЕДЕЛЯЮЩАЯ ФУНКЦИЯ

ОПТИМАЛЬНАЯ (БАЙЕСОВСКАЯ) ПРОЦЕДУРА КЛАССИФИКАЦИИ

ОПТИМИЗАЦИОННЫЕ ФОРМУЛИРОВКИ СТАТИСТИЧЕСКИХ ЗАДАЧ

ОРДИНАЛЬНАЯ (ПОРЯДКОВАЯ) ШКАЛА

ОРТОГОНАЛЬНЫЕ ПОЛИНОМЫ ЧЕБЫШЕВА

ОТБОР БЕСПОВТОРНЫЙ

ОТБОР ИНФОРМАТИВНЫХ ТИПООБРАЗУЮЩИХ ПРИЗНАКОВ

ОТКЛОНЕНИЕ СРЕДНЕКВАДРАТИЧЕСКОЕ

ОТНОСИТЕЛЬНАЯ ЧАСТОТА (ЧАСТОСТЬ)

ОТНОШЕНИЕ ПРАВДОПОДОБИЯ

ОЦЕНИВАНИЕ РОБАСТНОЕ

ОЦЕНКА ДИСПЕРСИИ

ОЦЕНКА ДОСТАТОЧНАЯ

ОЦЕНКА ИНТЕРВАЛЬНАЯ

ОЦЕНКА НЕСМЕЩЁННАЯ

ОЦЕНКА РОБАСТНАЯ

ОЦЕНКА СОСТОЯТЕЛЬНАЯ

ОЦЕНКА СТАТИСТИЧЕСКАЯ

ОЦЕНКА СТАТИСТИЧЕСКАЯ

ОЦЕНКА ТОЧЕЧНАЯ

ОЦЕНКА ЭФФЕКТИВНАЯ

ОЦИФРОВКА

ОШИБКА АППРОКСИМАЦИИ

ОШИБКА ВТОРОГО РОДА

ОШИБКА ВЫБОРКИ

ОШИБКА ИЗМЕРЕНИЯ

ОШИБКА ПЕРВОГО РОДА

П

ПАКЕТ ПРИКЛАДНЫХ ПРОГРАММ MICROSOFT EXCEL

ПАКЕТ ПРИКЛАДНЫХ ПРОГРАММ STATISTICA

ПАКЕТ ПРИКЛАДНЫХ СТАТИСТИЧЕСКИХ ПРОГРАММ (SPSS)

ПАКЕТ ЭКОНОМЕТРИЧЕСКИЙ STATA

ПАКЕТ ЭКОНОМЕТРИЧЕСКИЙ E-VIEWS

ПАНЕЛЬНЫЕ ДАННЫЕ

ПАРАЛЛЕЛЬНЫЕ КЛАСТЕР-ПРОЦЕДУРЫ

ПАРАМЕТР АДАПТАЦИИ

ПАРАМЕТРИЗАЦИЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

ПАРАМЕТРИЗАЦИЯ РЕГРЕССИОННОЙ МОДЕЛИ

ПАРАМЕТРИЧЕСКИЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

ПАРАМЕТРЫ СДВИГА И МАСШТАБА

ПАРНЫЕ СРАВНЕНИЯ

ПАРНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

ПАССИВНЫЙ ЭКСПЕРИМЕНТ

ПЕРВИЧНАЯ СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ

ПЕРИОД УПРЕЖДЕНИЯ ПРОГНОЗА

ПЕРИОДОГРАММА

ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА

ПЛОТНОСТЬ ВЕРОЯТНОСТИ

ПЛОТНОСТЬ ПОТОКА СОБЫТИЙ

ПЛОТНОСТЬ РАСПРЕДЕЛЕНИЯ

ПОКАЗАТЕЛИ КАЧЕСТВЕННЫЕ

ПОКАЗАТЕЛИ КЛАССИФИКАЦИОННЫЕ (НОМИНАЛЬНЫЕ)

