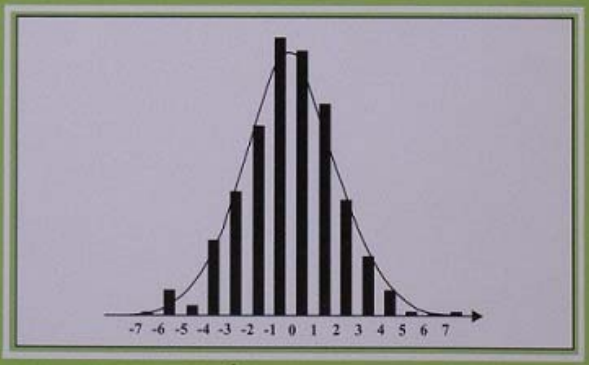
ФИНАНСОВАЯ АКАДЕМИЯ
ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ

А.В. БРАИЛОВ

Лекции по математической статистике



МОСКВА · 2007

ФГОУ ВПО ФИНАНСОВАЯ АКАДЕМИЯ
ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ
(ФИНАКАДЕМИЯ)

КАФЕДРА МАТЕМАТИКИ И ФИНАНСОВЫХ ПРИЛОЖЕНИЙ

А.В. БРАИЛОВ

Лекции по математической статистике

*РЕКОМЕНДОВАНО УЧЕНЫМ СОВЕТОМ ПО СПЕЦИАЛЬНОСТИ
"МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ЭКОНОМИКЕ"
ФИНАКАДЕМИИ В КАЧЕСТВЕ УЧЕБНОГО ПОСОБИЯ
ДЛЯ СТУДЕНТОВ*

МОСКВА · 2007

УДК 512.2(075.8)
ББК 22.172я73
Б87

Рецензенты:
В.Н. Тутубалин, д.физ.-мат.н, проф.
(МГУ им. М.В. Ломоносова, кафедра теории вероятностей)
В.П. Носко, к.физ.-мат.н.
(МГУ им. М.В. Ломоносова, лаборатория теории вероятностей)

Б87 Браилов А.В. Лекции по математической статистике. М.: Финакадемия, 2007. 172 с.

ISBN 978-5-7942-0527-5

В основу пособия положен курс лекций по математической статистике для студентов факультета математических методов в экономике Финансовой академии при Правительстве Российской Федерации. Издание содержит основные разделы математической статистики: теорию выборки, точечные и интервальные оценки параметров распределений, статистическую проверку гипотез, метод наименьших квадратов и линейную модель парной регрессии.

В приложениях читатель познакомится с описанием разработанной автором компьютерной программы «Матричный калькулятор» и примером ее применения для расчета некоторых характеристик финансовых временных рядов.

Издание предназначено для студентов экономических специальностей вузов, изучающих математическую статистику.

УДК 512.2(075.8)
ББК 22.172я73

Учебное издание
Андрей Владимирович Браилов
Лекции по математической статистике

Редактор и корректор *О.В. Платонова*. Художественный редактор *В.А. Селин*
Техническое редактирование и компьютерная верстка *А.В. Браилов, Л.Б. Галкина*

Подписано в печать 25.04.2007. Формат 60x84 1/16. Печать офсетная.
Усл. п.л. 10,0. Усл. кр.-отт. 10,23. Уч.-изд. л. 5,90. Тираж 500 экз. Заказ № 5549.

Финакадемия
125468, Москва, Ленинградский просп., 49

Отпечатано в ФГУП "Производственно-издательский комбинат ВИНИТИ"
140010, Московская обл., г. Люберцы, Октябрьский просп., 403

ISBN 978-5-7942-0527-5

Содержание

Предисловие

Настоящее учебное пособие – результат обработки лекций, которые автор в течение ряда последних лет читал студентам академии, обучающимся по специальности «Математические методы в экономике».

Предполагается, что читатель в дополнение к стандартным курсам математического анализа и линейной алгебры хорошо знаком с теорией вероятностей в объеме, предусмотренном программой упомянутой специальности. Достаточный для понимания набор сведений содержится в изданном в 2002 году пособии «Теория вероятностей. Курс лекций», написанном автором совместно с В.А. Бабайцевым и А.С. Солодовниковым.

Лекции 1–3 посвящены эмпирическим характеристикам и выборкам из конечной генеральной совокупности. Точечные и интервальные оценки параметров распределений рассматриваются в лекциях 4–9. Статистической проверке гипотез посвящены лекции 10–15. В заключительной лекции изучаются метод наименьших квадратов и линейная модель парной регрессии.

По мнению автора, полноценное изучение математической статистики невозможно в отрыве от реальных статистических данных, а с учетом специализации академии – без финансовых данных. Поскольку работа с реальными данными требует применения компьютерной техники, возникает задача согласования теоретического курса статистики с доступными программными средствами. Одним из возможных путей решения этой задачи является разработка специальных компьютерных программ, более простых, нежели известные математические пакеты (MATLAB, Statistica, …), но достаточных для выполнения лабораторных и курсовых работ, а также научных исследований на начальной стадии. Такой программой мог бы стать разработанный автором «Матричный калькулятор», описание которого содержится в Приложении 1. Применению «Матричного калькулятора» для расчета некоторых характеристик финансовых временных рядов посвящено Приложение 2.

Лекция 1
Эмпирические характеристики

Математическая статистика – это представитель целого семейства дисциплин, в названии которых есть слово «статистика». Вот некоторые из них: социальная статистика, финансовая статистика, экономическая статистика, статистика отраслей народного хозяйства и т.д. Основная задача математической статистики состоит в том, чтобы обеспечить конкретные статистические дисциплины надежным теоретическим фундаментом.

§1.1. Основные эмпирические характеристики признака

Первым понятием, с которого мы начнем изложение математической статистики, будет понятие *признака*. В сущности, признак – это то же самое, что функция, только без явной привязки к некоторой области определения. Вместо термина признак иногда используется равнозначный термин *переменная*. Желая подчеркнуть аналогию с теорией вероятностей, мы будем обозначать признаки так же, как и случайные величины – большими латинскими буквами: X, Y и т.д. С каждым признаком может быть связано произвольное количество функций, заданных на различных множествах. Обозначаться все эти функции будут тем же символом, что и соответствующий им признак.

Пример 1.1. Пусть Ω_1 – множество жителей Москвы, Ω_2 – множество жителей Владивостока, X – возраст человека. С признаком X связаны две функции X с областями определения Ω_1 и Ω_2.

Определение. *Область определения конкретной функции, связанной с признаком, называется* (***статистической***) ***совокупностью***, *а число ее элементов –* ***объемом***.

Далее совокупности обозначаются при помощи большой греческой буквы «омега», например, Ω, $\hat{\Omega}$ или Ω_n. В этой лекции предполагается, что объем совокупности конечен.

Рассмотрим признак X, заданный на совокупности $\Omega = \{\omega_1,...,\omega_n\}$. Пусть $x_1 = X(\omega_1)$, …, $x_n = X(\omega_n)$ – его значения.

Определение. ***Эмпирическим средним***, *или* ***средним значением признака в совокупности*** Ω *называется среднее арифметическое всех его значений в этой совокупности*

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Обозначение эмпирического среднего определяется способом записи *значения* признака: сверху добавляется черточка, а номер значения отбрасывается. Если, например, значения признака X в совокупности Ω_1 обозначаются $x_{1,1}, x_{1,2},..., x_{1,n}$, а в совокупности Ω_2 – $x_{2,1}, x_{2,2},..., x_{2,m}$, то средние значения X в Ω_1 и Ω_2 будут обозначены $\overline{x}_1$ и $\overline{x}_2$.

Определение. ***Эмпирической дисперсией***, *или* ***дисперсией признака*** X ***в совокупности*** Ω *называется среднее арифметическое квадратов отклонений его значений от эмпирического среднего*

$$D(X) = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_n - \overline{x})^2}{n},$$

при этом корень $\sigma = \sqrt{D(X)}$ *называется* ***стандартным отклонением*** *признака* X *в совокупности* Ω.

Обозначение эмпирической дисперсии, в отличие от среднего, происходит от обозначения *признака*. Так, для совокупности Ω_m дисперсия X будет обозначена $D_m(X)$ или $\sigma_m^2(X)$.

Эмпирические начальные и *центральные моменты k*-го порядка признака X определяются соотношениями:

$\nu_k(X) = \dfrac{x_1^k + x_2^k + ... + x_n^k}{n}$ – начальный момент и

$\mu_k(X) = \dfrac{(x_1 - \overline{x})^k + (x_2 - \overline{x})^k + ... + (x_n - \overline{x})^k}{n}$ – центральный момент.

В этих формулах $k = 1, 2, ...$ – *порядок* эмпирического момента.

Наконец, *эмпирическая функция распределения* $F(x)$ определяется так:

$$F(x) = \frac{\{\textit{число элементов } \omega \in \Omega, \textit{ для которых } X(\omega) < x\}}{n}.$$

Обозначения эмпирических моментов и функции распределения также зависят от обозначения статистической совокупности. Если, например, признак X определен на совокупности $\hat{\Omega}$, эмпирические моменты обозначаются $\hat{\nu}_k$ и $\hat{\mu}_k$, а функция распределения – $\hat{F}(x)$.

Введенные эмпирические понятия обладают всеми свойствами своих теоретико-вероятностных аналогов. Чтобы понять, почему так происходит, представим признак X как случайную величину. Рассмотрим для этого опыт, состоящий в том, что из Ω случайным равновероятным образом извлекается один элемент $\omega \in \Omega$. Таким образом, $X = X(\omega)$ – случайная величина на вероятностном пространстве $(\Omega, \mathcal{F}, P)$, где $\mathcal{F}$ – алгебра всех подмножеств в Ω, $P(A) = m/n$ – отношение числа элементов в $A \in \mathcal{F}$ к объему Ω. Для элементарных событий ω_i, очевидно, имеем $P(\omega_1) = ... = P(\omega_n) = 1/n$. Отсюда получаем

$$E(X) = X(\omega_1)P(\omega_1) + ... + X(\omega_n)P(\omega_n) = \bar{x},$$

т.е. для данного вероятностного пространства $E(X)$ – это эмпирическое среднее признака X.

Аналогично находим:

$$E(X^k) = \sum_{i=1}^{n} X^k(\omega_i)P(\omega_i) = \frac{x_1^k + ... + x_n^k}{n} = \overline{x^k}$$

– начальный эмпирический момент порядка k,

$$E\{[X - E(X)]^k\} = ... = \frac{(x_1 - \bar{x})^k + ... + (x_n - \bar{x})^k}{n} = \overline{(x - \bar{x})^k}$$

– центральный эмпирический момент порядка k,

$$F_X(x) = P(X < x) = \sum_{x_i < x} P(\omega_i) = \sum_{x_i < x} \frac{1}{n}$$

– эмпирическая функция распределения.

Таким образом, все введенные ранее эмпирические понятия просто совпадают со своими теоретико-вероятностными прототипами, если считать признак случайной величиной, определенной на вероятностном пространстве с равновероятными элементарными исходами. Например, хорошо известное в теории вероятностей тождество

$$D(X) = E(X^2) - E^2(X)$$

применительно к признаку X со значениями $x_1, ..., x_n$ дает следующее соотношение для эмпирической дисперсии:

$$D(X) = \overline{x^2} - \bar{x}^2.$$

Отметим, что искусственное вероятностное пространство $(\Omega, \mathcal{F}, P)$ введено только для того, чтобы обосновать правомерность применения свойств случайных величин к признакам и их эмпирическим характеристикам.

§1.2. Вариационный ряд и эмпирические квантили

Пусть, как и раньше, $x_1, ..., x_n$ – значения (возможно с повторениями) некоторого признака X в совокупности объема n. Упорядочив эти числа по неубыванию, получим *вариационный ряд* признака

$$x_{(1)} \le x_{(2)} \le ... \le x_{(n-1)} \le x_{(n)}.$$

Разность между наибольшим и наименьшим значением $x_{(n)} - x_{(1)}$ называется *размахом* признака.

Порядковый центр (середина) вариационного ряда называется *эмпирической медианой* и определяется формулой

$$Me = \begin{cases} x_{(k+1)}, & \text{если } n = 2k+1, \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right), & \text{если } n = 2k. \end{cases}$$

Пусть, например, в некоторой группе из 20 студентов на экзамене были получены следующие оценки:

4, 2, 5, 5, 3, 4, 3, 5, 5, 3, 2, 3, 4, 4, 3, 4, 4, 2, 4, 5.

Тогда вариационный ряд оценок имеет вид:

$$2 \le 2 \le 2 \le 3 \le 3 \le 3 \le 3 \le 3 \le 4 \le 4 \le 4 \le 4 \le 4 \le 4 \le 4 \le 5 \le 5 \le 5 \le 5 \le 5.$$

Следовательно, медиана $Me = \frac{1}{2}\left(x_{(10)} + x_{(11)}\right) = \frac{1}{2}(4+4) = 4.$

Эмпирические квантили порядка p определяются как приближенные решения уравнения $F(x) = p$, где $F(x)$ – эмпирическая функция распределения. В частности эмпирическая медиана является эмпирическим квантилем (или квантилью) уровня ½. Другим примером эмпирических квантилей служат *квартили*:

$$F(Q_1) \approx \frac{1}{4} \text{ и } F(Q_3) \approx \frac{3}{4}.$$

Квартиль Q_1 можно представлять как точку, разделяющую первую и вторую четверти вариационного ряда. Соответственно, Q_3 можно представлять как точку, разделяющую третью и четвертую четверти вариационного ряда. Разность $Q_3 - Q_1$ (*квартильный размах*) используется в качестве числовой характеристики степени разброса значений признака, альтернативной стандартному отклонению.

§1.3. Распределение частот

Пусть X – признак в совокупности Ω объема n. Список всех его значений образует ряд из n чисел. Удалив из него одинаковые числа и пронумеровав заново то, что осталось, получим последовательность $x_1, ..., x_s$, $s \le n$.

Определение. *Количество n_i элементов $\omega \in \Omega$, для которых $X(\omega) = x_i$ называется* ***частотой*** *значения x_i. Отношение n_i / n называется* ***относительной частотой*** *x_i. При этом таблица частот значений*

x_1	x_2	...	x_s
n_1	n_2	...	n_s

называется ***частотным распределением***, *а таблица относительных частот*

x_1	x_2	...	x_s
$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_s}{n}$

(1.1)

называется ***эмпирическим распределением*** *признака*.

Введенные ранее эмпирические характеристики признака легко находятся по таблицам частот:

- эмпирическое среднее

$$\bar{x} = \frac{x_1 n_1 + ... + x_s n_s}{n}, \tag{1.2}$$

- эмпирический начальный момент порядка k

$$v_k = \frac{x_1^k n_1 + ... + x_s^k n_s}{n}, \tag{1.3}$$

- эмпирический центральный момент порядка k

$$\mu_k = \frac{(x_1 - \bar{x})^k n_1 + ... + (x_s - \bar{x})^k n_s}{n}, \tag{1.4}$$

- эмпирическая дисперсия

$$D(X) = \frac{(x_1 - \bar{x})^2 n_1 + ... + (x_s - \bar{x})^2 n_s}{n}, \tag{1.5}$$

- эмпирическая функция распределения

$$F(x) = \frac{1}{n} \sum_{x_i < x} n_i. \tag{1.6}$$

Определение. *Пусть* $W_1, ..., W_s$ *– положительные, а* $x_1, ..., x_s$ *– произвольные числа. Отношение*

$$\bar{x} = \frac{x_1 W_1 + ... + x_s W_s}{W_1 + ... + W_s}$$

называется ***взвешенным средним*** *чисел* $x_1, ..., x_s$, *при этом число* W_i *называется* ***весом*** x_i.

Из этого определения следует, что правые части формул (1.2) – (1.5) являются взвешенными средними соответствующих чисел. Например, эмпирическое среднее $\bar{x}$ – это взвешенное по частоте среднее значений признака X, эмпирическая дисперсия $D(X)$ – это взвешенное по частоте среднее квадратов отклонений чисел $x_1, ..., x_s$ от их взвешенного среднего $\bar{x}$ и т.д.

Заметим также, что формулы (1.3) – (1.6) можно интерпретировать как соотношения для характеристик дискретной случайной величины, распределенной по закону (1.1).

§1.4. Эмпирический коэффициент корреляции

Пусть $x_i = X(\omega_i)$ и $y_i = Y(\omega_i)$, $\omega_i \in \Omega$. – значения признаков X и Y на совокупности $\Omega = \{\omega_1, ..., \omega_n\}$. *Эмпирическая ковариация* определяется формулой

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \qquad (1.7)$$

Для произвольной функции $z = f(x,y)$ определим признак $Z = f(X,Y)$, значениями которого будут числа

$$z_i = f(x_i, y_i),\ i = 1, ..., n.$$

Может так быть, что среди пар $(x_1, y_1), ..., (x_n, y_n)$ некоторые пары часто повторяются. В этом случае для вычисления среднего $\bar{z}$ нет необходимости n раз вычислять функцию $f(x,y)$. Поступим следующим образом. Так же как и в предыдущем параграфе, в рядах $x_1, ..., x_n$ и $y_1, ..., y_n$ удалим одинаковые значения, а то, что осталось, пронумеруем заново. В результате получим последовательности $x_1, ..., x_r$ и $y_1, ..., y_s$, где $r \le n$ и $s \le n$.

Таблицей сопряженности, или *совместным частотным распределением признаков* X и Y, называется следующая таблица:

	$Y = y_1$	$Y = y_2$	...	$Y = y_s$
$X = x_1$	n_{11}	n_{12}	...	n_{1s}
...	...	...	...	...
$X = x_r$	n_{r1}	n_{r2}	...	n_{rs}

, (1.8)

где n_{ij} – *частота пары* (x_i, y_j), т.е. число элементов $\omega \in \Omega$, для которых $X(\omega) = x_i$, а $Y(\omega) = y_j$.

Нетрудно понять, что сумма всех частот n_{ij} в таблице (1.8) равна объему совокупности n, а статистические распределения признаков X и Y получаются при помощи простых арифметических операций. Так, например, распределение X получается путем суммирования частот в каждой строке таблицы (1.8).

Для вычисления среднего $\bar{z}$ можно воспользоваться основанной на распределении (1.8) формулой

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} f(x_i, y_j) n_{ij}. \tag{1.9}$$

Частным случаем (1.9) является следующее соотношение, применяемое при вычислении эмпирической ковариации:

$$\mathrm{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} (x_i - \bar{x})(y_j - \bar{y}) n_{ij}. \tag{1.10}$$

Эмпирический коэффициент корреляции признаков определяется тем же соотношением, что и коэффициент корреляции случайных величин:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{D(X) D(Y)}}. \tag{1.11}$$

Пример 1.2. В совокупности 16 студентов определены два признака: X – оценка по математике и Y – оценка по иностранному языку. Совместное статистическое распределение оценок задано таблицей:

	$X=2$	$X=3$	$X=4$	$X=5$
$Y=3$	1	0	1	0
$Y=4$	2	4	4	2
$Y=5$	0	1	0	1

.

Требуется найти эмпирический коэффициент корреляции $\rho(X,Y)$.

Решение. Сначала находим статистические распределения признаков:

значение X	2	3	4	5
частота	3	5	5	3

и

значение Y	3	4	5
частота	2	12	2

.

Затем последовательно вычисляем

$$\bar{x}=\frac{1}{16}(2\times 3+3\times 5+4\times 5+5\times 3)=3{,}5;$$

$$D(X)=\frac{1}{16}\left((2-3{,}5)^2\times 3+(3-3{,}5)^2\times 5+(4-3{,}5)^2\times 5+(5-3{,}5)^2\times 3\right)=1$$

$$\bar{y}=\frac{1}{16}(3\times 2+4\times 12+5\times 2)=4;$$

$$D(Y)=\frac{1}{16}\left((3-4)^2\times 2+(4-4)^2\times 12+(5-4)^2\times 2\right)=\frac{1}{4};$$

$$\mathrm{Cov}(X,Y)=\frac{1}{16}\left((1{,}5)+(-0{,}5)+(-0{,}5)+(1{,}5)\right)=\frac{1}{8};$$

$$\rho(X,Y)=\frac{1/8}{\sqrt{1/4}}=0{,}25.$$

Лекция 2
Межгрупповая дисперсия и интервальные характеристики

§2.1. Межгрупповая дисперсия

Пусть X и Y – признаки в совокупности Ω объема n,

y_1	y_2	...	y_s
n_1	n_2	...	n_s

– статистическое распределение Y. Значения Y разбивают совокупность Ω на s групп:

$$\Omega_1 = \{\omega \in \Omega:\ Y(\omega) = y_1\};\ \ldots\ ; \Omega_s = \{\omega \in \Omega:\ Y(\omega) = y_s\}.$$

Элемент *i*-ой группы с номером *j* обозначим ω_{ij}. В результате имеем

$$\Omega_i = \{\omega_{i1}, \omega_{i2}, ..., \omega_{in_i}\}.$$

Приписав каждому элементу ω_{ij} вероятность $1/n$, мы будем считать, что X и Y – случайные величины, заданные на пространстве элементарных событий Ω. Следовательно, распределением случайной величины Y будет следующая таблица:

y_1	y_2	...	y_s
$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_s}{n}$

.

Из теории вероятностей известно равенство

$$E(X) = E(E(X \mid Y)). \qquad (2.1)$$

Посмотрим, что означает это равенство применительно к значениям признака X. Левая часть (2.1) – это эмпирическое среднее признака X,

$$E(X) = \sum_{i,j} x_{ij} P(\omega_{ij}) = \frac{1}{n} \sum_{i,j} x_{ij} = \bar{x},$$

где $x_{ij} = X(\omega_{ij})$ – значение признака на элементе *i*-ой группы с номером *j*. Правую часть (2.1) найдем в два этапа. Сначала получим выражение для условного математического ожидания

$$E(X \mid Y = y_k) = \sum_{i,j} x_{ij} P(\omega_{ij} \mid Y = y_k) = \sum_{j} x_{kj} P(\omega_{kj} \mid Y = y_k) = \\ = \sum_{j} x_{kj} \frac{1/n}{n_k/n} = \frac{1}{n_k} \sum_{j=1}^{n_k} x_{kj} \equiv \bar{x}_k, \tag{2.2}$$

где $\bar{x}_k$ – так называемое *k*-е *групповое среднее*. Другими словами $\bar{x}_k$ – среднее значение признака X в совокупности Ω_k. Теперь мы можем представить правую часть (2.1) в виде $E(E(X \mid Y)) = \sum_{i=1}^{s} \bar{x}_i \frac{n_i}{n}$. В итоге имеем

$$\bar{x} = \frac{\bar{x}_1 n_1 + \ldots + \bar{x}_s n_s}{n_1 + \ldots + n_s}, \tag{2.3}$$

т.е. *среднее значение признака во всей совокупности является взвешенным средним групповых средних*.

Следующая задача – получить аналогичное представление для равенства

$$D(X) = E(D(X \mid Y)) + D(E(X \mid Y)), \tag{2.4}$$

справедливого для любых случайных величин X и Y. Левая часть данного равенства – это эмпирическая дисперсия X во всей совокупности

$$D(X) = \frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2. \tag{2.5}$$

Правая часть (2.4) – сумма двух слагаемых, каждое из которых рассмотрим по отдельности. Чтобы найти $E(D(X \mid Y))$, найдем прежде *k*-е возможное значение случайной величины $D(X \mid Y)$. Имеем

$$D(X \mid Y = y_k) = \sum_{i=1}^{s} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 P(\omega_{ij} \mid Y = y_k) =$$

$$= \sum_{j=1}^{n_k} (x_{kj} - \bar{x}_k)^2 \frac{1}{n_k} = D_k(X),$$

где $D_k(X)$ – эмпирическая дисперсия X в k-ой группе. Мы будем также говорить, что $D_k(X) = \sigma_k^2$ – это *k-я групповая дисперсия*, а дисперсию $D(X) = \sigma^2$ называть *общей дисперсией*. Поскольку $P(Y = y_k) = \frac{n_k}{n}$,

$$E(D(X \mid Y)) = \frac{D_1(X)n_1 + \ldots + D_s(X)n_s}{n_1 + \ldots + n_s}, \qquad (2.6)$$

т.е. $E(D(X \mid Y)) \equiv \bar{\sigma}^2$ – взвешенное среднее групповых дисперсий. Для краткости, $\bar{\sigma}^2$ называется также *средней групповой дисперсией*.