ПОКАЗАТЕЛИ КОЛИЧЕСТВЕННЫЕ

ПОЛИГОН

ПОЛИНОМИАЛЬНАЯ РЕГРЕССИЯ

ПОЛНАЯ СИСТЕМА СОБЫТИЙ (ПОЛНАЯ ГРУППА СОБЫТИЙ)

ПОСЛЕДОВАТЕЛЬНАЯ СХЕМА НАБЛЮДЕНИЙ

ПОСЛЕДОВАТЕЛЬНЫЕ КЛАСТЕР-ПРОЦЕДУРЫ

ПОСЛЕДОВАТЕЛЬНЫЙ АНАЛИЗ

ПОСЛЕДОВАТЕЛЬНЫЙ КРИТЕРИЙ ВАЛЬДА

ПОТОК СОБЫТИЙ

ПОТОК СОБЫТИЙ БЕЗ ПОСЛЕДЕЙСТВИЯ

ПОТОК СОБЫТИЙ ОРДИНАРНЫЙ

ПОТОК СОБЫТИЙ ПРОСТЕЙШИЙ

ПРАВИЛО ТРЁХ СИГМ

ПРЕДИКТОРЫ В МОДЕЛИ РЕГРЕССИИ

ПРЕОБРАЗОВАНИЕ БОКСА-КОКСА

ПРИВЕДЁННАЯ ФОРМА МОДЕЛИ

ПРИНЦИП ОТНОШЕНИЯ ПРАВДОПОДОБИЯ

ПРИНЦИП ПРАКТИЧЕСКОЙ НЕВОЗМОЖНОСТИ МАЛОВЕРОЯТНЫХ СОБЫТИЙ

ПРОБИТ (PROBIT) - МОДЕЛЬ БИНАРНОГО ВЫБОРА

ПРОВЕРКА ВРЕМЕННОГО РЯДА НА СЛУЧАЙНОСТЬ КОЛЕБАНИЙ

ПРОГНОЗ ИНТЕРВАЛЬНЫЙ

ПРОГНОЗ ТОЧЕЧНЫЙ

ПРОГНОЗИРОВАНИЕ

ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ

ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ КОББА-ДУГЛАСА

ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ ЛЕОНТЬЕВСКОГО ТИПА

ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ ЛИНЕЙНАЯ

ПРОИЗВОДСТВЕННАЯ ФУНКЦИЯ С ПОСТОЯННОЙ ЭЛАСТИЧНОСТЬЮ ЗАМЕЩЕНИЯ

ПРОПУЩЕННЫЕ (СТЕРТЫЕ) НАБЛЮДЕНИЯ

ПРОСТОЙ СЛУЧАЙНЫЙ (СОБСТВЕННО-СЛУЧАЙНЫЙ) ОТБОР

ПРОСТРАНСТВЕННАЯ ВЫБОРКА

ПРОСТРАНСТВЕННО-ВРЕМЕННАЯ ВЫБОРКА

ПРОСТРАНСТВО ЭЛЕМЕНТАРНЫХ СОБЫТИЙ

ПРОТИВОПОЛОЖНОЕ (ДОПОЛНИТЕЛЬНОЕ) СОБЫТИЕ

ПРОЦЕДУРЫ КЛАССИФИКАЦИИ

ПРОЦЕДУРЫ КЛАССИФИКАЦИИ ИЕРАРХИЧЕСКИЕ

ПРОЦЕНТНАЯ ТОЧКА РАСПРЕДЕЛЕНИЯ

ПРЯМОЕ (КРОНЕКЕРОВО) ПРОИЗВЕДЕНИЕ МАТРИЦ

Р

РАВНОМЕРНО НАИБОЛЕЕ МОЩНЫЙ КРИТЕРИЙ

РАВНОТОЧНЫЕ ИЗМЕРЕНИЯ

РАЗЛОЖЕНИЕ ЭДЖВОРТА

РАЗМАХ

РАЗМАХ ВЫБОРКИ

РАНГ МАТРИЦЫ А

РАНГОВАЯ КОРРЕЛЯЦИЯ

РАСПОЗНАВАНИЕ ОБРАЗОВ

РАСПРЕДЕЛЕНИЕ БИНОМИАЛЬНОЕ

РАСПРЕДЕЛЕНИЕ «ХИ–КВАДРАТ»

РАСПРЕДЕЛЕНИЕ ВЕЙБУЛЛА

РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ

РАСПРЕДЕЛЕНИЕ ГИПЕРГЕОМЕТРИЧЕСКОЕ

РАСПРЕДЕЛЕНИЕ ГРАММА – ШАРЛЬЕ

РАСПРЕДЕЛЕНИЕ ДВУСТОРОННЕЕ ЭКСПОНЕНЦИАЛЬНОЕ ЛАПЛАСА

РАСПРЕДЕЛЕНИЕ КОШИ

РАСПРЕДЕЛЕНИЕ ЛОГАРИФМИЧЕСКИ-НОРМАЛЬНОЕ

РАСПРЕДЕЛЕНИЕ МАКСВЕЛЛА

РАСПРЕДЕЛЕНИЕ МАРГИНАЛЬНОЕ (ЧАСТНОЕ)

РАСПРЕДЕЛЕНИЕ НОРМАЛЬНОЕ (ГАУССОВСКОЕ)

РАСПРЕДЕЛЕНИЕ НОРМАЛЬНОЕ ДВУМЕРНОЕ

РАСПРЕДЕЛЕНИЕ НОРМАЛЬНОЕ МНОГОМЕРНОЕ

РАСПРЕДЕЛЕНИЕ ОТРИЦАТЕЛЬНОЕ БИНОМИАЛЬНОЕ

РАСПРЕДЕЛЕНИЕ ПАРЕТО

РАСПРЕДЕЛЕНИЕ ПОЛИНОМИАЛЬНОЕ (МУЛЬТИНОМИАЛЬНОЕ)

РАСПРЕДЕЛЕНИЕ ПУАССОНА

РАСПРЕДЕЛЕНИЕ РАВНОМЕРНОЕ (ПРЯМОУГОЛЬНОЕ)

РАСПРЕДЕЛЕНИЕ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

РАСПРЕДЕЛЕНИЕ СОВМЕСТНОЕ (МНОГОМЕРНОЕ)

РАСПРЕДЕЛЕНИЕ СТЬЮДЕНТА (t-РАСПРЕДЕЛЕНИЕ)

РАСПРЕДЕЛЕНИЕ УСЕЧЁННОЕ

РАСПРЕДЕЛЕНИЕ УСЛОВНОЕ

РАСПРЕДЕЛЕНИЕ ФИШЕРА (F – РАСПРЕДЕЛЕНИЕ)

РАСПРЕДЕЛЕНИЕ ХОТЕЛЛИНГА (T^2 - РАСПРЕДЕЛЕНИЕ)

РАСПРЕДЕЛЕНИЕ ЭКСПОНЕНЦИАЛЬНОЕ (ПОКАЗАТЕЛЬНОЕ)

РАСПРЕДЕЛЕНИЕ ЭМПИРИЧЕСКОЕ

РАССТОЯНИЕ МАХАЛАНОБИСА

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ «БЛИЖНЕГО СОСЕДА»

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ «ДАЛЬНЕГО СОСЕДА»

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ «СРЕДНЕЙ СВЯЗИ»

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ «ЦЕНТРОВ ТЯЖЕСТИ»

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ ИНФОРМАЦИОННОЕ

РАССТОЯНИЕ МЕЖДУ КЛАССАМИ ОБЪЕКТОВ ОБОБЩЁННОЕ (ПО КОЛМОГОРОВУ)