Возможные значения случайной величины $E(X \mid Y)$ уже были найдены выше (см. (2.2)), откуда следует, что

$$D(E(X \mid Y)) = (\bar{x}_1 - \bar{x})^2 \frac{n_1}{n} + \ldots + (\bar{x}_s - \bar{x})^2 \frac{n_s}{n}. \qquad (2.7)$$

Дисперсия $D(E(X \mid Y))$ называется *межгрупповой* и обозначается δ^2. Таким образом, из (2.4) вытекает, что *общая дисперсия равна сумме средней групповой и межгрупповой дисперсии*

$$\sigma^2 = \bar{\sigma}^2 + \delta^2.$$

Пример 2.1. Пусть Ω – совокупность студентов из примера 1.2. Рассмотрим разбиение $\Omega = \Omega_3 \cup \Omega_4 \cup \Omega_5$, где k-я группа Ω_k – это множество студентов, получивших по иностранному языку оценку $Y = k$. Требуется найти среднюю групповую и межгрупповую дисперсии признака X (оценки по математике).

Решение. Заметим, что строки таблицы частот из примера 1.2 – это, фактически, таблицы частот признака X в соответствующей группе. Используя данные частоты, получаем

$$\bar{x}_3 = \frac{1}{2}(2\times 1 + 4\times 1) = 3,\ \ \sigma_3^2 = \frac{1}{2}\left((2-3)^2\times 1 + (4-3)^2\times 1\right) = 1;$$

$$\bar{x}_4 = \frac{1}{12}\left(2\times 2 + 3\times 4 + 4\times 4 + 5\times 2\right) = 3{,}5;$$

$$\sigma_4^2 = \frac{1}{12}\left((2-3{,}5)^2\times 2 + (3-3{,}5)^2\times 4 + (4-3{,}5)^2\times 4 + (5-3{,}5)^2\times 2\right) = \frac{11}{12}$$

$$\bar{x}_5 = \frac{1}{2}(3\times 1 + 5\times 1) = 4,\ \ \sigma_5^2 = \frac{1}{2}\left((3-4)^2\times 1 + (5-4)^2\times 1\right) = 1;$$

Отсюда находим среднюю групповую дисперсию

$$E(D(X\,|\,Y)) = \bar{\sigma}^2 = \frac{1}{16}\left(2\sigma_3^2 + 12\sigma_4^2 + 2\sigma_5^2\right) = \frac{15}{16}.$$

Среднее значение $\bar{x} = 3{,}5$ было уже найдено в примере 1.2, поэтому

$$D(E(X\,|\,Y)) = \delta^2 = \frac{1}{16}\left((\bar{x}_3 - \bar{x})^2\times 2 + (\bar{x}_4 - \bar{x})^2\times 12 + (\bar{x}_5 - \bar{x})^2\times 2\right) = \frac{1}{16}.$$

В разобранном примере роль признака Y свелась к тому, чтобы задать разбиение основной совокупности на группы. Если такое разбиение задано без привлечения признака Y, то при решении задач, связанных с межгрупповой дисперсией, лучше обойтись без его явного использования.

Пример 2.2. Пусть некоторая совокупность разбита на две равные по объему группы. Предположим, что в первой группе среднее значение признака $\bar{x}_1 = 10$, дисперсия $\sigma_1^2 = 15$, а во второй группе $\bar{x}_2 = 20$, $\sigma_2^2 = 25$. Найдите среднее значение и дисперсию признака во всей совокупности.

Решение. Сначала находим среднее, потом дисперсию:

$$\bar{x} = \bar{x}_1\frac{n_1}{n} + \bar{x}_2\frac{n_2}{n} = 10\cdot\frac{1}{2} + 20\cdot\frac{1}{2} = 15,$$

$$\sigma^2 = \bar{\sigma}^2 + \delta^2 = \left(\sigma_1^2\frac{n_1}{n} + \sigma_2^2\frac{n_2}{n}\right) + \left((\bar{x}_1 - \bar{x})^2\frac{n_1}{n} + (\bar{x}_2 - \bar{x})^2\frac{n_2}{n}\right) =$$

$$= 20 + 25 = 45.$$

§2.2. Интервальные характеристики признака

Пусть $(a_1, b_1), \ldots, (a_s, b_s)$ – набор попарно непересекающихся интервалов, покрывающих все значения признака X в совокупности Ω объема n.

Определение. ***Частотой интервала*** (a_i, b_i) *называется число тех элементов* $\omega \in \Omega$, *для которых* $X(\omega) \in (a_i, b_i)$; *интервал* (a_i, b_i) *при этом называется* i*-м* ***интервалом группировки***.

Таблицей интервальных частот, или *интервальным статистическим распределением*, называется следующая таблица:

a_1	b_1	n_1
a_2	b_2	n_2
...	...	...
a_s	b_s	n_s

,

в которой n_i – частота интервала (a_i, b_i), $i = 1, 2, ..., s$. Поскольку интервалы группировки не пересекаются и покрывают все значения признака, сумма интервальных частот равна объему совокупности, $\sum_{i=1}^{s} n_i = n$.

Для графического представления таблицы интервальных частот применяется *гистограмма* – геометрическая фигура, составленная из прямоугольников, основаниями которых служат отрезки $[a_i, b_i]$ на оси абсцисс. Если для представления интервальной частоты используется высота прямоугольника, гистограмма называется *гистограммой частот*. Если же для представления интервальной частоты используется площадь прямоугольника, гистограмма называется *гистограммой плотности частот*.

Гистограммы частот используются обычно в тех случаях, когда все интервалы группировки имеют равную длину. Если же это условие не выполняется, следует использовать гистограмму плотности частот.

В качестве примера приведем гистограмму частот двухдневных процентных изменений индекса Токийской фондовой биржи

Никкей (Nikkei-225) за период с понедельника 02.04.2001 по четверг 31.03.2005. По данным, полученным с сайта **http://www.rbc.ru**, в этом периоде имеется 947 дней, для которых определен индекс на момент закрытия биржи: $x_1, x_2, ..., x_{947}$. Таким образом, имеется 473 двухдневных отрезков, для которых определены относительные изменения индекса:

$$r_1 = \frac{x_3 - x_1}{x_1} = 1{,}17\%;\ r_2 = \frac{x_5 - x_3}{x_3} = 3{,}26\%;\ \ldots;\ r_{473} = \frac{x_{947} - x_{945}}{x_{945}} = 0{,}6\%.$$

Результат подсчета интервальных частот указан в таблице 2.1, в которой a_i, b_i – границы (в процентах) i-го интервала группировки, n_i – частота, $i = 1, 2, ..., 16$.

Табл. 2.1.

a_i(%)	b_i(%)	n_i	a_i(%)	b_i(%)	n_i	a_i(%)	b_i(%)	n_i	a_i(%)	b_i(%)	n_i
-8	-7	0	-4	-3	24	0	1	93	4	5	8
-7	-6	1	-3	-2	40	1	2	77	5	6	1
-6	-5	5	-2	-1	66	2	3	38	6	7	0
-5	-4	3	-1	0	97	3	4	19	7	8	1

Соответствующая гистограмма изображена на рис. 2.1.

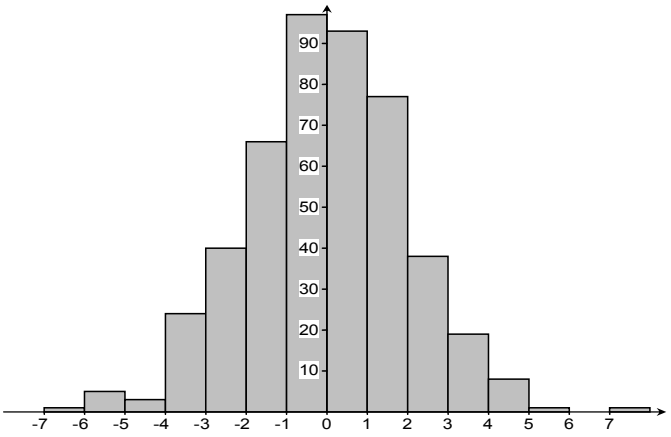


Рис. 2.1.

Построение гистограмм обычно осуществляется при помощи компьютерных программ, которые слегка «зауживают» прямоугольники в гистограмме. Такой прием помимо достижения определенного визуального эффекта позволяет разместить на заднем плане гистограммы дополнительные графические элементы. Так, на рис. 2.2 гистограмма, иллюстрирующая табл. 2.1, составлена из прямоугольников, ширина которых в два раза меньше интервала группировки. Эта гистограмма изображена на фоне графика плотности нормального распределения, параметры которого совпадают с соответствующими характеристиками совокупности процентных изменений индекса, что позволяет на глаз оценить «нормальность» эмпирического распределения. Позже, в главе, посвященной проверке статистических гипотез, вопрос о соответствии эмпирического распределения теоретическому будет рассмотрен достаточно подробно.

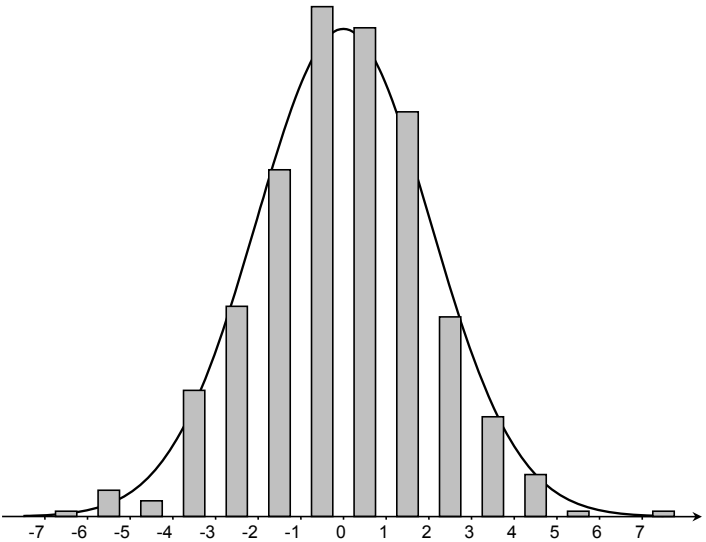


Рис. 2.2.

Обозначим *i*-й интервал группировки через Δ_i, а его середину через $x_i^* = (a_i + b_i)/2$. Определим признак X на множестве $\Omega^* = \{\Delta_1, ..., \Delta_s\}$, положив $X(\Delta_i) \equiv x_i^*$. Рассмотрим опыт, состоя-

щий в том, что из Ω^* случайным образом выбирается один из интервалов группировки и при этом вероятность выбора интервала Δ_i равна его относительной частоте. В этом опыте X – случайная величина, ее характеристики называются (*эмпирическими*) *интервальными характеристиками* признака X.

К основным эмпирическим интервальным характеристикам относятся:

- интервальное среднее $\bar{x}^* = \frac{1}{n}\sum_{i=1}^{s} x_i^* n_i$,
- интервальная дисперсия $D^*(X) = \frac{1}{n}\sum_{i=1}^{s} (x_i^* - \bar{x}^*)^2 n_i$,
- интервальное стандартное отклонение $\sigma^*(X) = \sqrt{D^*(X)}$.

Отметим, что эти и другие эмпирические интервальные характеристики вычисляются как характеристики *интервального распределения*

X	x_1^*	x_2^*	...	x_s^*
P	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_s}{n}$

. (2.8)

Интервальные эмпирические характеристики широко применялись в докомпьютерную эпоху, поскольку для их вычисления требуется сравнительно небольшое число арифметических операций. И сейчас необходимость применения интервальных характеристик возникает всякий раз, когда исследователь располагает только таблицей интервальных частот, а исходное эмпирическое распределение недоступно. В такой ситуации вместо эмпирических характеристик используются их интервальные аналоги. Естественно возникает вопрос: насколько велики ошибки, возникающие при замене $\bar{x}$ на $\bar{x}^*$, $D(X)$ на $D^*(X)$, …?

Оценим величину модуля $|\bar{x} - \bar{x}^*|$. Заметим, что интервалы Δ_i порождают разбиение совокупности Ω на s групп:

$$\Omega_1 = \{\omega \in \Omega : X(\omega) \in \Delta_1\},\ \ldots,\ \Omega_s = \{\omega \in \Omega : X(\omega) \in \Delta_s\}.$$

Пусть h_i – длина Δ_i, а $\bar{x}_i$ – среднее значение признака в i-ой группе, $i = 1, 2, ..., s$. Имеем

$$|\bar{x} - \bar{x}^*| = |\sum_{i=1}^{s} \bar{x}_i \frac{n_i}{n} - \sum_{i=1}^{s} \bar{x}_i^* \frac{n_i}{n}| \le \frac{1}{n} \sum_{i=1}^{s} |\bar{x}_i - \bar{x}_i^*| n_i \le \frac{1}{2} \bar{h},$$

где $\bar{h} = \frac{1}{n} \sum_{i=1}^{s} h_i n_i$ – взвешенная средняя длина интервалов группировки.

Оценка разности $\sigma^2 - \sigma^{*2}$ производится при дополнительных предположениях о наборе интервалов группировки и форме исходного эмпирического распределения признака. Предположим сначала, что все интервалы имеют одинаковую длину h, а распределение на каждом интервале является приближенно равномерным. В этом случае для дисперсии признака имеем

$$\sigma^2 = \bar{\sigma}^2 + \delta^2 \approx \frac{1}{12} h^2 + \delta^2 \approx \sigma^{*2} + \frac{1}{12} h^2. \qquad (2.9)$$

Наличие в правой части формулы (2.10) слагаемого $h^2/12$ связано с тем, что дисперсия равномерного распределения на отрезке длины h равна $h^2/12$. Действительно, по предположению эмпирическое распределение на каждом интервале группировки является приближенно равномерным. Следовательно, средняя групповая дисперсия $\bar{\sigma}^2 \approx \sigma_i^2 \approx h^2/12$. Далее, i-ое групповое среднее $\bar{x}_i \approx \bar{x}_i^*$ в силу равномерности эмпирического распределения на i-ом интервале группировки. Отсюда, одновременно с равенством $\delta^2 \approx \sigma^{*2}$, вытекает и соотношение (2.10).

Рассмотрим теперь случай, когда все интервалы примыкают друг к другу:

$$\Delta_1 = (a_1, a_1 + h), \ \Delta_2 = (a_1 + h, a_1 + 2h), \ ...$$

Если распределение признака X удовлетворяет некоторым условиям «гладкости» эмпирической функции распределения, можно показать, что выполняется приближенное соотношение

$$\sigma^2 \approx \sigma^{*2} - \frac{1}{12} h^2, \qquad (2.10)$$

отличающееся от (2.9) противоположным знаком. Правая часть (2.10) называется *поправкой Шеппарда*. С учетом поправки Шеппарда в качестве приближенного значения неизвестной эмпирической дисперсии σ^2 используется выражение $\sigma^{*2} - \frac{h^2}{12}$. Поправка Шеппарда чаще всего применяется в тех случаях, когда эмпирическая функция распределения хорошо приближается функцией распределения нормального закона.

§2.3. Разложение в ряд по центральным моментам среднего значения дифференцируемой функции

Пусть $x_1, x_2, ..., x_n$ – значения признака в некоторой совокупности объема n, $\bar{x}$ – среднее значение. Предположим, что ряд Тейлора некоторой функции $f(x)$ с центром в точке $\bar{x}$ сходится к этой функции на интервале, содержащем все значения признака

$$f(x) = f(\bar{x}) + \sum_{k=1}^{\infty} \frac{1}{k!} f^{(k)}(\bar{x})(x - \bar{x})^k. \qquad (2.11)$$

Подставив в равенстве (2.11) вместо x поочередно $x_1, x_2, ..., x_n$, взяв сумму полученных равенств и разделив ее на n, получим

$$\overline{f(x)} = f(\bar{x}) + \sum_{k=1}^{\infty} \frac{1}{k!} f^{(k)}(\bar{x})\overline{(x - \bar{x})^k}. \qquad (2.12)$$

Заметив, что $\overline{(x - \bar{x})} = 0$, а $\overline{(x - \bar{x})^k} = \mu_k$ – эмпирический центральный момент порядка k, запишем (2.12) в виде

$$\overline{f(x)} = f(\bar{x}) + \sum_{k=2}^{\infty} \frac{1}{k!} f^{(k)}(\bar{x})\mu_k. \qquad (2.13)$$

Именно эта формула и называется разложением среднего значения функции $\overline{f(x)}$ в ряд по центральным моментам. Если, например, положить $f(x) = e^x$, то получим следующее разложение:

$$\overline{e^x} = e^{\bar{x}} + \sum_{k=2}^{\infty} \frac{1}{k!} e^{\bar{x}} \mu_k = e^{\bar{x}} \left(1 + \sum_{k=2}^{\infty} \frac{\mu_k}{k!} \right). \qquad (2.14)$$

В тех случаях, когда все модули $|x_i|$ во много раз меньше 1, ряд в правой части (2.14) быстро сходится, и имеет место приближенное равенство

$$\overline{e^x} \approx e^{\bar{x}} \left(1 + \frac{\mu_2}{2} \right). \qquad (2.15)$$

Рассмотрим пример. Пусть p_i – цена некоторого актива в конце *i*-го периода, где $i = 1, \ldots, n$. Доходность актива за *i*-й период определяется формулой

$$r_i = \frac{p_i - p_{i-1}}{p_{i-1}},$$

при этом, логарифмическая доходность актива (доходность, связанная с непрерывно начисляемыми процентами) за *i*-й период определяется формулой

$$h_i = \ln\left(\frac{p_i}{p_{i-1}} \right). \qquad (2.16)$$

Обычная доходность r_i связана с логдоходностью простым соотношением

$$r_i = e^{h_i} - 1. \qquad (2.17)$$

Применяя (2.15) и (2.17), получаем приближенное соотношение

$$\bar{r} \approx e^{\bar{h}} \left(1 + \frac{\sigma^2}{2} \right), \qquad (2.18)$$

где $\sigma^2 = \mu_2$ – эмпирическая дисперсия логдоходности.

Пример 2.3. В году было 200 торговых дней по некоторому активу. Цена актива колебалась таким образом, что в течение 100 дней (1-я группа) она росла, а в течение других 100 дней (2-я группа) цена падала. Для дневной логдоходности известно, что в 1-ой группе $\sigma_1 = \bar{h}_1 = 0{,}01$, а во 2-ой группе $\sigma_2 = 0{,}01$ и

$\bar{h}_2 = -0{,}01$. Необходимо приближенно найти сумму дневных доходностей $\sum_{i=1}^{200} r_i$.

Решение. Сначала находим среднюю дневную логдоходность за год

$$\bar{h} = \bar{h}_1 \frac{100}{200} + \bar{h}_2 \frac{100}{200} = 0.$$

Затем – дисперсии:

$$\delta^2 = (\bar{h}_1 - \bar{h})^2 \frac{1}{2} + (\bar{h}_2 - \bar{h})^2 \frac{1}{2} = 10^{-4}, \quad \bar{\sigma}^2 = \sigma_1^2 \frac{1}{2} + \sigma_2^2 \frac{1}{2} = 10^{-4},$$

$$\sigma^2 = \bar{\sigma}^2 + \delta^2 = 2 \cdot 10^{-4}.$$

Применяя (2.18), получим $\bar{r} \approx e^{\bar{h}}(1 + \frac{\sigma^2}{2}) - 1 = 10^{-4}$. Следовательно,

$$\sum r_i = 200 \cdot \bar{r} = 2\%.$$

Лекция 3
Повторные и бесповторные выборки

Начиная с этой лекции все рассматриваемые совокупности разделяются на *генеральные* и *выборочные*.

Определение. *Совокупность, из которой извлекаются элементы, называется* ***генеральной****, тогда как совокупность, образованная отобранными элементами, называется* ***выборочной****.*

§2.4. Повторные и бесповторные выборки

По способу формирования выборочные совокупности разделяются на несколько видов. В первую очередь мы изучим *простые выборки* – выборки, способ формирования которых не предполагает предварительного разбиения генеральной совокупности.

Повторной выборкой называется совокупность, образованная по следующей схеме: сначала из генеральной совокупности случайным равновероятным образом извлекается один элемент; затем этот элемент возвращается в генеральную совокупность и все повторяется, пока не будет отобрано необходимое число элементов. *Бесповторной выборкой* называется совокупность, образованная по аналогичной схеме, но с одним отличием – отобранные элементы в генеральную совокупность не возвращаются.

Характерной особенностью бесповторной выборки является то, что она состоит из различных элементов. Напротив, в состав повторной выборки могут входить одинаковые элементы генеральной совокупности.

Предположим, что из генеральной совокупности Ω объема N извлекается выборка $\hat{\Omega}$ объема n. Пусть X – некоторый признак на Ω. Поскольку все элементы $\hat{\Omega}$, независимо от вида выборки, являются также элементами Ω, признак X определен и на совокупности $\hat{\Omega}$.

Обозначим $x_{01}, x_{02}, ..., x_{0N}$ значения признака X в генеральной совокупности и $X_1, X_2, ..., X_n$ – значения X в выборке. Далее значения $x_{01}, ..., x_{0N}$ рассматриваются как числа, а $X_1, ..., X_n$ – как случайные величины. Именно поэтому первые обозначены строчными буквами, вторые – прописными.

Определение. ***Генеральными*** (*соответственно* ***выборочными***) ***характеристиками*** *признака* X *называют эмпирические характеристики признака* X *в генеральной* (*соответственно выборочной*) *совокупности*.

Например:

$\bar{x}_0 = \frac{1}{N}(x_{01} + ... + x_{0N})$ – генеральное среднее (число);

$\overline{X} = \frac{1}{n}(X_1 + ... + X_n)$ – выборочное среднее (случайная величина);

$D(X) = \frac{1}{N}\sum_{i=1}^{N}(x_{0i} - \bar{x}_0)^2$ – генеральная дисперсия (число);

$\hat{D}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ – выборочная дисперсия (случайная величина).

Теорема 2.1. *Пусть* $X_1, ..., X_n$ *– значения признака* X *в выборке,* $\bar{x}_0$ *– генеральное среднее, а* $D(X)$ *– генеральная дисперсия. Тогда для выборочного среднего* $\overline{X}$ *имеем*:

1) *в случае повторной или бесповторной выборки*

$$E(\overline{X}) = \bar{x}_0;$$

2) *в случае повторной выборки*

$$D(\overline{X}) = \frac{D(X)}{n};$$

3) *в случае бесповторной выборки*

$$D(\overline{X}) = \frac{D(X)}{n}\frac{N-n}{N-1},$$

где N *– объем генеральной совокупности.*

Доказательство. Пусть

x_1	x_2	...	x_s
N_1	N_2	...	N_s

(2.19)

– статистическое распределение признака X в генеральной совокупности. Из классического определения вероятности находим $P(X_1 = x_j) = N_j / N$. Следовательно, вероятностным распределением X_1 будет таблица

x_1	x_2	...	x_s
$\frac{N_1}{N}$	$\frac{N_2}{N}$	...	$\frac{N_s}{N}$

. (2.20)

В случае повторной выборки, очевидно, эта же таблица будет и распределением всех X_i, $i = 1, ..., n$. Точно так же (2.20) служит распределением всех X_i и в случае бесповторной выборки. Чтобы в этом убедиться, достаточно представить бесповторную выборку как результат одновременного извлечения всех ее элементов. Сле-

довательно, и для повторной, и для бесповторной выборки $E(X_i) = \bar{x}_0$ и $D(X_i) = D(X)$, $i = 1, ..., n$. Отсюда вытекает утверждение 1) теоремы:

$$E(\overline{X}) = \frac{1}{n} E(X_1 + ... + X_n) = \bar{x}_0.$$

В случае повторной выборки случайные величины $X_1, ..., X_n$ независимы, и

$$D(\overline{X}) = \frac{1}{n^2} D(X_1 + ... + X_n) = \frac{n}{n^2} D(X_1) = \frac{1}{n} D(X),$$

что доказывает утверждение 2).