РАССТОЯНИЕ МЕЖДУ ОБЪЕКТАМИ

РАССТОЯНИЕ ХЕММИНГА

РАСЩЕПЛЕНИЕ СМЕСИ РАСПРЕДЕЛЕНИЙ

РЕАЛИЗАЦИЯ ВРЕМЕННОГО РЯДА

РЕАЛЬНЫЙ КОМПЛЕКС УСЛОВИЙ

РЕГРЕССИОННЫЙ АНАЛИЗ

РЕГРЕССИЯ ТИПОЛОГИЧЕСКАЯ

РЕДУЦИРОВАННАЯ МАТРИЦА

РЕЗКО ВЫДЕЛЯЮЩИЕСЯ НАБЛЮДЕНИЯ

РЕЗУЛЬТИРУЮЩАЯ ПЕРЕМЕННАЯ

РЕПРЕЗЕНТАТИВНОСТЬ ВЫБОРКИ (ПРЕДСТАВИТЕЛЬНОСТЬ ВЫБОРКИ)

РЕШАЮЩЕЕ ПРАВИЛО

РИСКОВАЯ НАДБАВКА

РИСКОВАЯ ПРЕМИЯ

РЯД ВАРИАЦИОННЫЙ

С

САМООРГАНИЗУЮЩИЕСЯ КАРТЫ КОХОНЕНА

СЕРИЙНАЯ (ГНЕЗДОВАЯ) ВЫБОРКА

СЖАТИЕ МАССИВОВ ИНФОРМАЦИИ

СИММЕТРИЧНАЯ МАТРИЦА

СИСТЕМА ОДНОВРЕМЕННЫХ УРАВНЕНИЙ (СОУ)

СИТУАЦИОННЫЙ АНАЛИЗ

СКОЛЬЗЯЩЕГО СРЕДНЕГО МЕТОД

СКОШЕННОСТЬ ВАРИАЦИОННОГО РЯДА

СЛОЖЕНИЯ ВЕРОЯТНОСТЕЙ ТЕОРЕМЫ

СЛУЦКОГО ТЕОРЕМА

СЛУЧАЙНАЯ (СТОХАСТИЧЕСКАЯ) МАТРИЦА

СЛУЧАЙНАЯ ВЕЛИЧИНА

СЛУЧАЙНАЯ ВЕЛИЧИНА ДИСКРЕТНАЯ

СЛУЧАЙНАЯ ВЕЛИЧИНА КОЛИЧЕСТВЕННАЯ

СЛУЧАЙНАЯ ВЕЛИЧИНА МНОГОМЕРНАЯ (ВЕКТОРНАЯ)

СЛУЧАЙНАЯ ВЕЛИЧИНА НЕПРЕРЫВНАЯ

СЛУЧАЙНАЯ ВЕЛИЧИНА НОМИНАЛЬНАЯ

СЛУЧАЙНАЯ ВЕЛИЧИНА ОДНОМЕРНАЯ (СКАЛЯРНАЯ)

СЛУЧАЙНАЯ ВЕЛИЧИНА ОРДИНАЛЬНАЯ (ПОРЯДКОВАЯ)

СЛУЧАЙНОЕ БЛУЖДАНИЕ (ПРОЦЕСС СЛУЧАЙНОГО БЛУЖДАНИЯ)

СЛУЧАЙНОЕ СОБЫТИЕ

СЛУЧАЙНЫЕ ЧИСЛА (ТАБЛИЦА)

СЛУЧАЙНЫЙ ВЕКТОР

СЛУЧАЙНЫЙ ПРОЦЕСС (СЛУЧАЙНАЯ ФУНКЦИЯ)

СМЕСЬ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ

СМЕСЬ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ ДВУХ НОРМАЛЬНЫХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ

СМЕСЬ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ МНОГОМЕРНЫХ НОРМАЛЬНЫХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ

СНИЖЕНИЕ РАЗМЕРНОСТИ ПРИЗНАКОВОГО ПРОСТРАНСТВА

СОБСТВЕННОЕ (ХАРАКТЕРИСТИЧЕСКОЕ) ЗНАЧЕНИЕ (ЧИСЛО) МАТРИЦЫ

СОБСТВЕННЫЙ (ХАРАКТЕРИСТИЧЕСКИЙ) ВЕКТОР МАТРИЦЫ

СОВОКУПНОСТЬ ГЕНЕРАЛЬНАЯ

СОСТОЯТЕЛЬНОСТЬ ОЦЕНКИ

СПЛАЙН (СПЛАЙН-ФУНКЦИЯ)

СПОСОБЫ ОРГАНИЗАЦИИ ВЫБОРКИ

СРАВНЕНИЕ ДВУХ ГЕНЕРАЛЬНЫХ ДИСПЕРСИЙ

СРАВНЕНИЕ НЕСКОЛЬКИХ ВЕРОЯТНОСТЕЙ

СРАВНЕНИЕ НЕСКОЛЬКИХ ГЕНЕРАЛЬНЫХ СРЕДНИХ

СРЕДНЕЕ ЗНАЧЕНИЕ ВЫБОРОЧНОЕ

СРЕДНЕЕ ЗНАЧЕНИЕ ГАРМОНИЧЕСКОЕ

СРЕДНЕЕ ЗНАЧЕНИЕ ГЕОМЕТРИЧЕСКОЕ

СРЕДНЕКВАДРАТИЧЕСКОЕ ОТКЛОНЕНИЕ

СРЕДНЯЯ МЕРА ВНУТРИКЛАССОВОГО РАССЕЯНИЯ

СТАТИСТИКА БОКСА-ПИРСА

СТАТИСТИКА ЛЬЮНГА-БОКСА

СТАТИСТИЧЕСКАЯ ГИПОТЕЗА

СТАТИСТИЧЕСКАЯ ЗАВИСИМОСТЬ (ВЕРОЯТНОСТНАЯ, СТОХАСТИЧЕСКАЯ)

СТАТИСТИЧЕСКАЯ НЕЗАВИСИМОСТЬ

СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

СТАТИСТИЧЕСКИ ЗНАЧИМАЯ СВЯЗЬ

СТАТИСТИЧЕСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ

СТАТИСТИЧЕСКИЕ МЕТОДЫ ПРОГНОЗИРОВАНИЯ

СТАТИСТИЧЕСКИЙ КОНТРОЛЬ КАЧЕСТВА

СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ СЛУЧАЙНЫХ ПРОЦЕССОВ

СТАЦИОНАРНОСТЬ В УЗКОМ СМЫСЛЕ

СТАЦИОНАРНОСТЬ В ШИРОКОМ СМЫСЛЕ

СТАЦИОНАРНЫЙ СЛУЧАЙНЫЙ ПРОЦЕСС

СТЕПЕНЬ СОГЛАСОВАННОСТИ МНЕНИЙ ЭКСПЕРТОВ

СТЕПЕНЬ ТЕСНОТЫ СТАТИСТИЧЕСКОЙ СВЯЗИ

СТРАХОВАЯ НАГРУЗКА

СТРАХОВАЯ ПРЕМИЯ

СТРАХОВОЙ ВЗНОС (СТРАХОВОЙ ПЛАТЁЖ)

СТРАХОВОЙ ТАРИФ (ТАРИФНАЯ СТАВКА)

СТРАХОВЫЕ РЕЗЕРВЫ

СТРУКТУРНАЯ ФОРМА МОДЕЛИ

СХОДИМОСТЬ ПО ВЕРОЯТНОСТИ

СЮРВЕЙЕР

Т

ТАБЛИЦА (МАТРИЦА) «ОБЪЕКТ-СВОЙСТВО»