Далее считаем, что $X_1, ..., X_n$ – значения признака в бесповторной выборке. Поскольку X_i и X_j зависимы, у нас нет оснований предполагать, как это было в случае повторной выборки, что $\mathrm{Cov}(X_i, X_j) = 0$. Прежде чем найти выражение для $\mathrm{Cov}(X_i, X_j)$, заметим, что данная ковариация одинакова для всех пар (i, j), для которых $i \neq j$. Это следует из того, что бесповторную выборку можно представить как результат одновременного извлечения всех выборочных элементов, и, следовательно, совместное распределение случайных величин X_i и X_j $(i \neq j)$ совпадает с совместным распределением X_1 и X_2. Предположим, что объемы выборочной и генеральной совокупностей равны. Тогда выборочное среднее $\overline{X}$ является неслучайной величиной, поскольку $\overline{X} = \bar{x}_0$. Следовательно,

$$D(\overline{X}) = \mathrm{Cov}(\overline{X}, \overline{X}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{Cov}(X_i, X_j) =$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{Cov}(X_i, X_i) + \frac{1}{N^2} \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) =$$

$$= \frac{1}{N^2} \left(ND(X) + N(N-1)\mathrm{Cov}(X_1, X_2) \right) = 0.$$

Отсюда находим ковариацию X_1 и X_2

$$\mathrm{Cov}(X_1, X_2) = -\frac{D(X)}{N-1}.$$

Завершая доказательство 3), заметим, что для любого n

$$D(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j) =$$

$$= \frac{1}{n}\big(D(X) + (n-1)\mathrm{Cov}(X_1, X_2)\big) =$$

$$= \frac{1}{n}\left(D(X) - (n-1)\frac{D(X)}{N-1}\right) = \frac{D(X)}{n}\frac{N-n}{N-1}.$$

Важнейшим приложением теоремы 2.1 является то, что она позволяет определить точность приближенного равенства

$$\overline{x}_0\,(\text{генеральное среднее}) \approx \overline{X}\ \ (\text{выборочное среднее}). \quad (2.21)$$

Величину характерной ошибки в (2.21) можно определить, по крайней мере, двумя способами:

1) как $E(|\overline{X} - \overline{x}_0|)$, т.е. как *среднюю линейную ошибку*, или

2) как $\sqrt{E(|\overline{X} - \overline{x}_0|^2)}$, т.е. как *среднюю квадратичную ошибку*.

Второе определение значительно более популярно, поскольку для выборки любого вида $\overline{x}_0 = E(\overline{X})$, и, следовательно,

$$\sqrt{E(|\overline{X} - \overline{x}_0|^2)} = \sqrt{E\{(\overline{X} - E(\overline{X}))^2\}} = \sqrt{D(\overline{X})} = \sigma_{\overline{X}}.$$

Определение. ***Средней ошибкой выборки*** *назовем среднеквадратичную ошибку в равенстве* (2.21), *равную* $\sigma_{\overline{X}}$.

Как мы знаем,

$$D(\overline{X}) = \frac{D(X)}{n} \quad \textit{или} \quad D(\overline{X}) = \frac{D(X)}{n}\frac{N-n}{N-1}$$

в зависимости от вида выборки (повторная или бесповторная). В обоих случаях выполняется неравенство $\sigma_{\overline{X}} \le \sigma_X / \sqrt{n}$, поэтому при $n \to \infty$, $\sigma_{\overline{X}} \to 0$.

Пример 2.4. Значение признака X в генеральной совокупности задано следующей таблицей:

значения	3 – 23	23 – 43	43 – 63
частоты	20	60	20

.

Из этой совокупности извлекается бесповторная выборка объема 25. Пусть $\overline{x}_0$ – генеральное, а $\overline{X}$ – выборочное среднее. Найдите среднеквадратичную ошибку в приближенном равенстве $\overline{x}_0 \approx \overline{X}$. При вычислении генеральной дисперсии используйте поправку Шеппарда.

Решение. Составим интервальное распределение (2.8)

X	13	33	53
P	0,2	0,6	0,2

.

Затем найдем интервальное среднее $\overline{x}^*$ и интервальную дисперсию σ^{*2}:

$$\overline{x}^* = 13 \cdot 0{,}2 + 33 \cdot 0{,}6 + 53 \cdot 0{,}2 = 33;$$

$$\sigma^{*2} = (13-33)^2 \cdot 0{,}2 + (33-33)^2 \cdot 0{,}6 + (53-33)^2 \cdot 0{,}2 = 400 \cdot 0{,}4 = 160.$$

С учетом поправки Шеппарда

$$\sigma^2 \approx \sigma^{*2} - h^2/12 = 160 - 400/12 \approx 126{,}7.$$

Следовательно, средняя ошибка выборки

$$\sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}} \approx \sqrt{\frac{126{,}7}{25}\frac{75}{99}} = \sqrt{\frac{126{,}7}{33}} \approx 1{,}96.$$

§2.5. Выборочная доля признака

Одновременно с распределением (2.19) признака X в генеральной совокупности, рассмотрим распределение X в выборочной совокупности

x_1	x_2	...	x_s
n_1	n_2	...	n_s

.

Определение. *Отношение* $p_i = N_i / N$ (*соответственно* $\hat{p}_i = n_i / n$) *называется* ***генеральной*** (*соответственно* ***выборочной***) ***долей значения*** x_i *признака* X.

Теорема 2.2. *Пусть* p *– генеральная, а* $\hat{p}$ *– выборочная доля какого-либо значения* x_1 *признака* X, $q = 1 - p$. *Тогда*:

1) *в случае повторной или бесповторной выборки*

$$E(\hat{p}) = p;$$

2) *в случае повторной выборки*

$$D(\hat{p}) = \frac{pq}{n};$$

3) *в случае бесповторной выборки*

$$D(\hat{p}) = \frac{pq}{n}\frac{N-n}{N-1}.$$

Доказательство. Определим вспомогательный признак Y:

$$Y = \begin{cases} 1, & \text{если } X = x_1, \\ 0, & \text{если } X \neq x_1. \end{cases}$$

Из определения генеральной доли вытекает, что эмпирическое распределение признака Y в генеральной совокупности имеет вид

значение	0	1
отн. частота	q	p

.

Следовательно, генеральное среднее $\overline{y}_0 = p$. Аналогично для выборочного среднего имеем $\overline{Y} = \hat{p}$, поэтому из теоремы 2.1 находим

$$E(\hat{p}) = E(\overline{Y}) = \overline{y}_0 = p.$$

что и доказывает утверждение 1).

Прежде чем доказывать утверждения 2) и 3), найдем генеральную дисперсию признака Y:

$$D(Y) = E(Y^2) - E^2(Y) = (0^2 \cdot q + 1^2 \cdot p) - p^2 = pq.$$

Теперь утверждения 2) и 3) теоремы 2.2 получаются как следствия соответствующих утверждений теоремы 2.1:

$$D(\hat{p}) = D(\overline{Y}) = \frac{D(Y)}{n} = \frac{pq}{n}$$

– в случае повторной выборки и

$$D(\hat{p}) = D(\overline{Y}) = \frac{D(Y)}{n}\frac{N-n}{N-1} = \frac{pq}{n}\frac{N-n}{N-1}$$

– в случае бесповторной выборки.

Следствием теоремы 2.2 является приближенное равенство

$$p \text{ (генеральная доля)} \approx \hat{p} \text{ (выборочная доля)}. \qquad (2.22)$$

Средняя ошибка в (2.22), так же как и для (2.21), определяется как средняя квадратичная ошибка:

$$\sqrt{E\{(\hat{p}-p)^2\}} = \sqrt{E\{(\hat{p}-E(\hat{p}))^2\}} = \sqrt{D(\hat{p})} = \sigma_{\hat{p}}.$$

Как правило, объем генеральной совокупности достаточно велик, так, что $N-1 \approx N$. Поэтому

$$D(\hat{p}) \approx \frac{pq}{n}\left(1-\frac{n}{N}\right).$$

Отсюда вытекает формула для средней ошибки доли признака:

$$\sigma_{\hat{p}} \approx \sqrt{\frac{pq}{n}\left(1-\frac{n}{N}\right)}. \qquad (2.23)$$

Пример 2.5. В выборах приняли участие 1000000 избирателей. Предполагая, что за наиболее популярного кандидата проголосует $\approx 50\%$ избирателей, найдите стандартное отклонение процента бюллетеней в его пользу среди первых 900000 обработанных бюллетеней.

Решение. Поскольку генеральная доля кандидата $p \approx 0{,}5$, имеем

$$\sigma_{\hat{p}} \approx \sqrt{\frac{0{,}5 \cdot 0{,}5}{900000}\left(1-\frac{900000}{1000000}\right)} = \frac{5}{3} \cdot 10^{-4} \approx 0{,}017\%.$$

Вследствие правила «трех сигм», можно с большой степенью надежности утверждать, что интервал $(\hat{p}-3\sigma_{\hat{p}}, \hat{p}+3\sigma_{\hat{p}})$ накроет p. Таким образом, есть достаточно веские основания предполагать, что после проверки 900000 бюллетеней выборочная доля кандидата будет совпадать с окончательным результатом вплоть до десятых долей процента.

§2.6. Пропорциональная выборка

Пусть генеральная совокупность Ω объема N разбита на s групп $\Omega_1, ..., \Omega_s$, объемы которых равны $N_1, ..., N_s$. Предположим, что из этих групп независимым образом извлекаются выборки $\hat{\Omega}_1, ..., \hat{\Omega}_s$ объемов $n_1, ..., n_s$. Объединенная выборка $\hat{\Omega} = \hat{\Omega}_1 \cup ... \cup \hat{\Omega}_s$ объема $n = n_1 + ... + n_s$ называется *сложной*, поскольку способ ее формирования зависит от предварительного разбиения генеральной совокупности. Если все выборки $\Omega_i (i = 1, ..., s)$ были повторными (соответственно бесповторными), совокупность $\hat{\Omega}$ называется *сложной повторной* (соответственно *сложной бесповторной*) *выборкой*. Сложная выборка называется *пропорциональной*, если выполнены условия:

$$\frac{n_1}{n} = \frac{N_1}{N}, \ \ldots, \ \frac{n_s}{n} = \frac{N_s}{N}.$$

Эти равенства можно записать по-другому:

$$n_1 = \alpha N_1, \ \ldots, n_s = \alpha N_s, \text{ где } \alpha = \frac{n}{N}.$$

Найдем математическое ожидание и дисперсию выборочного среднего сначала для случая повторной пропорциональной выборки:

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{s} E(n_i \overline{X}_i) = \frac{1}{n}\sum_{i=1}^{s} n_i \overline{x}_{0i} =$$

$$= \frac{1}{n}\sum_{i=1}^{s} \alpha N_i \overline{x}_{0i} = \sum_{i=1}^{s} \frac{N_i}{N} \overline{x}_{0i} = \overline{x}_0,$$

где $\overline{X}_i$ (соответственно $\overline{x}_{0i}$) – среднее значение признака в $\hat{\Omega}_i$ (соответственно в Ω_i), и

$$D(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{s} n_i^2 D(\overline{X}_i) = \frac{1}{n^2}\sum_{i=1}^{s} n_i D_i(X) =$$

$$= \frac{1}{n}\sum_{i=1}^{s} D_i(X)\frac{n_i}{n} = \frac{1}{n}\sum_{i=1}^{s} D_i(X)\frac{N_i}{N} = \frac{\overline{\sigma}^2}{n}, \qquad (2.24)$$

где $D_i(X)$ – дисперсия признака X в Ω_i, а $\overline{\sigma}^2$ – средняя групповая дисперсия X в генеральной совокупности.

Теперь находим $E(\overline{X}), D(\overline{X})$ для бесповторной пропорциональной выборки.

$$E(\overline{X}) = \overline{x}_0,$$

$$D(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{s} n_i^2 D(\overline{X}_i) = \frac{1}{n^2}\sum_{i=1}^{s} n_i D_i(X)\frac{N_i - n_i}{N_i - 1} \approx$$

$$\approx \frac{1}{n}\sum_{i=1}^{s} D_i(X)\frac{n_i}{n}(1-\alpha) = \frac{1}{n}\sum_{i=1}^{s} D_i(X)\frac{N_i}{N}(1-\alpha) = \frac{(1-\alpha)\overline{\sigma}^2}{n}. \qquad (2.25)$$

Напомним, что согласно теореме 2.1

$$D(\overline{X}) = \frac{\sigma^2}{n} \qquad (2.26)$$

для простой повторной выборки и

$$D(\overline{X}) = \frac{D(X)}{n}\frac{N-n}{N-1} \approx \frac{(1-\alpha)\sigma^2}{n} \qquad (2.27)$$

для простой бесповторной выборки, где $\sigma^2 \equiv D(X)$ – генеральная дисперсия.

Сравнивая формулы (2.24), (2.25) и (2.26), (2.27), приходим к важному выводу: *дисперсия выборочного среднего пропорциональной выборки составляет* $100\overline{\sigma}^2/\sigma^2$ *процентов от дисперсии выборочного среднего простой выборки*.

Пример 2.6. В некотором городе на выборах мэра кандидата А в государственном секторе поддерживает 90 % избирателей, а в частном – только 10 %; при этом количество избирателей в обоих секторах составляет 100 тыс. человек. Некоторая служба предполагает провести опрос 2000 избирателей с целью выявить степень поддержки этого кандидата. На сколько процентов средняя ошибка выборки в случае пропорционального отбора будет меньше, чем в случае простого отбора?

Решение. Находим: генеральную долю

$$p = 0{,}5 \cdot 0{,}9 + 0{,}5 \cdot 0{,}1 = 0{,}5,$$

генеральную дисперсию $\sigma^2 = pq = 0{,}25$ и среднюю групповую дисперсию

$$\overline{\sigma}^2 = 0{,}5 \times (0{,}9 \cdot 0{,}1) + 0{,}5 \times (0{,}1 \cdot 0{,}9) = 0{,}09.$$

Поскольку $\overline{\sigma} / \sigma = 0{,}6,$ средняя ошибка доли при пропорциональном отборе избирателей будет на 40 % меньше, чем при простом отборе.

Задача. Пусть X и Y – признаки в совокупности Ω, $\eta = \eta_{X|Y}$ – их корреляционное отношение. Из Ω извлекаются две выборки: сначала простая, затем – пропорциональная, с учетом разбиения на группы

$$\Omega_1 = \{\omega \in \Omega :\ Y(\omega) = y_1\};\ \ldots\ ; \Omega_s = \{\omega \in \Omega :\ Y(\omega) = y_s\}.$$

Докажите, что замена простой выборки на пропорциональную уменьшает дисперсию выборочного среднего $\overline{X}$ на $100\eta^2$ %.

Лекция 4
Выборки из распределения

В этой лекции мы приступаем к изучению выборок из распределения – выборок нового класса, существенно расширяющего класс повторных выборок. Что же касается бесповторных выборок,

то, по крайней мере, в случае генеральной совокупности большого объема, бесповторные выборки приближенно являются повторными, и выводы, справедливые в отношении повторных выборок, на практике применяются (часто без должного обоснования) также и для бесповторных выборок.

§4.1. Выборки из распределения

Пусть Ω – пространство элементарных событий, связанное с испытанием, в ходе которого случайная величина X получает определенное значение. Таким образом, $X = X(\omega)$ – функция от $\omega \in \Omega$. Предположим, что данное испытание повторяется в одинаковых условиях n раз и $\hat{\omega}_1, ..., \hat{\omega}_n$ ($\hat{\omega}_i \in \Omega$) – результаты этих испытаний. Рассмотрим множество $\hat{\Omega} = \{\hat{\omega}_1, ..., \hat{\omega}_n\}$ – аналог выборочной совокупности. Функция X определена как на Ω, так и на $\hat{\Omega}$ и, согласно принятой ранее терминологии, называется *признаком*.

Поскольку $\hat{\Omega}$ играет роль выборочной совокупности, все эмпирические характеристики X в $\hat{\Omega}$ далее называются *выборочными* характеристиками случайной величины X и обозначаются подобно соответствующим вероятностным характеристикам:

$\hat{D}(X)$ – выборочная дисперсия;

$\hat{\mu}_k$, $\hat{\nu}_k$ – выборочные моменты и т.д.

Исключение составляет выборочное среднее

$$\overline{X} = \hat{E}(X) = \frac{1}{n}\big(X(\hat{\omega}_1) + ... + X(\hat{\omega}_n)\big).$$

В рассматриваемой нами ситуации пространство элементарных событий Ω является аналогом генеральной совокупности, поэтому вероятностные характеристики случайной величины X далее интерпретируются как *генеральные* характеристики признака X. Например, $E(X)$ будет называться *генеральным средним*, дисперсия $D(X)$ – *генеральной дисперсией* и т.д.

Рассмотрим значения случайной величины X, принятые в отдельных независимых испытаниях:

$$X_1 = X(\hat{\omega}_1), ..., X_n = X(\hat{\omega}_n). \qquad (4.1)$$

Далее считаем, что $X_1, ..., X_n$ – это случайные величины, связанные со сложным опытом, состоящим из n простых испытаний. Ясно, что $X_1, ..., X_n$ независимы и распределены по тому же закону, что и случайная величина X.

Напомним, что в дискретном случае закон распределения задается, как правило, таблицей возможных значений и их вероятностей, в абсолютно непрерывном – функцией плотности, в общем случае – функцией распределения. Пусть $\mathcal{L}$ – какой-либо закон распределения вероятностей.

Определение. ***Выборкой объема n из распределения*** $\mathcal{L}$ *называется набор n независимых случайных величин, распределенных по закону* $\mathcal{L}$.

Набор (4.1) случайных величин $X_1, ..., X_n$ является основным примером выборки из распределения. Роль закона $\mathcal{L}$ в данном случае играет распределение случайной величины X.

Изученные в предыдущей главе повторные выборки являются выборками из распределения в следующем смысле. Предположим, что $\Omega = \{\omega_1, ..., \omega_N\}$ – генеральная совокупность, X – определенный в Ω признак, $\hat{\Omega} = \{\hat{\omega}_1, ..., \hat{\omega}_n\}$ – выборочная совокупность. Как уже отмечалось раньше (см. доказательство теоремы 3.1) для любого $i = 1,...,n$ распределение случайной величины $X_i = X(\hat{\omega}_i)$ совпадает с эмпирическим распределением X в Ω. Следовательно, значения X в $\hat{\Omega}$ образуют выборку из эмпирического распределения X в генеральной совокупности.

Пусть $X_1, ..., X_n$ – выборка из какого-либо распределения $\mathcal{L}$, ω – элементарное событие опыта, с которым связаны случайные величины $X_1, ..., X_n$. Набор чисел $x_1 = X_1(\omega)$, …, $x_n = X_n(\omega)$ называется *конкретной* выборкой, соответствующей *случайной* выборке $X_1, ..., X_n$. Упорядочив конкретную выборку по неубыванию, получим:

$$x_{(1)} \le x_{(2)} \le \ ... \ \le x_{(n-1)} \le x_{(n)}. \qquad (4.2)$$

Заметим, что $x_{(k)}$ – это приближенный квантиль выборочного распределения X порядка k/n. Набор чисел (4.2) зависит от выбора ω, поэтому каждое $x_{(k)}$ следует рассматривать как значение соответствующей случайной величины $X_{(k)}$, называемой *k*-ой *порядковой статистикой*. Нетрудно видеть, что $X_{(1)} = \min\{X_1, ..., X_n\}$ и $X_{(n)} = \max\{X_1, ..., X_n\}$, при этом порядковые статистики удовлетворяют аналогичным (4.2) неравенствам:

$$X_{(1)} \le X_{(2)} \le \ldots \le X_{(n-1)} \le X_{(n)}. \tag{4.3}$$

Пусть $F(x)$ – известная функция генерального распределения. Найдем $F_k(x)$ – функцию распределения *k*-ой порядковой статистики $X_{(k)}$ для выборки объема *n*. Обозначим через N_x случайную частоту интервала $(-\infty, x)$, т.е. число тех X_i $(i = 1,...,n)$, которые при исходе ω приняли значение меньше *x*. Из определения N_x следует, что

$$P(N_x = j) = C_n^j F^j(x)[1 - F(x)]^{n-j},$$
$$P(X_{(k)} < x) = P(N_x \ge k).$$

Отсюда получаем

$$F_k(x) = \sum_{j=k}^{n} P(N_x = j) = \sum_{j=k}^{n} C_n^j F^j(x)[1 - F(x)]^{n-j}. \tag{4.4}$$

Интересно отметить, что выражение для соответствующей плотности не содержит знака суммы. Действительно, если $f(x) = F'(x)$ – плотность генерального распределения, то плотность $f_k(x)$ *k*-ой порядковой статистики имеет вид

$$f_k(x) = \frac{d}{dx}\left(\sum_{j=k}^{n} C_n^j F^j(x)[1 - F(x)]^{n-j} \right)$$
$$= kC_n^k F^{k-1}(x)[1 - F(x)]^{n-k} f(x).$$

Заменяя kC_n^k на nC_{n-1}^{k-1}, получим

$$f_k(x) = C_{n-1}^{k-1} F^{k-1}(x)[1 - F(x)]^{n-k} \cdot nf(x). \tag{4.5}$$

Правая часть формулы (4.5) допускает простую интерпретацию:

$C_{n-1}^{k-1}F^{k-1}(x)[1-F(x)]^{n-k}$ – условная плотность $X_{(k)}$ в точке x;

$nf(x)$ – плотность условия, состоящего в том, что одна из случайных величин $X_1,...,X_n$ принимает значение x.

§4.2. Точечные статистические оценки

Пусть $X_1,...,X_n$ – выборка из распределения $\mathcal{L}$. Случайная величина вида $h(X_1,...,X_n)$, где $h(x_1,...,x_n)$ – какая-либо функция от выборочных значений, называется *статистикой*. Раньше мы уже встречались с порядковыми статистиками, теперь мы приступаем к изучению *точечных статистических оценок* – статистик, предназначенных для оценки параметров распределения.

Предположим, что генеральное распределение $\mathcal{L}$ зависит от параметра $\theta \in \Theta$, где Θ – множество допустимых значений параметра. Для того чтобы статистика $\hat{\theta} = h(X_1,...,X_n)$ могла быть (точечной) оценкой параметра θ, было бы желательно иметь $P(\hat{\theta} \approx \theta) \approx 1$ для всех $\theta \in \Theta$.

Тем не менее, в следующем определении под *статистической оценкой* понимается любая случайная величина вида $\hat{\theta} = h(X_1,...,X_n)$.

Определение. *Статистическая оценка* $\hat{\theta}$ *называется* ***несмещенной***, *если ее математическое ожидание* $E(\hat{\theta}) = \theta$ для всех $\theta \in \Theta$.

Поскольку статистическая оценка $\hat{\theta}$ используется для замены, как правило, неизвестного параметра θ, разность $\hat{\theta} - \theta$ далее интерпретируется как *ошибка оценки*, а ее математическое ожидание $E(\hat{\theta} - \theta)$ – как *систематическая ошибка*. Если, например, систематическая ошибка $E(\hat{\theta} - \theta) > 0$, то средняя ошибка при многократной замене θ на $\hat{\theta}$ по закону больших чисел также будет больше 0. Заметим, что несмещенную статистическую оценку

можно определить как оценку, не имеющую систематической ошибки.

Пусть $\nu_k = E(X^k)$ – генеральный, а $\hat{\nu}_k = \hat{E}(X^k) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ – выборочный начальный момент порядка k. Выборочный момент $\hat{\nu}_k$ является простейшим примером несмещенной оценки. Действительно,

$$E(\hat{\nu}_k) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^k) = \frac{n}{n}E(X_1^k) = \nu_k. \qquad (4.6)$$

В частности выборочное среднее является несмещенной оценкой генерального среднего:

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = E(X) = \nu_1. \qquad (4.7)$$

Рассмотрим теперь генеральные и выборочные центральные моменты:

$$\mu_k = E([X - E(X)]^k), \;\; \hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k.$$

Как правило, для центральных моментов $E(\hat{\mu}_k) \neq \mu_k$. Проще всего установить это для второго центрального момента (дисперсии).

Теорема 4.1. $E(\hat{\mu}_2) = \frac{n-1}{n}\mu_2$.

Доказательство. Используя $\hat{\mu}_2 = \hat{\nu}_2 - \hat{\nu}_1^2$, $\nu_2 - \nu_1^2 = \mu_2$, находим

$$\begin{aligned} E(\hat{\mu}_2) &= E(\hat{\nu}_2) - E(\hat{\nu}_1^2) \\ &= \nu_2 - \{D(\hat{\nu}_1) + E^2(\hat{\nu}_1)\} \\ &= \nu_2 - \left\{\frac{\mu_2}{n} + \nu_1^2\right\} = \frac{n-1}{n}\mu_2. \end{aligned}$$

Определение. ***Исправленная***, *или* ***несмещенная*** *выборочная дисперсия* s^2 *задается формулой*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Нетрудно проверить, что s^2 – несмещенная оценка генеральной дисперсии. Действительно

$$s^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}\hat{\mu}_2,$$

следовательно,

$$E(s^2) = \frac{n}{n-1}E(\hat{\mu}_2) = \frac{n}{n-1}\frac{n-1}{n}\mu_2 = \mu_2.$$

Замечание. Если генеральное среднее известно, $E(X) = a$, для оценки дисперсии используется статистика s_0^2, определяемая соотношением

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2.$$

Эта статистика также является несмещенной оценкой дисперсии:

$$E(s_0^2) = \frac{1}{n}\sum_{i=1}^{n}E\{(X_i - a)^2\} = \frac{1}{n}\sum_{i=1}^{n}D(X_i) = \mu_2.$$

Обсудив понятие несмещенности, перейдем теперь к рассмотрению понятия точности оценки. По аналогии со средней (квадратичной) ошибкой выборки под точностью оценки будем понимать корень из математического ожидания подходящей случайной величины, в данном случае – $(\hat{\theta} - \theta)^2$. Эта величина $\sqrt{E\{(\hat{\theta} - \theta)^2\}}$ называется *средней квадратичной ошибкой оценки*.