ТАБЛИЦА СОПРЯЖЕННОСТИ

ТАБЛИЦЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

ТЕОРЕМА (ФОРМУЛА) БАЙЕСА

ТЕОРЕМА БЕРНУЛЛИ

ТЕОРЕМА МУАВРА-ЛАПЛАСА

ТЕОРЕМА ПУАССОНА

ТЕОРЕМА СЛУЦКОГО

ТЕОРЕМА ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ

ТЕОРЕМЫ СЛОЖЕНИЯ ВЕРОЯТНОСТЕЙ

ТЕОРЕМЫ УМНОЖЕНИЯ ВЕРОЯТНОСТЕЙ

ТЕОРИЯ АСИМПТОТИЧЕСКОГО ОЦЕНИВАНИЯ

ТЕОРИЯ ВЕРОЯТНОСТЕЙ

ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ

ТЕОРИЯ ОШИБОК

ТЕОРИЯ СТАТИСТИЧЕСКИХ РЕШЕНИЙ

ТЕСНОТА СТАТИСТИЧЕСКОЙ СВЯЗИ

ТЕСТ БРЕУША-ПАГАНА НА ГЕТЕРОСКЕДАСТИЧНОСТЬ ОСТАТКОВ

ТЕСТЫ ДИКИ–ФУЛЛЕРА

ТОЛЕРАНТНЫЕ ГРАНИЦЫ

ТОЧЕЧНАЯ ОЦЕНКА

ТОЧНОСТЬ ИНТЕРВАЛЬНОЙ ОЦЕНКИ

ТОЧНОСТЬ ПРОГНОЗА

ТРЕНД

ТРЕНД ЛИНЕЙНЫЙ

ТРЕНД СТЕПЕННОЙ

ТРЕНД ЭКСПОНЕНЦИАЛЬНЫЙ

ТРЁХШАГОВЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (3МНК)

У

УМНОЖЕНИЯ ВЕРОЯТНОСТЕЙ ТЕОРЕМЫ

УНИФИКАЦИЯ ТИПОВ ПЕРЕМЕННЫХ

УРАВНЕНИЕ РЕГРЕССИИ (ФУНКЦИЯ РЕГРЕССИИ)

УРЕЗАНИЕ РАСПРЕДЕЛЕНИЯ

УРОВЕНЬ ЗНАЧИМОСТИ КРИТЕРИЯ

УСЛОВНАЯ ВЕРОЯТНОСТЬ

УСЛОВНОЕ РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ

УСТОЙЧИВОСТЬ СТАТИСТИЧЕСКАЯ

УСТОЙЧИВЫЕ СТАТИСТИЧЕСКИЕ ВЫВОДЫ

Ф

ФАКТОРНАЯ ДИСПЕРСИЯ

ФИКТИВНАЯ ПЕРЕМЕННАЯ

ФОРМУЛА БАЙЕСА

ФУНКЦИЯ МОЩНОСТИ КРИТЕРИЯ

ФУНКЦИЯ ПЛОТНОСТИ ВЕРОЯТНОСТИ

ФУНКЦИЯ ПРАВДОПОДОБИЯ

ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

ФУНКЦИЯ СТАНДАРТНОГО НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

ФУНКЦИЯ ХАРАКТЕРИСТИЧЕСКАЯ

Х

ХАРАКТЕРИСТИКИ ВЫБОРОЧНЫЕ

ХЕММИНГОВО РАССТОЯНИЕ

ХОТЕЛЛИНГА РАСПРЕДЕЛЕНИЕ

ХРАНИЛИЩЕ ДАННЫХ (ХД)