Определение. *Статистическая оценка называется* ***эффективной*** *в некотором классе оценок, если в этом классе при фиксированном объеме выборки она имеет наименьшую среднюю квадратичную ошибку.*

Таким образом, оценка $\hat{\theta}$ является эффективной в классе $\mathcal{K}$, если для любой оценки $\hat{\theta}' \in \mathcal{K}$ выполняется неравенство

$$E\{(\hat{\theta}-\theta)^2\} \le E\{(\hat{\theta}'-\theta)^2\}. \qquad (4.8)$$

Предположим, что оценки $\hat{\theta}$ и $\hat{\theta}'$ несмещенные. Тогда (4.8) эквивалентно более простому неравенству $D(\hat{\theta}) \le D(\hat{\theta}')$. В дальнейшем мы будем в основном рассматривать классы $\mathcal{K}$ несмещенных оценок, а для таких классов эффективность оценки $\hat{\theta} \in \mathcal{K}$ означает, что $\hat{\theta}$ имеет минимальную дисперсию среди всех оценок класса $\mathcal{K}$.

Теорема 4.2. *Если генеральная дисперсия существует, а генеральное среднее* $E(X) = a \neq 0$, *то выборочное среднее* $\overline{X}$ *является его эффективной оценкой в классе всех линейных несмещенных оценок вида*

$$\hat{a} = c_1 X_1 + c_2 X_2 + \ldots + c_n X_n.$$

Доказательство. Вследствие $E(X) = a \neq 0$, несмещенность означает, что среднее арифметическое коэффициентов $c_1, \ldots, c_n$ равно

$$\overline{c} = \frac{c_1 + \ldots + c_n}{n} = \frac{1}{n}.$$

По условию генеральная дисперсия $\sigma^2 < \infty$. Следовательно,

$$D(\hat{a}) = \sum_{i=1}^{n} c_i^2 D(X_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 = n\sigma^2 \left(\frac{1}{n} \sum_{i=1}^{n} c_i^2 \right) = n\sigma^2 \overline{c^2}.$$

Применяя неравенство Йенсена [10, часть 2] к выпуклой функции $f(c) = c^2$, получим

$$D(\hat{a}) = n\sigma^2 \overline{c^2} \ge n\sigma^2 \overline{c}^2 = n\sigma^2 \left(\frac{1}{n} \right)^2 = \frac{\sigma^2}{n} = D(\overline{X}),$$

что означает эффективность выборочного среднего $\overline{X}$ в классе всех линейных несмещенных оценок $\hat{a}$ генерального среднего $a = E(X)$.

Определение. *Статистическая оценка называется* ***состоятельной***, *если* $p\lim\hat{\theta} = \theta$ *при* $n \to \infty$.

Пусть $X_1, ..., X_n$ – выборка из распределения Бернулли

X	0	1
P	$1-p$	p

,

$\hat{p}$ – доля единиц в выборочной совокупности. Согласно одной из форм закона больших чисел (теореме Бернулли) $\hat{p}$ сходится по вероятности к p. Следовательно, $\hat{p}$ – состоятельная оценка параметра p распределения Бернулли.

Отметим, что на практике довольно часто используются смещенные и неэффективные статистические оценки, тогда как несостоятельные оценки обычно не применяются.

Лекция 5
Состоятельные оценки и метод моментов

Теорема 5.1. *Пусть последовательность случайных величин* $\{X_n\}$ *такова, что при* $n \to \infty$ *дисперсия* $D(X_n) \to 0$, *а* $E(X_n) \to \theta$, *где* θ *– некоторое число. Тогда*:

$$p\lim_{n\to\infty} X_n = \theta.$$

Доказательство. Фиксируем $\varepsilon > 0$. Так как $E(X_n) \to \theta$, то существует n_0, такое, что для $n > n_0$ выполняется неравенство $|E(X_n) - \theta| < \frac{\varepsilon}{2}$. Для всякого $n > n_0$ определим события

$$A_n = \{|X_n - \theta| \geq \varepsilon\}, \quad B_n = \{|X_n - E(X_n)| \geq \frac{\varepsilon}{2}\}.$$

Из неравенства "треугольника"

$$|X_n - E(X_n)| + |E(X_n) - \theta| \geq |X_n - \theta|$$

в случае A_n и $n > n_0$ получаем, что

$$|X_n - E(X_n)| \geq |X_n - \theta| - |E(X_n) - \theta| \geq \varepsilon - \frac{\varepsilon}{2} = \frac{\varepsilon}{2}.$$

Следовательно, событие A_n влечет B_n и $P(A_n) \leq P(B_n)$.

Применяя неравенство Чебышева, при $n \to \infty$ имеем

$$P(B_n) \leq \frac{D(X_n)}{(\varepsilon/2)^2} \to 0.$$

Отсюда находим

$$\lim_{n\to\infty} P(|X_n - \theta| < \varepsilon) = 1 - \lim_{n\to\infty} P(A_n) = 1,$$

что означает $p\lim\limits_{n\to\infty} X_n = \theta.$

Пример 5.1. Пусть $X_1, ..., X_n$ – выборка из равномерного распределения на отрезке $[0, 1]$. Требуется найти $p\lim\sqrt{\overline{X}}$, где

$$\sqrt{\overline{X}} = \frac{1}{n}\left(\sqrt{X_1} + \sqrt{X_2} + ... + \sqrt{X_n}\right)$$

Решение. Сначала находим

$$E\left(\sqrt{X_i}\right) = \int_0^1 \sqrt{x}\,dx = \frac{2}{3},$$

$$E\left(\sqrt{\overline{X}}\right) = \frac{1}{n} E\left(\sqrt{X_1} + \sqrt{X_2} + ... + \sqrt{X_n}\right) = \frac{2}{3}.$$

Далее, используя

$$D\left(\sqrt{X_i}\right) = \int_0^1 \left(\sqrt{x} - \frac{2}{3}\right)^2 dx < +\infty,$$

находим, что $D\left(\sqrt{\overline{X}}\right) = \frac{1}{n} D(X_1) \to 0$ при $n \to \infty$ и, вследствие теоремы 5.1,

$$p\lim\sqrt{\overline{X}} = \frac{2}{3}.$$

Пусть $\hat{\theta} = h_n(X_1,...,X_n)$ – оценка параметра θ распределения, определенная при любом объеме выборки n. Из теоремы 5.1 немедленно вытекает следующее

Следствие (*достаточное условие состоятельности*). *Для того чтобы оценка* $\hat{\theta}$ *была состоятельной достаточно, чтобы выполнялись условия*:

1) $\lim\limits_{n\to\infty} D(\hat{\theta}) = 0,$

2) $\lim\limits_{n\to\infty} E(\hat{\theta}) = \theta.$

Заметим, что для несмещенных оценок достаточное условие состоятельности, фактически, сводится только к первому требованию, поскольку второе выполняется тривиально.

Рассмотрим выборочное среднее $\overline{X}$. Известно (4.7), что $\overline{X}$ несмещенная оценка генерального среднего. Вследствие того, что при $n \to \infty$

$$D(\overline{X}) = \frac{D(X)}{n} \to 0,$$

из достаточного условия получаем состоятельность $\overline{X}$ как оценки генерального среднего.

Для доказательства состоятельности выборочных моментов как оценок соответствующих генеральных моментов нам потребуются две вспомогательные теоремы.

Теорема 5.2. *Если существует* $E(|Y|)$, *то существует и* $E(Y)$.

Доказательство проведем только для случая, когда существует плотность распределения $f(y)$. Имеем

$$E(|Y|) = \int\limits_{-\infty}^{\infty} |y| f(y)dy < +\infty,$$

Следовательно, интеграл $\int_{-\infty}^{\infty} yf(y)dy$ сходится абсолютно, что и означает существование $E(Y)$.

Теорема 5.3. *Если для некоторого* $m > 0$ *существует* $E(|X|^m)$, *то для любого* $k \in [0, m]$ *существует* $E(X^k)$.

Доказательство. Как и прежде, предположим, что существует плотность распределения $f(x)$. Тогда

$$E(|X|^m) = \int_{-\infty}^{\infty} |x|^m f(x)dx < +\infty,$$

если существует $E(|X|^m)$. Следовательно,

$$\begin{aligned} E(|X|^k) &= \int_{-\infty}^{\infty} |x|^k f(x)dx \\ &= \int_{|x|<1} |x|^k f(x)dx + \int_{|x|>1} |x|^k f(x)dx \\ &\le 1 + E(|X|^m) < +\infty. \end{aligned}$$

Отсюда и из теоремы 5.2 получаем существование $E(X^k)$.

Теорема 5.4. *Если генеральное распределение имеет начальный момент* ν_{2m}, *то выборочный момент* $\hat{\nu}_k$ *является состоятельной оценкой генерального момента* ν_k *при любом* $k \le m$.

Доказательство. Известно (4.6), что $E(\hat{\nu}_k) = \nu_k$, поэтому достаточно доказать, что $D(\hat{\nu}_k) \to 0$, когда объем выборки $n \to \infty$. Существование моментов ν_{2k} и ν_k следует из теоремы 5.3, поэтому при $n \to \infty$ имеем

$$\begin{aligned} D(\hat{\nu}_k) &= \frac{1}{n^2}\left\{D(X_1^k) + D(X_2^k) + \ldots + D(X_n^k)\right\} \\ &= \frac{1}{n} D(X_1^k) = \frac{1}{n}\left\{E(X_1^{2k}) - E^2(X_1^k)\right\} \\ &= \frac{1}{n}\left\{\nu_{2k} - \nu_k^2\right\} \to 0. \end{aligned}$$

Теорема 5.5 (*теорема Слуцкого*). *Если для последовательностей случайных величин* $\{X_n^{(1)}\},...,\{X_n^{(k)}\}$ *существуют конечные пределы*

$$p\lim_{n\to\infty} X_n^{(1)} = a_1, \dots, p\lim_{n\to\infty} X_n^{(k)} = a_k,$$

то для функции $f(x_1,...,x_k)$ *непрерывной в точке* $(a_1,...,a_k)$ *имеем*

$$p\lim_{n\to\infty} f(X_n^{(1)},...,X_n^{(k)}) = f(a_1,...,a_k). \qquad (5.1)$$

Доказательство. Поскольку $f(x_1,...,x_k)$ непрерывна в $(a_1,...,a_k)$, для всякого $\varepsilon > 0$ существует такое $\delta > 0$, что

$$\left(|x_1 - a_1| < \delta, ..., |x_k - a_k| < \delta\right) \Rightarrow |f(x_1,...,x_k) - f(a_1,...,a_k)| < \varepsilon. \qquad (5.2)$$

Для ε и δ удовлетворяющих (5.2), определим события:

$$A_n = \left\{|f(X_n^{(1)},...,X_n^{(k)}) - f(a_1,...,a_n)| \le \varepsilon\right\},$$

$$B_n^1 = \left\{|X_n^{(1)} - a_1| < \delta\right\}, \ \dots \ , B_n^k = \left\{|X_n^{(k)} - a_k| < \delta\right\}.$$

По условию теоремы для $i = 1,...,k$ вероятность $P(B_n^i) \to 1$ при $n \to \infty$. Отсюда с учетом $P(B_n^1...B_n^k) \ge 1 - P(\overline{B}_n^1) - ... - P(\overline{B}_n^k)$ имеем

$$\lim_{n\to\infty} P(B_n^1...B_n^k) = 1. \qquad (5.3)$$

Завершая доказательство, заметим, что (5.2) влечет $P(A_n) \ge P(B_n^1...B_n^k)$, откуда совместно с (5.3) получаем $P(A_n) \to 1$ при $n \to \infty$, что эквивалентно соотношению (5.1).

Теорема 5.6. *Если для генерального распределения существует начальный момент* ν_{2m}, *то выборочный центральный момент* $\hat{\mu}_k$ *является состоятельной оценкой* μ_k *при* $k \le m$.

Доказательство. Нетрудно видеть, что центральный момент любого порядка μ_k можно представить в виде $\mu_k = f_k(\nu_1, ..., \nu_k)$, где f_k – многочлен степени k. Например,

$$\mu_2 = \nu_2 - \nu_1^2,$$

$$\mu_3 = E([X - \nu_1]^3) = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3, \ \dots$$

При этом выборочные моменты удовлетворяют тем же самым соотношениям: $\hat{\mu}_k = f_k(\hat{\nu}_1, ..., \hat{\nu}_k)$. Поскольку многочлен f_k – непрерывная функция, из теорем 5.4, 5.5 получаем

$$\underset{n\to\infty}{p\lim}\,\hat{\mu}_k = \underset{n\to\infty}{p\lim}\, f_k(\hat{\nu}_1, ..., \hat{\nu}_k) =$$

$$= f_k\left(\underset{n\to\infty}{p\lim}\,\hat{\nu}_1, ..., \underset{n\to\infty}{p\lim}\,\hat{\nu}_k\right) = f_k(\nu_1, ..., \nu_k) = \mu_k.$$

Пусть генеральное распределение зависит от параметров $\theta_1, ..., \theta_k$. Тогда все генеральные моменты являются функциями от этих параметров:

$$\text{(I)} \quad \begin{cases} \nu_1 = g_1(\theta_1, ..., \theta_k), \\ ..... \\ \nu_k = g_k(\theta_1, ..., \theta_k). \end{cases}$$

Явный вид функций $g_i(\theta_1, ..., \theta_k)$ определяется способом задания генерального распределения. Если, скажем, распределение задано плотностью распределения вероятностей $f(x; \theta_1, ..., \theta_k)$, то для $i = 1, ..., k$ имеем

$$g_i(\theta_1, ..., \theta_k) = \int_{-\infty}^{\infty} x^i f(x; \theta_1, ..., \theta_k) dx.$$

Предположим, что систему (I) можно разрешить в виде системы

$$\text{(II)} \quad \begin{cases} \theta_1 = h_1(\nu_1, ..., \nu_k), \\ ..... \\ \theta_k = h_k(\nu_1, ..., \nu_k), \end{cases}$$

где $h_1, ..., h_k$ – непрерывные функции.

Определение. *Оценки параметров* $\theta_1, ..., \theta_k$, *заданные формулами*

$$\hat{\theta}_{1,MM} = h_1(\hat{\nu}_1, ..., \hat{\nu}_k),$$
$$.....$$
$$\hat{\theta}_{k,MM} = h_k(\hat{\nu}_1, ..., \hat{\nu}_k),$$

называются ***оценками метода моментов***.

Пример 5.2. Нормальное распределение $N(a,\sigma^2)$ имеет два параметра: $a = E(X)$ и $b = \sigma^2 = D(X)$. Требуется найти оценки этих параметров, используя метод моментов.

Решение. Находим системы (I) и (II):

$$\text{(I)} \begin{cases} \nu_1 = a, \\ \nu_2 = b + a^2, \end{cases} \quad \Rightarrow \quad \text{(II)} \begin{cases} a = \nu_1, \\ b = \nu_2 - \nu_1^2. \end{cases}$$

Заменяя генеральные моменты выборочными, получаем оценки:

$$\hat{a}_{MM} = \hat{\nu}_1, \quad \hat{b}_{MM} = \hat{\nu}_2 - \hat{\nu}_1^2 = \hat{\mu}_2.$$

Теорема 5.7. *Если распределение зависит от параметров* $\theta_1, ..., \theta_k$ *и при любом допустимом наборе их значений распределение имеет начальный момент порядка* $2k$, *то оценки метода моментов параметров* $\theta_1, ..., \theta_k$ *являются состоятельными*.

Доказательство. Применяя теоремы 5.4 и 5.5, имеем

$$\operatorname*{p\lim}_{n\to\infty} \hat{\theta}_{j,MM} = \operatorname*{p\lim}_{n\to\infty} h_j(\hat{\nu}_1, ..., \hat{\nu}_k) =$$
$$= h_j\left(\operatorname*{p\lim}_{n\to\infty} \hat{\nu}_1, ..., \operatorname*{p\lim}_{n\to\infty} \hat{\nu}_k \right) = h_j(\nu_1, ..., \nu_k) = \theta_j,$$

что означает состоятельность оценки $\hat{\theta}_{j,MM}$ для $j = 1, ..., k$.

Заметим, что для биномиального распределения и распределения Пуассона, для геометрического распределения и нормального закона, для показательного распределения и равномерного распределения на отрезке существуют начальные моменты любого порядка. Следовательно, для этих распределений состоятельность оценок метода моментов вытекает из теоремы 5.7.

Лекция 6
Метод максимального правдоподобия

В этой лекции помимо метода максимального правдоподобия рассматривается понятие информации Фишера, а также доказывается связанное с ним неравенство Рао–Крамера.

§6.1. Метод максимального правдоподобия

Предположим, что распределение случайной величины X (генеральное распределение) зависит от параметра $\theta \in \Theta$, где θ – число или вектор, Θ – множество допустимых значений параметра.

Имеется два основных случая:

1) X – дискретная случайная величина и $p(x;\theta) = P(X = x)$ – вероятность попадания X в точку x.

2) X – непрерывная случайная величина и $f(x;\theta)$ – плотность вероятности X в точке x.

В первом случае функция правдоподобия L определяется формулой

$$L(x_1, ..., x_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta), \tag{6.1}$$

во втором – формулой

$$L(x_1, ..., x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{6.2}$$

Пусть $X_1, ..., X_n$ – выборка из распределения случайной величины X. В дискретном случае $L(x_1, ..., x_n;\theta)$ – это вероятность попадания случайной точки $(X_1, ..., X_n)$ в неслучайную точку $(x_1, ..., x_n)$,

$$L(x_1, ..., x_n; \theta) = P\big((X_1, ..., X_n) = (x_1, ..., x_n)\big).$$

В непрерывном случае $L(x_1, ..., x_n; \theta)$ – это плотность распределения случайного вектора $(X_1, ..., X_n)$ в точке $(x_1, ..., x_n)$,

$$\int ... \int_B L(x_1, ..., x_n; \theta) dx_1 ... dx_n = P\big((X_1, ..., X_n) \in B\big), \quad B \subset \mathbf{R}^n.$$

В обоих случаях областью определения функции правдоподобия является $V^n \times \Theta$, где V – множество возможных значений X.

Определение. Пусть $X_1, ..., X_n$ *– выборка из распределения, зависящего от* (*многомерного*) *параметра* $\theta \in \Theta$, $h(X_1, ..., X_n)$ *– точка глобального максимума функции правдоподобия на* Θ *при фиксированных* $X_1, ..., X_n$. *Случайная* (*многомерная*) *величина* $\hat{\theta}_{МП} = h(X_1, ..., X_n)$ *называется* ***оценкой максимального правдоподобия*** *параметра* θ.

Пример 6.1. Требуется найти закон распределения оценки $\hat{\theta}_{МП}$ параметра θ распределения Бернулли

X	0	1
P	$1-\theta$	θ

,

где множество Θ допустимых значений параметра – отрезок $[0, 1]$.

Решение. Функция вероятности имеет вид

$$p(x; \theta) = P(X = x) = \begin{cases} \theta, & \text{если} \quad x = 1, \\ 1-\theta, & \text{если} \quad x = 0. \end{cases}$$

Следовательно, функции правдоподобия

$$L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta) = \theta^k (1-\theta)^{n-k},$$

где k – число единиц среди $x_1, ..., x_n$.

Найдем максимум $L(x_1, ..., x_n; \theta)$ при фиксированных $x_1, ..., x_n$. Приравнивая нулю производную

$$\frac{\partial}{\partial\theta}L(x_1,...,x_n;\theta)=\theta^{k-1}(1-\theta)^{n-k-1}(k-n\theta),$$

находим, что точка максимума $\theta_{\max}\in\{0,1,\overline{x}\}$, где

$$\overline{x}=\frac{x_1+...+x_n}{n}=\frac{k}{n}.$$

Сравнивая затем значения $L(x_1,...,x_n;0)$, $L(x_1,...,x_n;1)$ и $L(x_1,...,x_n;\overline{x})$, заключаем, что $\theta_{\max}=\overline{x}$. Следовательно, $\hat{\theta}_{МП}=\overline{X}$ – выборочное среднее.

Теперь нетрудно найти распределение $\hat{\theta}_{МП}$:

$\hat{\theta}_{ML}$	0	$\frac{1}{n}$	...	$\frac{k}{n}$	...	1
P	$(1-\theta)^n$	$n\theta(1-\theta)^{n-1}$	...	$C_n^k\theta^k(1-\theta)^{n-k}$	...	θ^n

.

Следующий пример показывает целесообразность применения в некоторых случаях $\ln L$ *– логарифмической функции правдоподобия.*

Пример 6.2. Необходимо найти оценку $\hat{\lambda}_{МП}$ для параметра λ показательного распределения с плотностью

$$f(x;\lambda)=\lambda e^{-\lambda x}\quad(x\geq 0).$$

Решение. Находим сначала функцию правдоподобия

$$L(x_1,...,x_n;\lambda)=\prod_{i=1}^{n}\lambda e^{-\lambda x_i}=\lambda^n e^{-\lambda\sum x_i}.$$

Поскольку в области определения $L(x_1,...,x_n;\lambda)>0$, имеем

$$\frac{d}{d\lambda}L(x_1,...,x_n;\lambda)=0\ \Leftrightarrow\ \frac{d}{d\lambda}\ln L(x_1,...,x_n;\lambda)=0.$$

Находим производную логарифмической функции правдоподобия

$$\frac{d}{d\lambda}\ln L(x_1,...,x_n;\lambda)=\frac{d}{d\lambda}\left(n\ln\lambda-\lambda\sum_{i=1}^{n}x_i\right)=n\left(\frac{1}{\lambda}-\overline{x}\right).$$

Следовательно, $\lambda_{\max}=\left(\overline{x}\right)^{-1}$ и $\hat{\lambda}_{ML}=\left(\overline{X}\right)^{-1}$.

§6.2. Неравенство Рао–Крамера

Следующее неравенство

$$E^2(XY) \le E(X^2)E(Y^2) \tag{6.3}$$

называется *неравенством Коши–Буняковского для случайных величин*.

Докажем соотношение (6.3). Действительно, при любом $t \in \mathbf{R}$ случайная величина $(tX+Y)^2$ принимает только неотрицательные значения. Следовательно, ее математическое ожидание $E([tX+Y]^2) \ge 0$. Отсюда вытекает, что при любом $t \in \mathbf{R}$ выполняется неравенство

$$t^2E(X^2) + 2tE(XY) + E(Y^2) \ge 0. \tag{6.4}$$

Поскольку квадратный трехчлен в левой части (6.4) имеет не более одного действительного корня, его дискриминант $E^2(XY) - E(X^2)E(Y^2) \le 0$, что и означает справедливость неравенства (6.3).

Пусть $X_1, ..., X_n$ – выборка из распределения, зависящего от скалярного параметра $\theta \in \Theta$, $L(x_1, ..., x_n; \theta)$ – соответствующая функция правдоподобия.

Случаи дискретного (6.1) и непрерывного (6.2) распределений рассматриваются вполне аналогично, поэтому мы ограничимся только непрерывным распределением. В этом случае $L(x_1, ..., x_n; \theta)$ – плотность вероятности случайного вектора $(X_1, ..., X_n)$, следовательно,

$$\int ... \int_{\mathbf{R}^n} L(x_1, ..., x_n; \theta) dx_1 ... dx_n = 1.$$

Найдем математическое ожидание производной по θ логарифмической функции правдоподобия в точке $(X_1, ..., X_n)$:

$$\begin{aligned} E\left(\frac{\partial}{\partial\theta}\ln L(X_1,...,X_n;\theta)\right) &= \int ... \int_{\mathbf{R}^n} \frac{\partial}{\partial\theta}\ln L(x_1,...,x_n;\theta) L(x_1,...,x_n;\theta) dx_1 .. dx_n \\ &= \int ... \int_{\mathbf{R}^n} \frac{\partial}{\partial\theta} L(x_1,...,x_n;\theta) dx_1 .. dx_n \qquad (6.5) \\ &= \frac{\partial}{\partial\theta} \int ... \int_{\mathbf{R}^n} L(x_1,...,x_n;\theta) dx_1 .. dx_n = 0. \end{aligned}$$

Для несмещенной оценки $\hat{\theta} = h(X_1, ..., X_n)$ – параметра θ имеем

$$\theta = E(\hat{\theta}) = \int ... \int_{\mathbf{R}^n} h(x_1, ..., x_n) L(x_1, ..., x_n; \theta) dx_1 ... dx_n.$$

Дифференцируя данное неравенство по θ, получим

$$1 = \frac{\partial}{\partial \theta} \int ... \int_{\mathbf{R}^n} h(x_1, ..., x_n) L(x_1, ..., x_n; \theta) dx_1 ... dx_n$$

$$= \int ... \int_{\mathbf{R}^n} h(x_1, ..., x_n) \frac{\partial}{\partial \theta} \ln L(x_1, ..., x_n; \theta) L(x_1, ..., x_n; \theta) dx_1 ... dx_n$$

$$= E\left(\hat{\theta} \frac{\partial}{\partial \theta} \ln L(X_1, ..., X_n; \theta) \right).$$

С учетом (6.5) из последнего равенства следует, что

$$1 = E\left((\hat{\theta} - \theta) \frac{\partial}{\partial \theta} \ln L(X_1, ..., X_n; \theta) \right) = E^2\left((\hat{\theta} - \theta) \frac{\partial}{\partial \theta} \ln L(X_1, ..., X_n; \theta) \right).$$

Отсюда, используя неравенство Коши–Буняковского (6.3), получаем

$$1 \le E\left((\hat{\theta} - \theta)^2 \right) E\left\{ \left(\frac{\partial}{\partial \theta} \ln L(X_1, ..., X_n; \theta) \right)^2 \right\}.$$

Следовательно,

$$D(\hat{\theta}) \ge \frac{1}{I(\theta)}, \tag{6.6}$$

где

$$I(\theta) = E\left\{ \left(\frac{\partial}{\partial \theta} \ln L(X_1, ..., X_n; \theta) \right)^2 \right\} = D\left(\frac{\partial}{\partial \theta} \ln L(X_1, ..., X_n; \theta) \right).$$

Определение. *Функция* $I(\theta)$ *называется* ***информацией Фишера***, *а неравенство* (6.6) – ***неравенством Рао–Крамера***.

Смысл функции $I(\theta)$ состоит в том, что она характеризует количество информации о значении параметра θ распределения, содержащееся в выборке определенного объема. Если информация Фишера мала, то, как это следует из неравенства Рао-Крамера, не существует несмещенной оценки $\hat{\theta}$ с малой среднеквадратичной ошибкой в равенстве $\theta \approx \hat{\theta}$.

Пример 6.3. Запишите неравенство Рао–Крамера для оценки параметра a нормального распределения $N(a,\sigma^2)$ с плотностью

$$f(x;a,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (6.7)$$

Решение. Сначала находим функцию правдоподобия L,

$$L=\frac{1}{\sigma^n(2\pi)^{n/2}}\prod_{i=1}^{n}e^{-\frac{(x_i-a)^2}{2\sigma^2}},$$

затем – логарифмическую функцию правдоподобия $\ln L$,

$$\ln L=-n\ln\sigma-\frac{n}{2}\ln 2\pi-\sum_{i=1}^{n}\frac{(x_i-a)^2}{2\sigma^2}. \qquad (6.8)$$

Продифференцировав $\ln L$,

$$\frac{\partial}{\partial a}\ln L=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-a)=\frac{n}{\sigma^2}(\bar{x}-a),$$

находим информацию Фишера:

$$I(a)=D\left(\frac{\partial}{\partial a}\ln L\right)=D\left\{\frac{n}{\sigma^2}\left(\overline{X}-a\right)\right\}$$

$$=\frac{n^2}{\sigma^4}D\left(\overline{X}-a\right)=\frac{n^2}{\sigma^4}\frac{\sigma^2}{n}=\frac{n}{\sigma^2}.$$

Отсюда получаем неравенство

$$D(\hat{a})\geq\frac{\sigma^2}{n}, \qquad (6.9)$$

соответствующее неравенству Рао–Крамера (6.6).

Поскольку $\overline{X}$ имеет дисперсию $D(\overline{X})=\dfrac{\sigma^2}{n}$ и является несмещенной оценкой $E(X)$, из (6.9) следует, что *выборочное среднее – это эффективная оценка генерального среднего нормального распределения в классе всех несмещенных оценок*.

Пример 6.4. Запишите неравенство Рао–Крамера для оценки параметра σ^2 нормального распределения с плотностью (6.7).

Решение. Дифференцируя (6.8) по σ^2, имеем

$$\frac{\partial}{\partial \sigma^2} \ln L = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - a)^2.$$

Следовательно, информация Фишера о параметре σ^2

$$\begin{aligned} I(\sigma^2) &= \frac{n}{4\sigma^8} D\left\{(X-a)^2\right\} \\ &= \frac{n}{4\sigma^8} \left\{E\left((X-a)^4\right) - E^2\left((X-a)^2\right)\right\} \\ &= \frac{n}{4\sigma^8} \left\{3\sigma^4 - \left(\sigma^2\right)^2\right\} = \frac{n}{2\sigma^4}. \end{aligned}$$

Отсюда получаем неравенство Рао–Крамера для $\hat{\sigma}^2$:

$$D(\hat{\sigma}^2) \geq \frac{2\sigma^4}{n}.$$

Существует ли оценка $\hat{\sigma}^2$ с минимально возможной дисперсией $\frac{2\sigma^4}{n}$? Если генеральное среднее известно $E(X) = a$, такая оценка $\hat{\sigma}^2$ существует:

$$\hat{\sigma}^2 = s_0^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - a)^2.$$

Действительно,

$$\begin{aligned} D\left(s_0^2\right) &= \frac{1}{n^2} D\left(\sum_{i=1}^{n} (X_i - a)^2 \right) \\ &= \frac{1}{n} \left\{E\left((X-a)^4\right) - E^2\left((X-a)^2\right)\right\} \\ &= \frac{1}{n} \left\{3\sigma^4 - \sigma^4\right\} = \frac{2\sigma^4}{n}. \end{aligned}$$

Следовательно, $\hat{\sigma}^2 = s_0^2$ – эффективная оценка дисперсии σ^2 нормального распределения $N(a, \sigma^2)$ с известным генеральным средним a.

Лекция 7
Доверительные интервалы

Помимо рассмотренных в предыдущей главе точечных оценок в математической статистике изучаются так называемые *интервальные* оценки параметров распределений. В отличие от точечной оценки, интервальная оценка представляет собой случайный промежуток, накрывающий оцениваемый параметр с вероятностью близкой к 1.

Пусть $\vec{X}_n = (X_1, ..., X_n)$ – случайная выборка объема n из некоторого распределения, зависящего от параметра θ, $I(\vec{X}_n)$ – определяемый по выборке случайный интервал. Конечные интервалы $I(\vec{X}_n)$ задаются при помощи двух статистик: $I(\vec{X}_n) = \left(\underline{\theta}(\vec{X}_n), \overline{\theta}(\vec{X}_n)\right)$. Аналогично задаются и бесконечные интервалы: $I(\vec{X}_n) = \left(-\infty, \overline{\theta}(\vec{X}_n)\right)$, $I(\vec{X}_n) = \left(\underline{\theta}(\vec{X}_n), +\infty\right)$.

Для реализации $\vec{x}_n = (x_1, ..., x_n)$ случайной выборки $\vec{X}_n$ интервал $I(\vec{x}_n)$ является неслучайным и для одних реализаций он накрывает параметр θ, для других – не накрывает. Таким образом, $\{\theta \in I(\vec{X}_n)\}$ – случайное событие с некоторой вероятностью наступления $P\{\theta \in I(\vec{X}_n)\}$. Вероятность $P\{\theta \in I(\vec{X}_n)\}$ является показателем надежности интервальной оценки, и обычно стремятся к тому, чтобы эта вероятность была близка к 1.

§7.1. Понятие доверительного интервала

Отметим, что на практике истинное значение θ не известно и, следовательно, в общем случае невозможно вычислить $P\{\theta \in I(\vec{X}_n)\}$. Однако то, что невозможно в общем случае, удается сделать в частном случае. Действительно, существуют оценки $I(\vec{X}_n)$, для которых вероятность $P\{\theta \in I(\vec{X}_n)\}$ одинакова при лю-

бых допустимых θ. Именно для таких оценок вероятность $P\{\theta \in I(\vec{X}_n)\}$ определяется однозначно при неизвестном θ.

Определение. *Промежуток* $I(\vec{X}_n)$ *называется* ***γ-доверительным интервалом*** *для параметра* θ, *если при любом допустимом значении* θ *вероятность* $P\{\theta \in I(\vec{X}_n)\} = \gamma$. *Число* γ *при этом называется* ***доверительной вероятностью***.

В дальнейшем соотношение $\theta \in I(\vec{X}_n)$ (возможно записанное в виде одного или двух неравенств) также называется *γ-доверительным интервалом* или *γ-доверительной оценкой*.

Конечный γ-доверительный интервал

$$\underline{\theta}(\vec{X}_n) < \theta < \overline{\theta}(\vec{X}_n) \tag{7.1}$$

называется по-другому *двусторонней γ-доверительной оценкой*, а бесконечные γ-доверительные интервалы

$$\underline{\theta}(\vec{X}_n) < \theta \tag{7.2}$$

и

$$\theta < \overline{\theta}(\vec{X}_n) \tag{7.3}$$

– односторонними оценками. При этом интервал (7.2) называется *доверительной оценкой снизу*, а интервал (7.3) – *доверительной оценкой сверху*.

Замечание. Необходимо отметить, что неслучайный γ-доверительный интервал $I(\vec{x}_n)$ либо накрывает, либо не накрывает параметр θ. Соответственно, вероятность $P(\theta \in I(\vec{x}_n))$ равна либо 1, либо 0, поэтому в задачах на построение γ-доверительных интервалов по числовым данным вместо термина *доверительная вероятность* иногда употребляется термин *надежность*.

Для γ-доверительного интервала $(\underline{\theta}, \overline{\theta})$ его середина может рассматриваться как точечная оценка параметра θ,

$$\hat{\theta} = \frac{1}{2}\left(\underline{\theta} + \overline{\theta}\right). \tag{7.4}$$

Нетрудно видеть, что для оценки (7.4) абсолютная ошибка в приближенном равенстве $\theta \approx \hat{\theta}$ меньше половины длины интервала $(\underline{\theta}, \overline{\theta})$ с вероятностью γ. Вследствие этого, величина $\frac{1}{2}\left(\overline{\theta} - \underline{\theta}\right)$ называется *точностью* доверительной оценки (7.1).

§7.2. Построение доверительного интервала методом центральной статистики

Пусть $g_{\vec{x}_n}(\theta)$ – некоторая функция, зависящая от векторного параметра $\vec{x}_n$, $\vec{X}_n = (X_1, ..., X_n)$ – выборка объема n из распределения, зависящего от параметра θ. Заменяя $\vec{x}_n$ на случайный вектор $\vec{X}_n$, получим случайную величину $Y = g_{\vec{X}_n}(\theta)$. Если распределение Y не зависит от θ, случайная величина $Y = g_{\vec{X}_n}(\theta)$ называется *центральной статистикой*.

При построении γ-доверительного интервала *методом центральной статистики* предположим, что $g_{\vec{x}_n}(\theta)$ – непрерывная и возрастающая (убывающая) функция от θ при любом фиксированном $\vec{x}_n$. Тогда при любом $\vec{x}_n$ для функции $y = g_{\vec{x}_n}(\theta)$ существует обратная монотонная функция $g^{-1}_{\vec{x}_n}(y)$. Поскольку $g^{-1}_{\vec{x}_n}(g_{\vec{x}_n}(\theta)) = \theta$, для случайной величины Y имеет место аналогичное равенство

$$g^{-1}_{\vec{X}_n}(Y) = \theta. \qquad (7.5)$$

Предположим дополнительно, что распределение Y непрерывно. Используя такую центральную статистику Y, нетрудно построить γ-доверительный интервал для θ. Действительно, распределение Y не зависит от θ и непрерывно, поэтому даже не зная истинного значения θ, для любого γ можно подобрать интервал (a, b) так, чтобы центральная статистика Y попадала в этот интервал с заранее заданной вероятностью γ,

$$P(a < Y < b) = \gamma.$$

Подробнее процедура выбора интервала (a,b) рассматривается в §7.4, где, в частности, показано, что этот выбор можно произвести бесконечным числом способов. Сейчас же будем считать, что интервал (a,b) тем или иным образом уже выбран.

Предположим, для определенности, что функция $y = g_{\vec{x}_n}(\theta)$ является возрастающей при любом $\vec{x}_n$. Тогда функция $\theta = g^{-1}_{\vec{x}_n}(y)$ также является возрастающей, вследствие чего событие $\{a < Y < b\}$ эквивалентно событию $\{g^{-1}_{\vec{X}_n}(a) < g^{-1}_{\vec{X}_n}(Y) < g^{-1}_{\vec{X}_n}(b)\}$. Отсюда, с учетом (7.5), получаем вероятность двойного неравенства

$$P\{g^{-1}_{\vec{X}_n}(a) < \theta < g^{-1}_{\vec{X}_n}(b)\} = \gamma.$$

Следовательно, интервал $(\underline{\theta},\overline{\theta}) = (g^{-1}_{\vec{X}_n}(a),\, g^{-1}_{\vec{X}_n}(b))$ накрывает параметр θ с независящей от θ вероятностью γ, т.е. является γ-доверительным интервалом.

Аналогичные рассуждения показывают, что в случае, когда функция $y = g_{\vec{x}_n}(\theta)$ является убывающей при любом $\vec{x}_n$, γ-доверительным интервалом является интервал вида

$$(\underline{\theta},\overline{\theta}) = (g^{-1}_{\vec{X}_n}(b),\, g^{-1}_{\vec{X}_n}(a)).$$

Завершая описание метода центральной статистики, заметим, что этот метод применим и в тех случаях, когда накладываемые на центральную статистику дополнительные требования выполняются с вероятностью 1. Так, функция $y = g_{\vec{x}_n}(\theta)$ может и не быть возрастающей при любом $\vec{x}_n$. Однако, если для $\vec{X}_n = (X_1,...,X_n)$ функция $g_{\vec{X}_n}(\theta)$ является возрастающей с вероятностью 1, интервал $(g^{-1}_{\vec{X}_n}(a),\, g^{-1}_{\vec{X}_n}(b))$ все-таки будет γ-доверительным интервалом для θ.

§7.3. Квантили и процентные точки стандартных статистических распределений

Напомним определение некоторых распределений, играющих важную роль в математической статистике. Пусть $Z_1,...,Z_k$ – независимые, распределенные по стандартному нормальному закону $N(0,1)$ случайные величины. Распределение суммы квадратов $Z_1^2+...+Z_k^2$ называется распределением χ^2 с k степенями свободы и обозначается $\chi^2(k)$. Распределение $t(k)$ Стьюдента с k степенями свободы определяется как распределение случайной величины $X/\sqrt{Y/k}$, где X и Y независимы и $X \sim N(0,1), Y \sim \chi^2(k)$. Распределение $F(k,l)$ Фишера с k и l степенями свободы определяется как распределение отношения $\frac{X/k}{Y/l}$, где X и Y независимы и $X \sim \chi^2(k), Y \sim \chi^2(l)$. В теории вероятностей находится плотность $f(x)$ для каждого из распределений $N(0,1), \chi^2(k), t(k)$ и $F(k,l)$. Из явного вида $f(x)$ следует, что в случае распределений $N(0,1), t(k)$ плотность $f(x)>0$ и непрерывна на всей числовой оси, а для распределений $\chi^2(k)$ и $F(k,l)$ положительна и непрерывна на интервале $(0,+\infty)$ и тождественно равна нулю на $(-\infty,0)$. Из существования плотности $f(x)$ следует непрерывность функции распределения $F(x)=\int_{-\infty}^{x} f(t)dt$ для распределений всех четырех видов. Кроме того, и указанных условий положительности вытекает, что $F(x)$ возрастает от 0 до 1 на интервале $(-\infty,+\infty)$ для распределений $N(0,1), t(k),$ а для распределений $\chi^2(k), F(k,l)$ – на интервале $(0,+\infty)$.

Напомним также, что q-квантиль, или квантиль уровня q – это, по определению, корень уравнения $F(x)=q$. Из перечисленных свойств функции распределений $N(0,1), \chi^2(k), t(k)$ и $F(k,l)$ вытекает, что для любого $q \in (0,1)$ уравнение $F(x)=q$ имеет единственное решение. Следовательно, для этих распределений q-

квантиль определяется однозначно по уровню q. Квантили уровня $q = 1 - \alpha$ называются 100α-*процентными точками* и обозначаются следующим образом:

распределение	$N(0,1)$	$\chi^2(k)$	$t(k)$	$F(k,l)$
процентная точка	Z_α	$\chi^2_\alpha(k)$	$t_\alpha(k)$	$F_\alpha(k,l)$.

Предположим, что x_α является 100α-процентной точкой распределения непрерывной случайной величины X. Тогда x_α – это также квантиль уровня $(1-\alpha)$ и, следовательно,

$$P(X > x_\alpha) = 1 - P(X < x_\alpha) = 1 - (1-\alpha)) = \alpha.$$

Таким образом, соотношение $P(X > x_\alpha) = \alpha$ можно принять за определение процентной точки x_α. В дальнейшем везде, где могли бы использоваться как квантили так и процентные точки, мы, как правило, используем процентные точки.

Свойства процентных точек

1) $Z_{1-\alpha} = -Z_\alpha$;
2) $t_{1-\alpha}(k) = -t_\alpha(k)$;
3) $F_{1-\alpha}(k,l) = [F_\alpha(k,l)]^{-1}$;
4) $t_\alpha(k) \approx Z_\alpha,\ k > 30$;
5) $\chi^2_\alpha(k) \approx (Z_\alpha + \sqrt{2k-1})^2/2,\ k > 30$.

Первые три свойства легко доказываются. Выведем, например, свойство 1. Пусть $X \sim N(0,1)$. Тогда и $-X \sim N(0,1)$. Следовательно,

$$P(X > -Z_\alpha) = P(-X < Z_\alpha) = P(X < Z_\alpha) = 1 - \alpha.$$

Отсюда вытекает, что $-Z_\alpha$ является $100(1-\alpha)$-процентной точкой распределения $N(0,1)$.

Свойство 1 можно также доказать графически, используя четность функции $f(x)$ плотности распределения $N(0,1)$. Действительно, площадь под графиком плотности справа от Z_α равна α. Вследствие четности $f(x)$ площадь под графиком $f(x)$ слева от $-Z_\alpha$ также равна α. С учетом того, что площадь под всем графи-

ком $f(x)$ равна 1, отсюда получаем, что площадь под графиком $f(x)$ справа от $-Z_\alpha$ равна 1–α, т.е. $-Z_\alpha = Z_{1-\alpha}$.

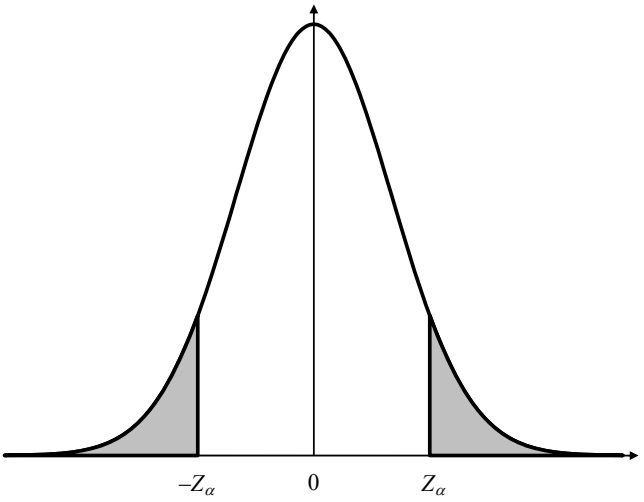


Рис. 7.1.

Лекция 8
Интервальные оценки параметров нормального распределения

Пусть $X_1, ..., X_n$ – выборка из нормального распределения $N(\mu, \sigma^2)$. Как известно, параметр μ равен математическому ожиданию, а σ^2 – дисперсии, $E(X_i) = \mu$, $D(X_i) = \sigma^2$. Поскольку распределение $N(\mu, \sigma^2)$ в отношении выборки $X_1, ..., X_n$ играет роль генеральной совокупности, назовем μ генеральным средним, а σ^2 – генеральной дисперсией. Значение случайной величины X_i бу-

дем интерпретировать как значение признака X на i-том элементе выборочной совокупности.

§7.4. Интервальная оценка математического ожидания

Вследствие того, что для выборочного среднего $\overline{X}$ его математическое ожидание $E(\overline{X})$ равно генеральному среднему, а дисперсия $D(\overline{X}) = \sigma^2 / n$, с учетом нормальности $\overline{X}$, имеем $\overline{X} \sim N(\mu, \sigma^2)$. Соответственно, случайная величина

$$Y = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

распределена по стандартному нормальному закону $N(0,1)$. Поскольку распределение Y не зависит от μ, Y – центральная статистика. Кроме того, зависимость Y от μ является убывающей, что позволяет использовать эту центральную статистику для построения доверительной интервальной оценки для генерального среднего μ (см. §7.2), если, конечно, значение σ известно.

1. Оценка генерального среднего при известной дисперсии

В качестве первого шага построения γ-доверительного интервала для μ методом центральной статистики выберем положительные δ и ε так, чтобы $\delta + \varepsilon = 1 - \gamma$. Рассмотрим интервал $(-Z_\delta, Z_\varepsilon)$. Поскольку $Y \sim N(0,1)$ и $-Y \sim N(0,1)$, имеем

$$P(Y \in (-Z_\delta, Z_\varepsilon)) = 1 - P(Y < -Z_\delta) - P(Y > Z_\varepsilon) =$$
$$= 1 - P(-Y > Z_\delta) - \varepsilon = 1 - \delta - \varepsilon = \gamma.$$

Согласно методу центральной статистики в качестве нижней доверительной границы для μ следует взять

$$\underline{\mu} = g^{-1}_{\overline{X}_n}(Z_\varepsilon),$$

где $g^{-1}_{\overline{X}_n}(y)$ – функция, обратная к убывающей функции

$$g_{\bar{X}_n}(\mu) = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}.$$

Другими словами, необходимо найти $\underline{\mu}$ из уравнения

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = Z_\varepsilon. \tag{7.6}$$

Решая (7.6), получим

$$\underline{\mu} = \overline{X} - Z_\varepsilon \frac{\sigma}{\sqrt{n}}.$$

Верхняя граница доверительного интервала имеет вид

$$\overline{\mu} = g^{-1}_{\bar{X}_n}(-Z_\delta),$$

поэтому значение $\overline{\mu}$ определяется уравнением

$$\frac{\overline{X} - \overline{\mu}}{\sigma / \sqrt{n}} = -Z_\delta. \tag{7.7}$$

Решая (7.7), находим

$$\overline{\mu} = \overline{X} + Z_\delta \frac{\sigma}{\sqrt{n}}.$$

Обозначив $\delta + \varepsilon$ через α, приходим к следующей $(1-\alpha)$-доверительной оценке генерального среднего μ нормально распределенного признака X:

$$\overline{X} - Z_\varepsilon \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_\delta \frac{\sigma}{\sqrt{n}}. \tag{7.8}$$

Заметим, что при фиксированной доверительной вероятности $\gamma = 1 - \alpha$ формула (7.8) дает бесконечное множество различных $(1-\alpha)$-доверительных оценок, поскольку в качестве δ можно взять любое число из интервала $(0, \alpha)$. При этом, как нетрудно видеть, граничные значения $\delta = 0$ и $\delta = \alpha$ приводят к односторонним оценкам. Если, например, $\delta \to 0$, то $Z_\delta \to +\infty$ и (7.8) превращается в $(1-\alpha)$-доверительную оценку снизу:

$$\overline{X} - Z_\alpha \frac{\sigma}{\sqrt{n}} < \mu. \tag{7.9}$$

Если же $\delta \to \alpha$, то $Z_\varepsilon \to +\infty$ и (7.8) превращается в $(1-\alpha)$-доверительную оценку сверху:

$$\mu < \overline{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}. \tag{7.10}$$

Определение. *Двусторонняя доверительная оценка* $\underline{\theta} < \theta < \overline{\theta}$ *называется* ***симметричной по вероятности***, *если*

$$P(\theta < \underline{\theta}) = P(\theta < \overline{\theta}). \tag{7.11}$$

Симметричные по вероятности доверительные оценки используются наиболее часто, поскольку они легко находятся, а во многих случаях соответствующий γ-доверительный интервал $(\underline{\theta}, \overline{\theta})$ имеет наименьшую длину среди прочих γ-доверительных интервалов, обеспечивая тем самым наилучшую точность оценки.