Ц

ЦЕЛЕВАЯ ФУНКЦИЯ

ЦЕЛЕНАПРАВЛЕННОЕ ПРОЕЦИРОВАНИЕ

ЦЕНЗУРИРОВАНИЕ ВЫБОРКИ

ЦЕНТР РАССЕИВАНИЯ

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

ЦЕНТРАЛЬНЫЕ МОМЕНТЫ ВЫБОРОЧНЫЕ

ЦЕНТРАЛЬНЫЙ МОМЕНТ ПОРЯДКА Q

ЦЕНТРОИДНЫЙ МЕТОД

ЦЕПИ МАРКОВА

ЦЕПИ МАРКОВА НЕПРИВОДИМЫЕ

ЦЕПИ МАРКОВА ПЕРИОДИЧЕСКИЕ

ЦЕПИ МАРКОВА ЭРГОДИЧЕСКИЕ

ЦЕПОЧЕЧНЫЙ ЭФФЕКТ

Ч

ЧАСТНАЯ (МАРГИНАЛЬНАЯ) ПЛОТНОСТЬ ВЕРОЯТНОСТИ

ЧАСТНАЯ АВТОКОРРЕЛЯЦИОННАЯ ФУНКЦИЯ

ЧАСТНАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

ЧАСТНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

ЧАСТОСТЬ

ЧАСТОТА ОТНОСИТЕЛЬНАЯ

ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ

Ш

ШКАЛА ИЗМЕРЕНИЙ

ШКАЛИРОВАНИЕ МНОГОМЕРНОЕ

Э

ЭВРИСТИЧЕСКИЕ МЕТОДЫ СНИЖЕНИЯ РАЗМЕРНОСТИ

ЭКЗОГЕННЫЕ ПЕРЕМЕННЫЕ

ЭКОНОМЕТРИКА

ЭКОНОМИКО-МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

ЭКСПЕРТНАЯ СИСТЕМА (ЭС)

ЭКСПЕРТНОЕ УПОРЯДОЧЕНИЕ

ЭКСПЕРТНО-СТАТИСТИЧЕСКИЙ МЕТОД

ЭКСПОНЕНЦИАЛЬНОЕ (ПОКАЗАТЕЛЬНОЕ) РАСПРЕДЕЛЕНИЕ

ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ

ЭКСТРАПОЛЯЦИЯ

ЭКСТРЕМАЛЬНАЯ ГРУППИРОВКА ПРИЗНАКОВ

ЭКСТРЕМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ КЛАССИФИКАЦИИ

ЭКСТРЕМАЛЬНЫЕ (ОПТИМАЛЬНЫЕ) СВОЙСТВА ГЛАВНЫХ КОМПОНЕНТ

ЭЛЕМЕНТАРНОЕ СОБЫТИЕ

ЭМПИРИЧЕСКАЯ (ВЫБОРОЧНАЯ) ДИСПЕРСИЯ

ЭМПИРИЧЕСКАЯ (ВЫБОРОЧНАЯ) ФУНКЦИЯ ПЛОТНОСТИ

ЭМПИРИЧЕСКАЯ (ВЫБОРОЧНАЯ) ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

ЭМПИРИЧЕСКИЕ АНАЛОГИ НАЧАЛЬНЫХ МОМЕНТОВ

ЭМПИРИЧЕСКИЕ АНАЛОГИ ЦЕНТРА ГРУППИРОВАНИЯ

ЭМПИРИЧЕСКИЕ АНАЛОГИ ЦЕНТРАЛЬНЫХ МОМЕНТОВ

ЭНДОГЕННЫЕ ПЕРЕМЕННЫЕ

ЭРГОДИЧЕСКОЕ СВОЙСТВО

ЭТАПЫ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКОГО МОДЕЛИРОВАНИЯ

ЭТАПЫ СТАТИСТИЧЕСКОГО ИССЛЕДОВАНИЯ ЗАВИСИМОСТЕЙ

ЭФФЕКТ СЛУЦКОГО-ЮЛА

ЭФФЕКТИВНОСТЬ КРИТЕРИЯ АСИМПТОТИЧЕСКАЯ

ЭФФЕКТИВНОСТЬ ОЦЕНКИ

ЭФФЕКТИВНЫЙ КРИТЕРИЙ

Ю

ЮЛА-УОКЕРА УРАВНЕНИЯ