Условие (7.11) для оценки (7.8) означает, что $\delta = \varepsilon$. Следовательно, симметричная по вероятности $(1-\alpha)$-доверительная оценка генерального среднего имеет вид:

$$\overline{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}. \tag{7.12}$$

2. Оценка генерального среднего при неизвестной дисперсии

Рассмотренные выше оценки (7.8)–(7.10), (7.12) при неизвестной дисперсии σ^2 практически бесполезны. Чтобы найти оценки, не использующие σ^2, заменим в центральной статистике Y стандартное отклонение σ на s – корень из исправленной выборочной дисперсии

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Полученная таким образом статистика

$$T = \frac{\overline{X} - \mu}{s / \sqrt{n}} \tag{7.13}$$

является центральной. Этот факт является очевидным следствием следующей теоремы, которую мы приводим без доказательства.

Теорема 7.1. *Если* $X_1,...,X_n$ *независимы и распределены по нормальному закону* $N(\mu,\sigma^2)$, *то отношение* (7.13) *распределено по закону Стьюдента с* $n-1$ *степенями свободы*, $T \sim t(n-1)$.

Построение доверительных интервалов для μ, основанное на центральной статистике T, вполне аналогично проведенному ранее построению доверительных интервалов, основанному на статистике Y. Отличие состоит лишь в том, что σ заменяется на s, а вместо процентных точек стандартного нормального распределения используются процентные точки распределения Стьюдента. Приведем наиболее употребительные доверительные оценки μ при неизвестной дисперсии σ^2.

Двусторонняя симметричная по вероятности $(1-\alpha)$-доверительная оценка генерального среднего имеет вид:

$$\overline{X} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}. \qquad (7.14)$$

Соответственно, односторонние оценки генерального среднего записываются следующим образом:

$$\overline{X} - t_{\alpha}(n-1)\frac{s}{\sqrt{n}} < \mu \qquad (7.15)$$

– $(1-\alpha)$-доверительная оценка снизу;

$$\mu < \overline{X} + t_{\alpha}(n-1)\frac{s}{\sqrt{n}} \qquad (7.16)$$

– $(1-\alpha)$-доверительная оценка сверху.

При рассмотрении всех дальнейших примеров, если не оговорено противное, двусторонние доверительные оценки предполагаются симметричными по вероятности.

Пример 7.1. Распределенная по нормальному закону $N(\mu,100)$ случайная величина X в 4 независимых испытаниях приняла значения 3, 7, 9 и 5. Требуется найти 90%-ую доверительную оценку μ.

Решение. Значения X до проведения испытаний можно рассматривать как независимые случайные величины $X_1, X_2, X_3, X_4 \sim N(\mu,100)$. Таким образом, $\vec{X}_4 = (X_1, X_2, X_3, X_4)$ – выборка из распределения $N(\mu,100)$. В условиях примера дана реализация этой выборки

$$\vec{x}_4 = (x_1, x_2, x_3, x_4) = (3, 7, 9, 5).$$

Поскольку генеральное стандартное отклонение $\sigma = 10$, вычислив выборочное среднее $\bar{x} = 6$, из формулы (7.12) получим оценку

$$6 - 5Z_{\alpha/2} < \mu < 6 + 5Z_{\alpha/2}.$$

Для требуемой доверительной вероятности $\gamma = 0{,}9$ имеем $\alpha = 1 - \gamma = 0{,}1$ и $\alpha/2 = 0{,}05$. Из таблицы значений функции Лапласа $\Phi(x)$ находим

$$Z_{0{,}05} = \Phi^{-1}(0{,}45) \approx 1{,}64.$$

Следовательно, искомая 0,9-доверительная оценка имеет вид

$$-2{,}2 < \gamma < 14{,}2.$$

При рассмотрении нескольких примеров, связанных с динамикой цен на активы (обычно, акции), мы будем исходить из модели *логарифмически нормального случайного блуждания*, известной также как модели *геометрического броуновского движения с дискретным временем*. Суть этой модели сводится к следующему: цена актива S_t в момент времени $t = 0,1,...,n$ рассматривается как случайная величина, при этом считается, что логарифмы относительных изменений цены

$$h_t = \ln \frac{S_t}{S_{t-1}}, \quad t = 1,...,n$$

независимы и распределены по нормальному закону $N(\mu,\sigma^2)$ (историю вопроса и обзор более сложных моделей см. в [11]).

Величина h_t характеризует доходность актива с точки зрения непрерывного начисления процентов и в дальнейшем называется *логарифмической доходностью* (*logarithmical return*) *за пери-*

од $[t-1, t]$. Среднее квадратичное отклонение σ называется *волатильностью* актива (за время 1).

Пример 7.2. Логарифмическая доходность за 5 дневных периодов (т.е. от закрытия до закрытия торгов) составила: 0,02; 0,01; –0,03; 0,01; –0,02. Предположим, что эти данные являются реализацией случайной выборки из $N(\mu, \sigma^2)$ с неизвестными параметрами μ и σ^2. Требуется найти 95-процентный доверительный интервал для μ.

Решение. Сначала находим выборочное среднее и исправленную дисперсию:

$$\bar{h} = \frac{1}{5}(0{,}02 + 0{,}01 - 0{,}03 + 0{,}01 - 0{,}02) = -0{,}002;$$

$$s^2 = \frac{1}{4}\{(0{,}02 - \bar{h})^2 + (0{,}01 - \bar{h})^2 + \ldots + (-\,0{,}02 - \bar{h})^2\} = 0{,}00047.$$

Поскольку в условии примера не указан вид интервала, найдем симметричный по вероятности доверительный интервал. Соответствующая процентная точка $t_{0{,}025}(4) = 2{,}776$. Из (7.14) получаем искомый доверительный интервал:

$$-0{,}0289 < \mu < 0{,}0249. \qquad (7.17)$$

Выбор симметричного по вероятности доверительного интервала не всегда является самым лучшим решением. Так, осторожный инвестор, которого больше волнуют возможные потери, чем аналогичные по величине возможные прибыли, в условиях примера 7.2 может предпочесть доверительную оценку μ снизу.

Пример 7.3. В условиях предыдущего примера требуется найти 0,95-доверительную оценку снизу математического ожидания μ логарифмической доходности.

Решение. Находим процентную точку

$$t_\alpha(n-1) = t_{0{,}05}(4) = 2{,}132.$$

Подставив $t_\alpha(n-1)$ и найденные при решении примера 7.2 значения s и $\bar{h}$ в (7.15), получим 0,95-доверительную оценку μ снизу:

$$-0{,}0227 < \mu. \qquad (7.18)$$

Сравнивая решения примеров 7.2 и 7.3, видим, что нижняя граница доверительного интервала (7.17) оказалась меньше нижней границы (7.18). Таким образом, (7.18) дает более точную оценку μ снизу, чем (7.17) (но при этом не дает никакой оценки μ сверху).

§7.5. Интервальная оценка дисперсии

Пусть $X_1, ..., X_n$ – выборка из нормального распределения $N(\mu, \sigma^2)$. Как и в предыдущем разделе μ называется генеральным средним, а σ^2 – генеральной дисперсией. Так же как интервальная оценка генерального среднего зависела от того, известна или нет генеральная дисперсия, так и интервальная оценка генеральной дисперсии производится различным образом в зависимости от того, известно или нет генеральное среднее.

1. Оценка дисперсии при известном генеральном среднем

При известном μ существует эффективная точечная оценка дисперсии

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Определим при помощи s_0^2 центральную статистику T_0 соотношением

$$T_0 = \frac{ns_0^2}{\sigma^2}. \qquad (7.19)$$

Вследствие того, что T_0 можно представить в виде суммы

$$T_0 = Z_1^2 + ... + Z_n^2$$

квадратов независимых случайных величин

$$Z_i = \frac{X_i - \mu}{\sigma}, \quad i = 1,...,n,$$

распределенных по закону $N(0,1)$, распределением T_0 является $\chi^2(n)$.

Поскольку $P(s_0^2 > 0) = 1$, зависимость T_0 от σ^2 является убывающей с вероятностью 1, что позволяет использовать T_0 в качестве центральной статистики для построения доверительной оценки σ^2. Приведем основные, полученные методом центральной статистики $(1-\alpha)$-доверительные оценки дисперсии:

$$\frac{ns_0^2}{\chi^2_{\alpha/2}(n)} < \sigma^2 < \frac{ns_0^2}{\chi^2_{1-\alpha/2}(n)} \tag{7.20}$$

– симметричная по вероятности оценка σ^2;

$$\sigma^2 < \frac{ns_0^2}{\chi^2_{1-\alpha}(n)} \tag{7.21}$$

– доверительная оценка σ^2 сверху и

$$\frac{ns_0^2}{\chi^2_{\alpha}(n)} < \sigma^2 \tag{7.22}$$

– доверительная оценка σ^2 снизу.

2. Оценка дисперсии при неизвестном генеральном среднем

Предположим, что генеральное среднее μ не известно. Новая центральная статистика T, так же как и T_0, строится на основе несмещенной точечной оценки σ^2, но на этот раз используется исправленная выборочная дисперсия s^2:

$$T = \frac{(n-1)s^2}{\sigma^2}.$$

Распределение центральной статистики T устанавливается следующей теоремой, которая приводится без доказательства.

Теорема 7.2. *Если* $X_1,...,X_n$ *независимы и распределены по нормальному закону* $N(\mu,\sigma^2)$, *то отношение*

$$T = \frac{(X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{\sigma^2}$$

распределено по закону χ^2 *с* $n-1$ *степенями свободы*.

Применяя метод центральной статистики, получим (1–α)-доверительные оценки дисперсии при неизвестном генеральном среднем:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \tag{7.23}$$

– симметричная по вероятности оценка σ^2;

$$\sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)} \tag{7.24}$$

– (1–α)-доверительная оценка σ^2 сверху и

$$\frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)} < \sigma^2 \tag{7.25}$$

– (1–α)-доверительная оценка σ^2 снизу.

Пример 7.4. По результатам 250 наблюдений дневной логарифмической доходности h некоторой акции найдено выборочное стандартное отклонение $\hat{\sigma}(h) = 0{,}02$. Требуется, используя модель геометрического броуновского движения с дискретным временем, построить двустороннюю 0,95-доверительную оценку волатильности σ.

Решение. Для применения (7.23) необходимо вычислить $(n-1)s^2$, но в условии примера дано $\hat{\sigma}$. Чтобы избежать вычисления s^2 заметим, что

$$(n-1)s^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 = n\hat{\sigma}^2.$$

В нашем примере $n = 250$. Следовательно,

$$249s^2 = 250\hat{\sigma}^2 = 250 \cdot 0{,}02^2 = 0{,}1.$$

Из (7.23) получаем двустороннюю оценку генеральной дисперсии σ^2

$$\frac{0{,}1}{\chi^2_{0{,}025}(249)} < \sigma^2 < \frac{0{,}1}{\chi^2_{0{,}975}(249)}.$$

Процентные точки с числом степеней $k = 249$ отсутствуют в таблице процентных точек распределения χ^2, поэтому воспользуемся приближенной формулой (свойство 5 на стр. 62)

$$\chi^2_\alpha(k) \approx \frac{(Z_\alpha + \sqrt{2k-1})^2}{2} = \frac{(Z_\alpha + 22{,}29)^2}{2},$$

что для $\alpha = 0{,}025$ и $\alpha = 0{,}975$ дает

$$\chi^2_{0{,}025}(249) \approx 294, \quad \chi^2_{0{,}975}(249) \approx 207.$$

Следовательно, оценка σ^2 имеет вид $0{,}00034 < \sigma^2 < 0{,}000483$. Отсюда получим $0{,}018 < \sigma < 0{,}022$ – симметричный 0,95-доверительный интервал для волатильности акции.

Лекция 9
Приближенные доверительные интервалы

В первой части лекции для среднего значения и доли признака в генеральной совокупности строятся приближенные доверительные интервалы. Во второй части рассматривается (точный) интервал предсказания.

§9.1. Приближенная доверительная оценка генерального среднего и доли

В некоторых случаях при заданном объеме выборки n не удается точно найти вероятность, с которой интервал $(\underline{\theta}, \overline{\theta})$ накрывает параметр θ распределения, однако предел вероятности

$$\lim_{n\to\infty} P\{\theta \in (\underline{\theta}, \overline{\theta})\} = \gamma$$

существует и может быть найден по имеющимся данным. Поскольку при больших n выполняется приближенное соотношение

$$P\{\theta \in (\underline{\theta}, \overline{\theta})\} \approx \gamma,$$

интервал $(\underline{\theta}, \overline{\theta})$ называется *приближенным* (*асимптотическим*) *γ-доверительным интервалом*.

1. Приближенная интервальная оценка генерального среднего

Пусть $X_1, ..., X_n$ – выборка из некоторого распределения с генеральным средним μ и дисперсией σ^2. Предположим сначала, что дисперсия σ^2 известна. Тогда имеет смысл рассмотреть отношение

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}},$$

которое уже использовалось для оценки μ в случае нормального генерального распределения в качестве центральной статистики. Теперь T не является центральной статистикой, поскольку ее распределение, вообще говоря, зависит от объема выборки n. Тем не менее в силу центральной предельной теоремы функция распределения T стремится к функции распределения стандартного нормального закона. Следовательно, полученные на основе T оценки (7.9), (7.10) и (7.12) являются различными вариантами приближенных $(1-\alpha)$-доверительных оценок μ.

Предположим теперь, что генеральная дисперсия σ^2 существует, но неизвестна. Заменим в формулах (7.9), (7.10) и (7.12) σ

на выборочное стандартное отклонение $\hat{\sigma} = \sqrt{\hat{D}(X)}$. В результате получим:

оценку генерального среднего снизу

$$\overline{X} - Z_{\alpha} \frac{\hat{\sigma}}{\sqrt{n}} < \mu, \tag{9.1}$$

оценку генерального среднего сверху

$$\mu < \overline{X} + Z_{\alpha} \frac{\hat{\sigma}}{\sqrt{n}}, \tag{9.2}$$

двустороннюю оценку

$$\overline{X} - Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} < \mu < \overline{X} + Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}. \tag{9.3}$$

Необходимо отметить, что без дополнительных предположений о виде генерального распределения неравенства (9.1) – (9.3), даже при больших n, могут выполняться с вероятностью, значительно отличающейся от $1-\alpha$. Минимальное требование, обеспечивающее асимптотическую $(1-\alpha)$-доверительность оценок (9.1) – (9.3), вероятно, состоит в том, чтобы существовал генеральный момент ν_4. Действительно, из существования ν_4 вытекает, что $\hat{\sigma}$ сходится по вероятности к σ вследствие чего, неравенства (9.1) – (9.3) являются асимптотическими $(1-\alpha)$-доверительными оценками μ.

Пример 9.1. В некотором городе население составляет 1000000 человек. Для случайно отобранных 625 жителей средний возраст составил 33 года, при среднем квадратичном отклонении 15 лет. Требуется найти приближенный 0,95-доверительный интервал для среднего возраста жителей города.

Решение. Поскольку $Z_{\alpha/2} = Z_{0,025} = 1{,}96$ и $\sqrt{625} = 25$, полученная из (9.3) приближенная 0,95-доверительная оценка среднего возраста μ имеет вид двойного неравенства $33 - 1{,}96 \cdot \frac{15}{25} < \mu < 33 + 1{,}96 \cdot \frac{15}{25}$, или $31{,}8 < \mu < 34{,}2$.

2. Приближенная интервальная оценка генеральной доли признака

Предположим, что признак X в генеральной совокупности $\Omega = \{\omega\}$ распределен по закону Бернулли

X	0	1
P_0	q	p

, (9.4)

где P_0 – вероятностная мера (относительная частота) на Ω. Другими словами, признак X принимает только значения 0 и 1, а доля (мера) тех элементов $\omega \in \Omega$, для которых $X(\omega) = 1$, равна p.

Пусть $X_1, ..., X_n$ – выборка объема n из распределения (9.4),

X	0	1
Отн. частота	$\hat{q}$	$\hat{p}$

(9.5)

– соответствующее выборочное распределение признака X. Заметим, что $\hat{q}$ и $\hat{p}$ – случайные числа, причем $\hat{p}$ – это доля тех $X_1, ..., X_n$, которые принимают значение 1, т.е.

$$\hat{p} = \frac{X_1 + ... + X_n}{n} = \overline{X}. \tag{9.6}$$

Поскольку для распределения Бернулли (9.5) дисперсия $\hat{\sigma}^2 = \hat{p}\hat{q}$, с учетом (9.6) из оценок (9.1) – (9.3) получаем следующие приближенные $(1-\alpha)$-доверительные оценки для $\mu = p$:

$$p > \hat{p} - Z_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}}, \tag{9.7}$$

$$p < \hat{p} + Z_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}}, \tag{9.8}$$

$$\hat{p} - Z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}. \tag{9.9}$$

Пример 9.2. В условиях предыдущего примера из отобранных жителей оказалось 125 учащихся. Требуется найти приближенный 95%-й доверительный для доли p учащихся среди всех жителей города.

Решение. Имеем $Z_{\alpha/2} = Z_{0,025} = 1{,}96$, $\hat{p} = 0{,}2$ и $\hat{q} = 0{,}8$. Следовательно, из (9.9) получаем оценку доли p

$$0{,}2 - 1{,}96\sqrt{\frac{0{,}2 \cdot 0{,}8}{625}} < p < 0{,}2 + 1{,}96\sqrt{\frac{0{,}2 \cdot 0{,}8}{625}},$$

или

$$0{,}169 < p < 0{,}231.$$

§9.2. Интервал предсказания

Пусть $X_1, ..., X_{n+1}$ – выборка из нормального распределения $N(\mu, \sigma^2)$. Будем считать, что X_i – результат наблюдения X в испытании, проводимом в момент времени $i = 1, ..., n+1$. Тогда имеет смысл следующая задача: *построить по данным* $X_1, ..., X_n$ ***интервал предсказания*** (L, H), *накрывающий* X_{n+1} *с доверительной вероятностью* γ.

Используя стандартные статистики

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

и

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

построим статистику

$$T = \frac{X_{n+1} - \overline{X}}{s\sqrt{1 + \frac{1}{n}}}.$$

Можно доказать, что T распределена по закону Стьюдента $t(n-1)$. Используя этот факт, построим двусторонний интервал предсказания для X_{n+1}, т.е. такой интервал (L, H), что $L = l(X_1, ..., X_n)$, $H = h(X_1, ..., X_n)$ и

$$P\{X_{n+1} \in (L, H)\} = \gamma,$$

где γ – заданная заранее доверительная вероятность.

Так же как и в методе центральной статистики, сначала выберем положительные $\delta > 0$ и $\varepsilon > 0$ так, чтобы $\delta + \varepsilon = 1 - \gamma$. Поскольку $T \sim t(n-1)$, имеем

$$P\{-t_\delta(n-1) < T < t_\varepsilon(n-1)\} = 1 - (\delta + \varepsilon) = \gamma.$$

Событие $-t_\delta(n-1) < T < t_\varepsilon(n-1)$ равносильно событию

$$\overline{X} - t_\delta(n-1)s\sqrt{1+\frac{1}{n}} < X_{n+1} < \overline{X} + t_\varepsilon(n-1)s\sqrt{1+\frac{1}{n}}. \qquad (9.10)$$

Следовательно, вероятность события (9.10) также равна γ.

Положив в (9.10) $\delta = \varepsilon = \alpha/2$, где $\alpha = 1 - \gamma$, получим симметричный по вероятности интервал предсказания

$$\overline{X} - t_{\alpha/2}(n-1)s\sqrt{1+\frac{1}{n}} < X_{n+1} < \overline{X} + t_{\alpha/2}(n-1)s\sqrt{1+\frac{1}{n}}. \quad (9.11)$$

Если при постоянной сумме $\delta + \varepsilon = \alpha$ устремить поочередно ε и δ к 0, получим следующие односторонние интервалы предсказания:

$$X_{n+1} > \overline{X} - t_\alpha(n-1)s\sqrt{1+\frac{1}{n}}, \qquad (9.12)$$

$$X_{n+1} < \overline{X} + t_\alpha(n-1)s\sqrt{1+\frac{1}{n}}. \qquad (9.13)$$

Пример 9.3. Пусть A и B – цены некоторых активов, причем разность $A - B \sim N(\mu, \sigma^2)$, с неизвестными μ и σ^2. В результате 5 торгов разность $A - B$ приняла значения: 8, 3, –2, 7 и 6. Сколько нужно зарезервировать денег M, чтобы, продав на очередных торгах актив B, их хватило для покупки актива A с надежностью[1] 0,95?

Решение. Строго говоря, требуется построить односторонний 95%-й интервал предсказания для $A - B$ вида $(-\infty, M)$. Из (9.13) находим

$$M = \overline{X} + t_{0,05}(4)s\sqrt{1+\frac{1}{5}} \approx 4{,}4 + 2{,}132 \cdot 4{,}037 \cdot 1{,}095 \approx 13{,}8.$$

[1] По поводу термина *надежность* см. замечание на стр. 59

Лекция 10
Статистическая проверка гипотез

Пусть $X_1,...,X_n$ – случайная выборка объема n из некоторого генерального распределения. Не ограничивая общности можно считать, что существует определенная схема испытаний, при осуществлении которой вычисляется случайная величина X, а $X_1,...,X_n$ – это те ее значения, которые X принимает в результате серии n независимых испытаний. Таким образом, случайные величины $X_1,...,X_n$ независимы и распределены по тому же закону, что и X.

Одна из основных задач математической статистики состоит в том, чтобы по реализации $x_1,...,x_n$ случайной выборки $X_1,...,X_n$ проверить определенную гипотезу о виде или параметрах генерального распределения. Существуют также задачи, в которых наряду с выборкой $X_1,...,X_n$ из некоторого распределения имеется выборка $Y_1,...,Y_m$ из другого распределения, и при этом требуется проверить определенные соотношения между данными распределениями. В частности, речь может идти о проверке гипотез, связанных с такими числовыми характеристиками распределений, как $E(X), E(Y), D(X)$ и $D(Y)$. Настоящая лекция и несколько следующих посвящены рассмотрению различных вопросов, связанных с проверкой подобного рода гипотез.

§9.3. Виды статистических гипотез и общая схема статистического критерия

Определение. ***Статистической гипотезой*** *называется любое утверждение о виде или параметрах генерального распределения. Статистическая гипотеза называется* ***параметрической***, *если она основана на предположении, что генеральное распределение известно с точностью до конечного числа параметров.*

Приведем несколько примеров параметрических гипотез:

1) $X \sim N(0,1)$ в предположении, что $X \sim N(\mu,\sigma^2)$.

2) $E(X) > 0$ в предположении, что $X \sim N(\mu,1)$.

3) $D(X) = D(Y)$ в предположении, что $X \sim N(\mu_x,\sigma_x^2), Y \sim N(\mu_y,\sigma_y^2)$.

Примеры непараметрических гипотез:

1) $X \sim N(0,1)$ в предположении, что $D(X)$ существует.

2) $E(X) > 0$ в предположении, что $E(X)$ существует.

3) $F_X(x) \equiv F_Y(y)$ в предположении, что функции распределения $F_X(x)$ и $F_Y(y)$ непрерывны.

Рассмотрим базисное предположение, состоящее в том, что генеральное распределение зависит от некоторого параметра $\theta \in \mathbf{R}^n$. В этом случае, очевидно, любая гипотеза о значениях θ будет параметрической.

Определение. *Параметрическая гипотеза называется* ***простой***, *если она имеет вид*: $\theta = \theta_0$, *где* θ_0, *– некоторое фиксированное значение параметра* θ. *Гипотеза вида*: $\theta \in \Theta$, *где* Θ *– какое-либо множество, содержащее, по меньшей мере, два различных элемента, называется* ***сложной***.

Пусть H_0 и H_1 – две взаимоисключающие статистические гипотезы. Гипотезу H_0 назовем *основной*, а гипотезу H_1 – *альтернативной*. Далее везде в качестве базисного предположения принимается утверждение о справедливости одной из этих гипотез.

Пример 9.4. Пусть $N(\mu,\sigma^2)$ – генеральное распределение со средним μ и дисперсией σ^2. Для основной гипотезы

$H_0: \ \mu = 0, \ \sigma^2 = 1;$

рассмотрим две альтернативные гипотезы:

1) $H_1: \ \mu = 0, \ \sigma^2 = 2;$

2) $H_1: \ \mu \neq 0, \ \sigma^2 = 1.$

Какая из этих гипотез является простой (сложной)? В чем состоит базисное предположение?

Решение. 1) В данном примере $\theta = (\mu,\sigma^2)$ – двумерный вектор. Для основной и для альтернативной гипотез компоненты

этого вектора определены однозначно, следовательно, H_0 и H_1 – простые гипотезы. Базисное предположение состоит в том, что генеральным распределением является $N(0,\sigma^2)$ с неизвестной дисперсией $\sigma^2 \in \{1; 2\}$.

2) H_1 – сложная гипотеза, поскольку μ может принимать различные значения. Базисное предположение состоит в том, что генеральным распределением является $N(\mu,1)$ с неизвестным средним μ.

Пусть K – некоторое подмножество в $\mathbf{R}^n$.

Определение. ***Статистическим критерием с критической областью*** K *называется правило, в соответствии с которым* H_0 *отвергается, если выборка* $(x_1,...,x_n) \in K$, *и принимается, если* $(x_1,...,x_n) \notin K$.

Поскольку H_0 и H_1 – взаимоисключающие гипотезы, принятие H_0 означает отклонение H_1. Напротив, отклонение H_0, вследствие базисного предположения, автоматически приводит к принятию H_1.

Как правило, критическая область задается при помощи неравенства:

$$K = \{(x_1,...,x_n) \in \mathbf{R}^n:\ t(x_1,...,x_n) > c\} \tag{9.14}$$

или

$$K = \{(x_1,...,x_n) \in \mathbf{R}^n:\ t(x_1,...,x_n) < c\}, \tag{9.15}$$

где $t(x_1,...,x_n)$ – подходящая функция от выборочных значений, а $c = const$ – некоторая константа. В дальнейшем критическую область вида (9.14) или (9.15) будем, для краткости, записывать в виде $t > c$ или $t < c$, соответственно. Нам также встретятся критические области вида

$$K = \{t < c_1\} \cup \{t > c_2\},$$

где c_1 и c_2 – некоторые константы, такие что $c_1 < c_2$. Во всех этих случаях числа c, c_1 и c_2 называются *критическими значениями*, а функция $t(x_1,...,x_n)$ – *статистикой критерия*.

Применение статистического критерия может привести к ошибкам двух различных типов:

Ошибка первого рода состоит в том, что отвергается верная гипотеза H_0.

Ошибка второго рода состоит в том, что отвергается верная гипотеза H_1.

Определение. *Вероятность ошибки первого рода называется* ***уровнем значимости*** *критерия и обозначается* α. *Вероятность ошибки второго рода обозначается* β, *а величина* $1-\beta$ *называется* ***мощностью*** *критерия*.

Нетрудно видеть, что для статистического критерия с критической областью $t > c$ или $t < c$ вероятности ошибок и мощность критерия находятся в соответствии с таблицей:

Таблица 9.1.

$K = \{t > c\}$	$K = \{t < c\}$
$\alpha = P_{H_0}\left(t(X_1,...,X_n) > c\right)$	$\alpha = P_{H_0}\left(t(X_1,...,X_n) < c\right)$
$\beta = P_{H_1}\left(t(X_1,...,X_n) \le c\right)$	$\beta = P_{H_1}\left(t(X_1,...,X_n) \ge c\right)$
$1-\beta = P_{H_1}\left(t(X_1,...,X_n) > c\right)$	$1-\beta = P_{H_1}\left(t(X_1,...,X_n) < c\right)$

В этой таблице $P_{H_i}(...)$ – вероятность события, вычисленная в предположении справедливости гипотезы $H_i, i = 1,2$. Заметим, что в данном подходе гипотезы H_0 и H_1 не рассматриваются как случайные события, поэтому символы P_{H_0} и P_{H_1} *не являются* обозначениями условной вероятности.

§9.4. Лемма Неймана–Пирсона

Предположим, что генеральное распределение имеет зависящую от параметра θ положительную при всех x плотность $f(x;\theta) > 0$. Пусть H_0 и H_1 – простые гипотезы вида $H_0: \theta = \theta_0$ и $H_1: \theta = \theta_1$.

Запишем функции правдоподобия, соответствующие этим гипотезам:

$$L_0(x_0,...,x_n) = f(x_1;\theta_0)\cdot...\cdot f(x_n;\theta_0),$$
$$L_1(x_0,...,x_n) = f(x_1;\theta_1)\cdot...\cdot f(x_n;\theta_1).$$

Теорема 9.1 (лемма Неймана–Пирсона). *Для любого* $\alpha \in (0,1)$ *существует такая константа* c_α, *что критерий с критической областью*

$$\frac{L_1(x_0,...,x_n)}{L_0(x_0,...,x_n)} > c_\alpha$$

является наиболее мощным критерием среди всех статистических критериев с какой-либо критической областью K, *предназначенных для проверки* H_0 *против* H_1 *с уровнем значимости* α.

Пример 9.5. Пусть $X_1,...,X_n$ – выборка из нормального распределения $N(\mu,\sigma^2)$ с известной дисперсией σ^2. Для проверки гипотезы $H_0: \mu = \mu_0$ против гипотезы $H_1: \mu = \mu_1$ определим статистику

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}. \qquad (9.16)$$

Требуется доказать, что в случае $\mu_1 > \mu_0$ критерий с критической областью $Z > Z_\alpha$ является наиболее мощным среди всех критериев по проверке H_0 против H_1 с уровнем значимости α. Аналогично, в случае $\mu_1 < \mu_0$ необходимо доказать, что наиболее мощным является критерий с критической областью $Z < -Z_\alpha$.

Решение. Записываем функции правдоподобия:

$$L_i(x_1,...,x_n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-\mu_i)^2}{2\sigma^2}} \cdot ... \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_n-\mu_i)^2}{2\sigma^2}} =$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu_i)^2 \right\}, \; i = 1,2.$$

Затем находим их отношение

$$\frac{L_1}{L_0} = \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} [(x_j - \mu_1)^2 - (x_j - \mu_0)^2] \right\} =$$

$$= \exp\left\{ \frac{\mu_1 - \mu_0}{2\sigma^2} \sum_{j=1}^{n} (2x_j - \mu_1 - \mu_0) \right\} =$$

$$= \exp\left\{ \frac{\mu_1 - \mu_0}{\sigma^2} \sum_{j=1}^{n} \left(n\bar{x} - \frac{n}{2}(\mu_0 + \mu_1) \right) \right\}.$$

При $\mu_1 > \mu_0$ критическая область $L_1 / L_0 > c_\alpha$ может быть задана также неравенством $\bar{x} > b_a$, где константа b_a, соответствующим образом выражена через константу c_α. В свою очередь неравенство $\bar{x} > b_a$, очевидно, эквивалентно неравенству

$$\frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > \frac{b_a - \mu_0}{\sigma / \sqrt{n}}. \qquad (9.17)$$

Поскольку при верной H_0 статистика (9.16) распределена по стандартному нормальному закону $N(0,1)$, а (9.17) задает критическую область наиболее мощного критерия с уровнем значимости α, правая часть неравенства (9.17) совпадает с процентной точкой Z_α распределения $N(0,1)$.

Аналогичным образом доказывается, что для $\mu_1 < \mu_0$ критическая область $L_1 / L_0 > c_\alpha$ может быть представлена как $Z < -Z_\alpha$.

Лекция 11
Проверка гипотезы об определенном значении параметра

§11.1. Проверка гипотезы об определенном значении параметра нормального распределения

Пусть $X_1,...,X_n$ – выборка из распределения $N(\mu,\sigma^2)$. Гипотеза об определенном значении генерального среднего состоит в том, что эта выборка получена при фиксированном значении $\mu=\mu_0$. Структура критерия по проверке данной гипотезы зависит от того, известна или нет генеральная дисперсия σ^2, а также от вида альтернативной гипотезы. Аналогично гипотеза об определенном значении дисперсии σ^2 состоит в том, что выборка $X_1,...,X_n$ получена при фиксированном значении $\sigma^2=\sigma_0^2$, а способ ее проверки зависит от того, известна или нет генеральное среднее μ и, конечно, от вида альтернативной гипотезы. Отметим, что во всех случаях прослеживается явное сходство с построением доверительных интервалов для соответствующих параметров нормального распределения.

1. Проверка гипотезы об определенном значении генерального среднего при известной дисперсии

Для проверки гипотезы H_0: $\mu=\mu_0$ против любой из трех альтернативных гипотез H_1: 1) $\mu>\mu_0$, 2) $\mu<\mu_0$ или 3) $\mu\neq\mu_0$ используется та же статистика, что и в примере 9.5

$$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}. \qquad (11.1)$$

Критическая область K выбирается в соответствии с таблицей

H_1	K
$\mu > \mu_0$	$Z > Z_\alpha$
$\mu < \mu_0$	$Z < -Z_\alpha$
$\mu \neq \mu_0$	$\lvert Z \rvert > Z_{\alpha/2}$

Как было показано в примере 9.5, в случае односторонней гипотезы ($\mu > \mu_0$ или $\mu < \mu_0$) соответствующий критерий является наиболее мощным вследствие леммы Неймана–Пирсона. Вместе с тем, можно показать, что применяемый в случае двусторонней альтернативы $\mu \neq \mu_0$ статистический критерий не является наиболее мощным ни при каком $\mu \neq \mu_0$. Тем не менее этот критерий также имеет уровень значимости α. Действительно, вероятность ошибки первого рода равна

$$P_{H_0}(|Z| > Z_{\alpha/2}) = P_{H_0}(Z > Z_{\alpha/2}) + P_{H_0}(Z < -Z_{\alpha/2}) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

2. Проверка гипотезы об определенном значении генерального среднего при неизвестной дисперсии

Если генеральная дисперсия σ^2 неизвестна, в качестве статистики критерия используется статистика

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}, \tag{11.2}$$

которая отличается от (11.1) заменой σ на s, где

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

В соответствии с видом альтернативы H_1 критическая область K выбирается по таблице

H_1	K
$\mu > \mu_0$	$t > t_\alpha(n-1)$
$\mu < \mu_0$	$t > -t_\alpha(n-1)$
$\mu \neq \mu_0$	$\lvert t \rvert > t_{\alpha/2}(n-1)$

в которой $t_\alpha(n-1) - 100\alpha$ -процентная точка распределения Стьюдента с $n-1$ степенями свободы.

Пример 11.1. Цена акции S_t в момент времени $t = 0,1,...,T$ задается моделью случайного блуждания:

$$\ln S_t = \ln S_0 + \sum_{i=1}^{t} X_i,$$

где

$$X_i = \ln \frac{S_i}{S_{i-1}}$$

– логарифмическая доходность акции за период $[i-1,i]$, а $X_1,...,X_T$ в совокупности образуют выборку из $N(\mu,\sigma^2)$. Предположим, что за период времени $T = 10$ средняя логарифмическая доходность составила $\bar{x} = 0,01$ при дисперсии $s_x^2 = 0,02^2$. Требуется при 5%-ном уровне значимости проверить гипотезу $H_0: \mu = 0$ против гипотезы $H_1: \mu > 0$.

Решение. Находим значение статистики

$$t = \frac{0,01-0}{0,03/3} = 1,5.$$

Затем по таблице процентных точек распределения Стьюдента находим критическое значение $t_{0,05}(9) = 1,833$. Поскольку $t < t_{0,05}(9)$, наблюдаемое значение статистики не принадлежит критической области. Следовательно, гипотеза $\mu = 0$ принимается.

3. Проверка гипотезы об определенном значении генеральной дисперсии

Способ проверки гипотезы $H_0: \sigma^2 = \sigma_0^2$ зависит от того, известно или нет генеральное среднее μ. Если μ известно, то используется статистика

$$\chi^2 = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - \mu)^2 = \frac{ns_0^2}{\sigma_0^2}, \qquad (11.3)$$

а критическая область K выбирается по таблице

H_1	K
$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_\alpha(m)$
$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha}(m)$
$\sigma^2 \neq \sigma_0^2$	$\{\chi^2 < \chi^2_{1-\alpha/2}(m)\} \cup \{\chi^2 > \chi^2_{\alpha/2}(m)\}$

, (11.4)

в которой $m = n$ – объем выборки, $\chi^2_\alpha(m)$ – 100α -процентная точка распределения χ^2 с m степенями свободы.

Если же генеральное среднее не известно, используется статистика

$$\chi^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(n-1)s^2}{\sigma_0^2}. \qquad (11.5)$$

При этом критическая область определяется все той же таблицей (11.4), но с другим числом степеней свободы $m = n - 1$.

Пример 11.2. Используя данные из примера 11.1, проверить при уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : \sigma = 0{,}01$ против гипотезы $H_1 : \sigma > 0{,}01$.

Решение. Вычисляем наблюдаемое значение статистики критерия:

$$\chi^2 = \frac{9 \cdot 0{,}02^2}{0{,}012} = 36.$$

Затем по таблице процентных точек распределения χ^2 находим критическое значение $\chi^2_{0{,}05}(9) = 16{,}92$. Поскольку значение статистики χ^2 принадлежит критической области, $\chi^2 > \chi^2_{0{,}05}(9)$, основная гипотеза отвергается в пользу гипотезы $\sigma > 0{,}01$.

§11.2. *P*-значение критерия

Описанная выше процедура проверки статистической гипотезы позволяет принять или отклонить основную гипотезу путем сравнения наблюдаемого значения статистики критерия с критическим значением, связанным с данным уровнем значимости, однако,

если уровень значимости будет другим, то придется вновь вычислить соответствующее критическое значение. Вводимое ниже понятие, ставшее популярным в связи с широким распространением статистических программ, позволяет решить вопрос о принятии или отклонении основной гипотезы одновременно для всех уровней значимости без вычисления критических значений.

Определение. *Для фиксированной реализации* $\vec{x}$ *случайной выборки* $\vec{X} = (X_1, ..., X_n)$ ***P-значением*** (***P-value***) *статистического критерия называется такое число* $PV(\vec{x})$, *что* $PV(\vec{x}) \geq \alpha$ *для любого уровня значимости* α, *при котором гипотеза* H_0 *принимается, и* $PV(\vec{x}) \leq \alpha$, *для любого уровня значимости* α, *при котором гипотеза* H_0 *отвергается*.

Предположим, что *P*-значение $PV(\vec{x})$ уже каким-либо способом найдено. Тогда решение о принятии (отклонении) H_0 для заданного α осуществляется на основе следующего простого правила: *если* $PV(\vec{x}) < \alpha$, *гипотеза* H_0 *отвергается, а если* $PV(\vec{x}) > \alpha$ *гипотеза* H_0 *принимается*.

Рассмотрим отдельно случай $PV(\vec{x}) = \alpha$. Как правило, критическую область можно представить в виде

$$K_\alpha = \{t(\vec{x}) > c(\alpha)\}, \tag{11.6}$$

где $c(\alpha)$ – непрерывная убывающая функция. Как нетрудно видеть, в этом случае $PV(\vec{x}) = c^{-1}(t(\vec{x}))$ и для $\alpha = PV(\vec{x})$ имеет место равенство

$$t(\vec{x}) = c(\alpha), \tag{11.7}$$

означающее, что H_0 принимается. Отсюда уже легко получить широко применяемую формулу:

$$PV(\vec{x}) = P_{H_0}(t(\vec{X}) > t(\vec{x})). \tag{11.8}$$

Действительно, при любом уровне значимости α из (11.6) имеем

$$P_{H_0}(t(\vec{X}) > c(\alpha)) = \alpha, \tag{11.9}$$

но с учетом (11.7) из (11.9) вытекает (11.8).

Совершенно аналогично доказывается, что в случае

$$K_\alpha = \{t(\vec{x}) < c(\alpha)\},$$

где $c(\alpha)$ – непрерывная возрастающая функция, *P*-значение удовлетворяет соотношению $PV(\vec{x}) = P_{H_0}(t(\vec{X}) < t(\vec{x}))$.

Пример 11.3. Пусть $\vec{x}$ – реализация случайной выборки $X_1,...,X_{100}$ из распределения $N(\mu,5^2)$, такая, что $\bar{x} = 3$. Требуется при уровне значимости $\alpha = 0{,}05$ проверить гипотезу $H_0: \mu = 0$, против альтернативы $H_1: \mu > 0$. Необходимо также вычислить *P*-значение критерия.

Решение. Для проверки гипотезы используем критическую область вида $K = \{Z > Z_\alpha\}$, где $Z = \bar{x}/5$, $Z_{0,05} = \Phi^{-1}(0{,}45) = 1{,}645$. Поскольку наблюдаемое значение статистики $Z_{\text{набл}} = 1{,}8 > 1{,}645$, гипотеза H_0 отвергается.

Прежде чем найти *P*-значение, заметим, что в случае Z-критерия формулу (11.8) можно представить в виде

$$P\text{-}value = P(Z > Z_{\text{набл}}) = \frac{1}{2} - \Phi(Z_{\text{набл}}). \qquad (11.10)$$

Под Z-критерием здесь понимается любой критерий, статистика которого $Z \sim N(0,1)$ при верной H_0. С учетом (11.10) находим

$$P\text{-}value = 0{,}5 - \Phi(1{,}8) = 0{,}0359.$$

Следовательно, при любом уровне значимости $\alpha > 0{,}0359$ нулевая гипотеза отвергается, что вполне согласуется приведенной раньше проверкой при заданном $\alpha = 0{,}05$.

На этом примере видно, что проверка гипотезы при помощи *P*-значения более информативна, нежели традиционная проверка с помощью критического значения. Тем не менее, выбор того или иного способа проверки, конечно, зависит от наличия соответствующих таблиц (или компьютерных программ).

Может сложиться впечатление, что *P*-значение является своеобразной характеристикой адекватности H_0, однако, это не совсем так. Нетрудно доказать, что при верной основной гипотезе *P*-

значение равномерно распределено на отрезке [0,1]. Поэтому вероятность получить малое *P*-значение, скажем, меньше α равно вероятности получить соответствующее большое значение (больше $1-\alpha$). Тем не менее, если H_0 не верна, наблюдаемые *P*-значения (при достаточно высокой мощности критерия) концентрируются около нуля.

Лекция 12
Сравнение параметров двух нормальных распределений

Пусть $\vec{X} = (X_1,...,X_m)$ – выборка из $N(\mu_x, \sigma_x^2)$, а $\vec{Y} = (Y_1,...,Y_n)$ – выборка из нормального распределения $N(\mu_y, \sigma_y^2)$. Далее считаем выборки $\vec{X}$ и $\vec{Y}$ *независимыми*, что означает независимость в совокупности $m+n$ случайных величин $X_1,...,X_m,Y_1,...,Y_n$.

§12.1. Сравнение генеральных средних

Способ проверки гипотез о соотношениях между генеральными средними совокупностей (распределений) $N(\mu_x, \sigma_x^2)$ и $N(\mu_y, \sigma_y^2)$ определяется тем, известны или нет дисперсии σ_x^2 и σ_y^2.

1. Сравнение генеральных средних при известной дисперсии

Предположим, что дисперсии σ_x^2 и σ_y^2 известны, генеральные средние μ_x и μ_y неизвестны. Рассматривается основная гипотеза $H_0: \mu_x = \mu_y$. Альтернативная гипотеза чаще всего имеет вид

1) $H_1: \mu_x > \mu_y$;

2) $H_1: \mu_x < \mu_y$;

3) $H_1: \mu_x \neq \mu_y$.

При проверке H_0 против H_1 вида 1), 2) или 3) используется одна и та же статистика

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}}. \qquad (12.1)$$

Утверждение 1. *Если верна* H_0, *то* $Z \sim N(0,1)$.

Доказательство. Статистика Z является линейной функцией от нормально распределенных, независимых случайных величин $X_1,...,X_m,Y_1,...,Y_n$. Следовательно, $Z \sim N(\mu_z, \sigma_z^2)$. Нетрудно видеть, что $\mu_z = C^{-1}\left(E(\overline{X}) - E(\overline{Y})\right)$, где $C = \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}$. При верной H_0 имеем

$$E(\overline{X}) - E(\overline{Y}) = \mu_x - \mu_y = 0,$$

Следовательно, $\mu_z = 0$. Далее находим

$$D(Z) = C^{-2}\left(\frac{1}{m}\sigma_x^2 + \frac{1}{n}\sigma_y^2\right) = 1,$$

что и завершает доказательство.

Пусть $Z_\alpha, Z_{\alpha/2}$ – соответствующие процентные точки стандартного нормального распределения $N(0,1)$. При проверке H_0 против H_1 применяется критерий с критической областью, определяемой по таблице

H_1	K
1) $\mu_x > \mu_y$	$Z > Z_\alpha$
2) $\mu_x < \mu_y$	$Z < -Z_\alpha$
3) $\mu_x \neq \mu_y$	$\lvert Z \rvert > Z_{\alpha/2}$

Утверждение 2. *Критерий с критической областью* 1) – 3) *имеет уровень значимости* α.

Доказательство. Утверждение для критической области 1) следует из утверждения 1 и определения процентной точки.

Если верна H_0, то $Z \sim -Z$ и

$$P_{H_0}(Z < -Z_\alpha) = P_{H_0}(-Z > Z_\alpha) = P_{H_0}(Z > Z_\alpha) = \alpha,$$

что доказывает утверждение для критической области 2).

Рассмотрим теперь случай критической области 3). Событие $|Z| > Z_{\alpha/2}$ является суммой несовместных событий $Z > Z_{\alpha/2}$ и $Z < -Z_{\alpha/2}$. Отсюда следует, что

$$P_{H_0}(|Z| > Z_{\alpha/2}) = P_{H_0}(Z > Z_{\alpha/2}) + P_{H_0}(Z < -Z_{\alpha/2}) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

Можно доказать, что во всех трех случаях при фиксированных μ_x и μ_y

$$\lim_{\substack{m\to\infty \\ n\to\infty}} P_{H_1}(\text{отвергается } H_0) = 1.$$

Таким образом, данные критерии имеют высокую мощность, по крайней мере, для выборок достаточно большого объема.

2. Сравнение генеральных средних при неизвестной дисперсии

Теперь предположим, что обе генеральные дисперсии неизвестны, но одинаковы, $\sigma_x^2 = \sigma_y^2 = \sigma^2$. С учетом этого равенства имеем

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} = \frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}. \qquad (12.2)$$

Поскольку значение σ неизвестно, попробуем заменить σ на корень из несмещенной оценки дисперсии. В данном случае имеется, по крайней мере, две такие оценки:

$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2$$

и

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Оказывается, что вместо того, чтобы использовать одну из них, лучше воспользоваться линейной комбинацией

$$s^2 = \frac{m-1}{m+n-2}s_x^2 + \frac{n-1}{m+n-2}s_y^2. \qquad (12.3)$$

Очевидно проверяется, что s^2 – несмещенная оценка σ^2. Кроме того, можно доказать, что

$$\frac{(m-1)s_x^2}{\sigma^2} = \frac{(X_1 - \overline{X})^2 + ... + (X_m - \overline{X})^2}{\sigma^2} \sim \chi^2(m-1) \qquad (12.4)$$

и

$$\frac{(n-1)s_y^2}{\sigma^2} = \frac{(Y_1 - \overline{Y})^2 + ... + (Y_n - \overline{Y})^2}{\sigma^2} \sim \chi^2(n-1). \qquad (12.5)$$

Отсюда с учетом независимости $\vec{X}$ и $\vec{Y}$ следует, что

$$\frac{(m-1)s_x^2 + (n-1)s_y^2}{\sigma^2} \sim \chi^2(m+n-2). \qquad (12.6)$$

Таким образом, коэффициенты при s_x^2 и s_y^2 в (12.3) можно интерпретировать как доли числа степеней свободы “иксов” и “игреков” в числе степеней свободы распределения случайной величины (12.6), которая отличается от оценки s^2 лишь постоянным множителем $\frac{m+n-2}{\sigma^2}$.

Заменив σ на s в записи статистики (12.2), получим новую статистику

$$T = \frac{\overline{X} - \overline{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Примем без доказательства следующую теорему.

Теорема 12.1. *Если верна* H_0, *то* $T \sim t(m+n-2)$, *где* $t(m+n-2)$ – *распределение Стьюдента с* $m+n-2$ *степенями свободы*.

В силу данной теоремы статистику T можно использовать для построения критерия по проверке H_0. Как и раньше рассматриваются три основных случая альтернативной гипотезы H_1, в соответствии с которыми выбирается критическая область K того или иного вида:

H_1	K
1) $\mu_x > \mu_y$	$T > t_\alpha(m+n-2)$
2) $\mu_x < \mu_y$	$T < -t_\alpha(m+n-2)$
3) $\mu_x \neq \mu_y$	$\lvert T \rvert > t_{\alpha/2}(m+n-2)$

(12.7)

Утверждение 3. *Критерий с критической областью* (12.7) *в любом из случаев* 1) – 3) *имеет уровень значимости* α.

Доказательство. 1) По теореме 12.1

$$P_{H_0}(T > t_\alpha(m+n-2)) = \alpha.$$

Таким образом, α – это действительно уровень значимости критерия 1).

2) Если в теореме 12.1 поменять местами $\vec{X}$ и $\vec{Y}$, то получим, что $-T \sim t(m+n-2)$. Отсюда имеем

$$P_{H_0}(T < -t_\alpha(m+n-2)) = P_{H_0}(-T > t_\alpha(m+n-2)) = \alpha.$$

3) Находим уровень значимости

$$P_{H_0}(|T| > t_{\alpha/2}(m+n-2)) = 2P_{H_0}(T > t_{\alpha/2}(m+n-2)) = \alpha.$$

§12.2. Сравнение дисперсий двух нормальных распределений

Пусть как и в предыдущем параграфе имеется две независимые выборки из нормальных распределений:

$$X_1,...,X_m \sim N(\mu_x, \sigma_x^2),$$
$$Y_1,...,Y_n \sim N(\mu_y, \sigma_y^2).$$

Далее считаем, что все четыре параметра μ_x, μ_y, σ_x^2 и σ_y^2 неизвестны. В качестве основной гипотезы примем $H_0 : \sigma_x^2 = \sigma_y^2$, а в качестве альтернативной – одну из трех гипотез:

1) $H_1 : \sigma_x^2 > \sigma_y^2$;

2) $H_1 : \sigma_x^2 < \sigma_y^2$;

3) $H_1 : \sigma_x^2 \neq \sigma_y^2$.

При построении критериев по проверке H_0 с заданным уровнем значимости α применяется следующая теорема.

Теорема 12.2. *Если верна* H_0, *то*

$$\frac{s_x^2}{s_y^2} \sim F(m-1, n-1),$$

где $s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2$, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, *а* $F(m-1,n-1)$ – *распределение Фишера с* $m-1$ *и* $n-1$ *степенями свободы*.

Доказательство. Определим величины:

$$\chi_x^2 = \frac{m-1}{\sigma_x^2} s_x^2, \quad \chi_y^2 = \frac{n-1}{\sigma_y^2} s_y^2.$$

Как уже ранее отмечалось (формулы (12.4) и (12.5)),

$$\chi_x^2 \sim \chi^2(m-1), \quad \chi_y^2 \sim \chi^2(n-1).$$

Поэтому при верной H_0 в силу независимости выборок $\vec{X}$ и $\vec{Y}$ имеем

$$\frac{s_x^2}{s_y^2} = \frac{(m-1)s_x^2/(m-1)}{\sigma_x^2} \Big/ \frac{(n-1)s_y^2/(n-1)}{\sigma_y^2} =$$

$$= \frac{\chi_x^2/(m-1)}{\chi_y^2/(n-1)} \sim F(m-1, n-1).$$

В соответствии с видом H_1 критическая область определяется следующими неравенствами:

H_1	Критическая область
1) $\sigma_x^2 > \sigma_y^2$	$\frac{s_x^2}{s_y^2} > F_\alpha(m-1, n-1)$
2) $\sigma_x^2 < \sigma_y^2$	$\frac{s_y^2}{s_x^2} > F_\alpha(n-1, m-1)$
3) $\sigma_x^2 \neq \sigma_y^2$	$\frac{s_1^2}{s_2^2} > F_{\alpha/2}(k_1, k_2)$

(12.8)

где символы s_1^2, s_2^2, k_1 и k_2 в зависимости от соотношения между s_x^2 и s_y^2 определяются таблицей

символ	$s_x^2 \geq s_y^2$	$s_x^2 < s_y^2$
s_1^2	s_x^2	s_y^2
s_2^2	s_y^2	s_x^2
k_1	$m-1$	$n-1$
k_2	$n-1$	$m-1$

Из теоремы 12.2 следует, что критерий с критической областью 1) или 2) имеет уровень значимости α. Критерий с критической областью 3) также имеет уровень значимости α, если только выполняется условие

$$F_{\alpha/2}(m-1, n-1) \cdot F_{\alpha/2}(n-1, m-1) > 1. \qquad (12.9)$$

Действительно, если верно (12.9), события

$$\frac{s_x^2}{s_y^2} > F_\alpha(m-1, n-1) \text{ и } \frac{s_y^2}{s_x^2} > F_\alpha(n-1, m-1)$$

несовместны. Следовательно,

$$P_{H_0}\left(\frac{s_1^2}{s_2^2} > F_{\alpha/2}(k_1,k_2)\right) = P_{H_0}\left(\frac{s_x^2}{s_y^2} > F_{\alpha/2}(m-1,n-1)\right) +$$

$$+ P_{H_0}\left(\frac{s_y^2}{s_x^2} > F_{\alpha/2}(n-1,m-1)\right) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

Заметим, что условие (12.9) нарушается только для достаточно больших α, которые на практике обычно не применяются.

Пример 12.1. После выхода негативной информации курс акций компании ABC упал на 10 %. До этого по результатам наблюдений за 30 периодов от закрытия до закрытия выборочное стандартное отклонение логарифмической доходности составило $s_x = 1{,}6\%$. В последующие после падения 25 периодов выборочное стандартное отклонение увеличилось до $s_y = 2{,}4\%$. Необходимо при 5-процентном уровне значимости проверить гипотезу о том, что генеральная дисперсия σ^2 не увеличилась. Предполагается, что дневная логдоходность (т.е. логарифмическая доходность за период от закрытия до следующего ближайшего закрытия торгов) распределена до выхода негативной информации по нормальному закону с параметрами (μ_x, σ_x^2), а после выхода – с параметрами (μ_y, σ_y^2).

Решение. Из условия ясно, что основная и альтернативная гипотезы имеют вид $H_0 : \sigma_x^2 = \sigma_y^2$ и $H_1 : \sigma_x^2 < \sigma_y^2$. Находим процентную точку распределения Фишера: $F_{0,05}(24,29) = 1{,}9$. Согласно (12.8) критическая область задается неравенством $s_y^2 / s_x^2 > 1{,}9$. Поскольку наблюдаемое значение статистики критерия

$$F_{\text{набл}} = \frac{s_y^2}{s_x^2} = \frac{0{,}024^2}{0{,}016^2} = 2{,}25$$

попадает в критическую область, гипотеза H_0 отклоняется в пользу H_1.

Лекция 13
Критерий согласия хи-квадрат

Среди множества различных статистических критериев имеется ряд критериев, называемых *критериями согласия*. Критерий согласия предназначен для проверки согласованности основной гипотезы H_0 с выборочными данными, однако, в отличие от рассмотренных ранее критериев, альтернативная гипотеза явным образом не выдвигается. При этом общая схема проверки гипотезы H_0, практически, остается неизменной. Так, если выборка $X_1,...X_n$ попадает в критическую область, гипотеза H_0 отвергается. Если же выборка оказывается вне критической области, говорят, что *нет оснований отклонить основную гипотезу*. Фактически, конечно, это означает, что гипотеза H_0 принимается, поскольку эмпирические данные согласуются с H_0.

1. Критерий хи-квадрат Пирсона

Производится серия повторных независимых испытаний, n – число испытаний, ω_t – элементарный исход испытания с номером $t = 1,...,n$. Поскольку испытания повторные, в качестве их общей вероятностной модели принимается одно и то же вероятностное пространство $(\Omega, \mathcal{F}, P)$, очевидно, что все элементарные исходы $\omega_t \in \Omega$.

Предположим, что $A_1,...,A_l \in \mathcal{F}$ – попарно несовместные события, такие что $A_1 + ... + A_l = \Omega$. В качестве H_0 примем гипотезу, состоящую в том, что вероятности событий A_i $(i = 1,...,l)$ заданы таблицей

событие	A_1	...	A_l
вероятность	p_1	...	p_l

. (13.1)

Пусть n_i – эмпирическая частота события A_i, т.е. число испытаний, в которых A_i наступило. Эквивалентно: n_i – число тех эле-

ментарных исходов ω_t, для которых $\omega_t \in A_i$. Исходными данными для критерия χ^2 Пирсона является таблица эмпирических частот

событие	A_1	...	A_l
частота	n_1	...	n_l

. (13.2)

Если основная гипотеза верна, согласно статистическому определению вероятности $\hat{p}_i \approx p_i$, где $\hat{p}_i = n_i / n$ – относительная частота события A_i. В качестве меры одновременной близости l пар чисел $(\hat{p}_i, p_i)$ можно принять любую сумму вида $c_1(\hat{p}_1 - p_1)^2 + ... + c_l(\hat{p}_l - p_l)^2$, в которой $c_i > 0$ – какие-либо положительные числа. К.Пирсон обнаружил, что если придать большие веса маловероятным событиям, положив $c_i = n / p_i$, то при неограниченном увеличении n распределение статистики

$$\chi^2 = \sum_{i=1}^{l} \frac{n}{p_i}(\hat{p}_i - p_i)^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i} \qquad (13.3)$$

перестает зависеть от конкретных значений вероятностей p_i и стремится к распределению хи-квадрат с $l-1$ степенями свободы.

Заметим, что при верной H_0, случайные величины n_i распределены по биномиальному закону с параметрами n и p_i, вследствие чего $np_i = E(n_i)$ называется *ожидаемой* (*теоретической*) *частотой* события A_i.

Опуская подробности, найдем асимптотическое (при $n \to \infty$) распределение статистики χ^2. Рассмотрим с этой целью l-мерный вектор

$$\vec{u}_0 = \left(\sqrt{p_1}, ..., \sqrt{p_l}\right)$$

Вектор $\vec{u}_0$ имеет единичную длину, вследствие чего может быть дополнен до ортонормированного базиса $\vec{u}_0, ..., \vec{u}_{l-1}$ в R^l. Определим l величин Z_k следующими формулами:

$$Z_0 = u_{0,1}X_1 + ... + u_{0,l}X_l,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$Z_{l-1} = u_{l-1,1}X_1 + ... + u_{l-1,l}X_l,$$

где $u_{k,i}$ – i-я компонента вектора $\vec{u}_k$, а X_i определяется соотношением

$$X_i = \frac{n_i - np_i}{\sqrt{np_i}}. \tag{13.4}$$

Из (13.3) и (13.4) следует, что

$$\chi^2 = \sum_{i=1}^{l} X_i^2, \tag{13.5}$$

а из ортонормированности базиса $\vec{u}_0,...,\vec{u}_{l-1}$ вытекает, что суммы квадратов величин $X_1,...,X_l$ и $Z_0,...,Z_{l-1}$ совпадают:

$$\sum_{i=1}^{l} X_i^2 = \sum_{k=0}^{l-1} Z_k^2. \tag{13.6}$$

Нетрудно видеть, что

$$Z_0 = \sqrt{p_1}\frac{n_1 - np_1}{\sqrt{np_1}} + ... + \sqrt{p_l}\frac{n_l - np_l}{\sqrt{np_l}} =$$

$$= \frac{1}{\sqrt{n}}\left(n_1 + ... + n_l - (p_1 + ... + p_l)\right) = 0.$$

Отсюда с учетом (13.5) и (13.6) получаем

$$\chi^2 = \sum_{k=1}^{l-1} Z_k^2. \tag{13.7}$$

Можно доказать, что случайный вектор $\vec{Z} = (Z_1,...,Z_{l-1})$ имеет нулевое математическое ожидание и единичную ковариационную матрицу. Если бы при этом вектор $\vec{Z}$ был еще и нормальным, то все его компоненты были бы независимыми и распределенными по стандартному нормальному закону $N(0,1)$. Отсюда вследствие (13.7) вытекало бы, что $\chi^2 \sim \chi^2(l-1)$. На самом деле, конечно, вектор $\vec{Z}$ не является нормальным ни при каком n. Однако применяя многомерный аналог центральной предельной теоремы, можно доказать, что при $n \to \infty$ распределение вектора $\vec{Z}$ стремится к $(l-1)$-мерному нормальному распределению. В итоге получаем,

что распределение статистики χ^2 Пирсона (13.3) при достаточно большом n близко к распределению χ^2 с $(l-1)$ степенями свободы.

Можно также доказать, что если гипотеза H_0 не верна, то при $n \to \infty$ вероятность $P(\chi^2 > c) \to 1$ для любого c, что в конечном счете определяет достаточно высокую мощность[1] критерия Пирсона, по крайней мере, для выборок большого объема.

В итоге приходим к заключению, что для проверки по эмпирическим данным (13.2) справедливости распределения (13.1) с асимптотическим уровнем значимости α можно использовать критерий согласия, основанный, на статистике χ^2 Пирсона (13.3) и критической области $\chi^2 > \chi^2_\alpha(l-1)$.

На практике данный критерий Пирсона применяется, если объем выборки $n > 50$ и все ожидаемые частоты $np_i > 5$. Несоблюдение данных условий обычно приводит к значительному отклонению фактического уровня значимости $P_{H_0}(\chi^2 > \chi^2_\alpha(l-1))$ от требуемого уровня α.

Пример 13.1. По результатам переписи населения установлен следующий возрастной состав

	до 50 лет	от 50 лет и старше
женщины	35 %	20 %
мужчины	35 %	10 %

Спустя несколько лет после переписи было отобрано случайным образом 1000 человек и для них подсчитано число мужчин и женщин в двух возрастных группах:

	до 50 лет	от 50 лет и старше
женщины	343	212
мужчины	343	102

[1] В данном случае под мощностью критерия согласия понимается мощность относительной какой-либо простой альтернативной гипотезы.

Требуется при уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что возрастной состав не изменился.

Решение. Рассмотрим события A_1, A_2, A_3, A_4, состоящие в том, что случайно отобранный человек оказался:

A_1 – женщиной до 50 лет,

A_2 – женщиной от 50 лет и старше,

A_3 – мужчиной до 50 лет,

A_4 – мужчиной от 50 лет и старше.

Очевидно, что события A_1, A_2, A_3, A_4 образуют полную группу, а минимальная ожидаемая частота $100 > 5$. С учетом малочисленности выборки относительно всего населения, считаем отборы людей независимыми испытаниями, поэтому проверяем гипотезу по критерию Пирсона. Вычисляем статистику критерия:

$$\chi^2 = \frac{(343-350)^2}{350} + \frac{(212-200)^2}{200} + \frac{(343-350)^2}{350} + \frac{(102-100)^2}{100} = 1{,}04.$$

Затем находим 5-процентную точку распределения хи-квадрат с 3 степенями свободы: $\chi^2_{0,05}(3) = 7{,}815$. Поскольку $1{,}04 < 7{,}815$, приходим к выводу, что данные выборочного исследования не обнаруживают значимого изменения возрастного состава населения.

2. Критерий хи-квадрат с оценкой параметров распределения

Случай полностью определенного гипотетического распределения (13.1) на практике встречается достаточно редко. Гораздо чаще встречаются ситуации, когда гипотетическая вероятностная мера P зависит от одного или нескольких параметров $\theta_1, ..., \theta_r$,

$$P(A) = P(A; \theta_1, ..., \theta_r),$$

что приводит к зависимости вероятностей p_i в (13.1) от $\theta_1, ..., \theta_r$:

$$\begin{aligned} p_1 &= p_1(\theta_1, ..., \theta_r) = P(A_1; \theta_1, ..., \theta_r), \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ p_l &= p_l(\theta_1, ..., \theta_r) = P(A_l; \theta_1, ..., \theta_r). \end{aligned} \qquad (13.8)$$

Проверяемая (возможно непараметрическая) гипотеза H часто формулируется как утверждение о том, что истинная вероят-

ностная мера $P_{\text{ист}}$ совпадает с гипотетической мерой P при определенных значениях параметров $\theta_1,...,\theta_r$. Тем не менее, фактической проверке подвергается не H, а вытекающая из H гипотеза H_0 о том, что распределение частот (13.2) согласуется с распределением (13.1) при определенном наборе значений параметров $\theta_1,...,\theta_r$ (13.8). Заметим, что при любой H гипотеза H_0 заведомо является параметрической гипотезой поскольку распределение (13.1) определяется конечным числом параметров.

Сложность гипотезы H_0 приводит к необходимости предварительной оценки параметров $\theta_1,...,\theta_r$ тем или иным способом. Предположим, что $\theta_1^*,...,\theta_r^*$ – оценки, полученные методом максимального правдоподобия по таблице эмпирических частот (13.2):

$$\begin{aligned} \theta_1^* &= \theta_1^*(n_1,...,n_l), \\ &\cdots\cdots\cdots\cdots\cdots \\ \theta_r^* &= \theta_r^*(n_1,...,n_l). \end{aligned} \tag{13.9}$$

Заменив в (13.3) вероятности p_i на их оценки $p_i^* = p_i(\theta_1^*,...,\theta_r^*)$, получим статистику вида

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - np_i^*)^2}{np_i^*}. \tag{13.10}$$

Произведение np_i^* по-прежнему интерпретируется как ожидаемая частота события A_i, несмотря на то, что np_i^* – случайная величина (в отличие от np_i).

Оказывается, что при выполнении некоторых достаточно общих условий (подробности в [6]) распределение статистики (13.10) при $n \to \infty$ стремится к распределению $\chi^2(l-r-1)$, в котором число степеней свободы по сравнению с аналогичным предельным распределением из предыдущего пункта уменьшилось на количество оцененных параметров.

Лекция 14
Проверка гипотезы о виде генерального распределения

§14.1. Проверка гипотезы о виде генерального распределения по критерию хи-квадрат

Предположим, что проверяется гипотеза о совпадении истинного распределения признака X в генеральной совокупности с некоторым гипотетическим распределением вида

X	x_1	...	x_l
P	p_1	...	p_l

, (14.1)

где $p_1,...,p_l$ известные положительные числа.

Для каждого значения x_i $(i=1,...,l)$ событие, состоящее в том, что X принимает значение x_i, обозначим A_i. Поскольку (14.1) – это закон распределения, события $A_1,...,A_l$ образуют полную группу.

Пусть $X_1,...,X_n$ – случайные значения X в серии n повторных независимых испытаний, n_i – частота конкретного значения x_i. Таким образом, n_i – это число тех значений среди $X_1,...,X_n$, которые равны x_i. Таблица наблюдаемых частот имеет вид

значение	x_1	...	x_l
частота	n_1	...	n_l

. (14.2)

Гипотезу о совпадении истинного распределения X с гипотетическим распределением (14.1) можно, очевидно, интерпретировать как гипотезу о соответствии наблюдаемых данных (14.2) распределению (14.1) и проверять ее по критерию χ^2 Пирсона, используя вместо таблиц (13.1) и (13.2) из предыдущего параграфа соответственно таблицы (14.1) и (14.2).

Пример 14.1. Четыре (возможно несимметричные) монеты подбрасываются 64 раза. Результаты бросков даны в следующей таблице:

число гербов	0	1	2	3	4
частота	5	8	32	14	5

.

Требуется при 5-процентном уровне значимости проверить гипотезу о распределении числа гербов по биномиальному закону с параметрами 4 и 0,5.

Решение. В соответствии с предполагаемым биномиальным распределением гипотетические вероятности $p_i = C_4^i \frac{1}{2^4}$, где i – число гербов в одном броске, $i = 0,...,4$. Находим p_i и соответствующие им ожидаемые частоты:

$$p_0 = \frac{1}{16}, p_1 = \frac{1}{4}, p_2 = \frac{3}{8}, p_3 = \frac{1}{4}, p_4 = \frac{1}{16};$$
$$np_0 = 4, np_1 = 16, np_2 = 24, np_3 = 16, np_4 = 4.$$

Как это часто случается, необходимое для применимости критерия Пирсона условие $np_i > 5$ оказалось нарушенным для крайних значений. Заменим исходную гипотезу о распределении числа гербов на какую-либо близкую по смыслу гипотезу, для которой условие $np_i > 5$ будет выполнено. На практике с этой целью производят объединение малочастотных событий с ближайшими событиями A_i, для которых $np_i > 5$. В нашем примере гипотезу о распределении числа гербов заменим на гипотезу о распределении вероятностей следующих событий:

A_1 = {число гербов равно 0 или 1},

A_2 = { число гербов равно 2},

A_3 = { число гербов равно 3 или 4},

Распределение частот событий A_1, A_2, A_3 представим в виде таблицы:

число гербов	0 – 1	2	3 – 4
наблюдаемые частоты	13	32	19
ожидаемые частоты	20	24	20

Находим значение статистики

$$\chi^2 = \frac{(13-20)^2}{20} + \frac{(32-24)^2}{24} + \frac{(19-20)^2}{20} = 5{,}17$$

и критические значения, соответствующие $\alpha = 0{,}05$,

$$\chi^2_{\text{крит}} = \chi^2_{0{,}05}(2) = 5{,}99.$$

Поскольку $\chi^2 < \chi^2_{\text{крит}}$, приходим к выводу, что значимого расхождения между наблюдаемыми и ожидаемыми частотами не обнаруживается, т.е. гипотеза о биномиальном распределении подтверждается (хотя и в несколько «огрубленном» виде).

1. Проверка гипотезы об определенном непрерывном распределении

Предположим, что распределение признака X в генеральной совокупности задается непрерывной функций $F(x)$, не содержащей каких-либо неизвестных параметров. Поскольку ожидаемая частота любого конкретного значения X равна 0, гипотеза H о распределении X уже в исходной постановке редуцируется к гипотезе H_0 о распределении вероятностей событий $A_1,...,A_l$, таких что $A_i = \{X \in \Delta_i\}$, где $\Delta_1,...,\Delta_l$ – некоторый набор попарно непересекающихся интервалов, $i = 1,...,l$. Дальнейшая проверка гипотезы H_0 проводится стандартным образом по критерию χ^2 Пирсона.

Заметим, что несмотря на наличие некоторых рекомендаций по выбору интервалов $\Delta_1,...,\Delta_l$ (предлагается, например, для всех интервалов кроме крайних устанавливать одинаковую длину) остается достаточно большой произвол в выборе этих интервалов. В результате весьма вероятной является ситуация, когда для заданного уровня значимости α при одном наборе интервалов $\Delta_1,...,\Delta_l$ гипотеза отвергается, а при другом наборе – принимается.

Указанное обстоятельство является недостатком критерия Пирсона с точки зрения проверки исходной гипотезы H, что, впрочем, не касается проверки гипотезы H_0. Таким образом, если редуцированная гипотеза H_0 имеет самостоятельный интерес, то для ее проверки имеет смысл воспользоваться критерием Пирсона, при этом для проверки исходной гипотезы H лучше применять другие критерии, например, критерий Колмогорова, о котором речь пойдет в § 14.2.

2. Проверка гипотезы о непрерывном распределении, зависящем от неизвестных параметров

Рассмотрим, наконец, наиболее сложный случай, когда непрерывная гипотетическая функция распределения $F(x) = F(x; \theta_1, ..., \theta_r)$ зависит от r неизвестных параметров $\theta_1, ..., \theta_r$. Проверяемая гипотеза H состоит в том, что истинная функция распределения $F_{\text{ист}}(x)$ совпадает с гипотетической функцией распределения при некоторых значениях параметров:

$$F_{\text{ист}}(x) = F(x; \theta_1^{\text{ист}}, ..., \theta_r^{\text{ист}}).$$

Заметим, что гипотеза H, несмотря на то, что гипотетическая функция распределения зависит от конечного числа параметров, не является параметрической! Действительно, базисное предположение состоит в том, что $F_{\text{ист}}(x)$ – произвольная функция распределения, а множество всех функций распределения нельзя задать конечным числом параметров. Тем не менее, проверка H сводится к проверке параметрической гипотезы H_0 путем задания набора непересекающихся интервалов $\Delta_1, ..., \Delta_l$, для которых сумма гипотетических вероятностей

$$\sum_{i=1}^{l} P(X \in \Delta_i) = 1$$

при любых допустимых значениях параметров $\theta_1, ..., \theta_r$. Дальнейшая проверка гипотезы H_0 производится по критерию χ^2 Пирсона с оценкой параметров распределения (см. п.2 предыдущей лекции).

Следует отметить, что оценка параметров $\theta_1,...,\theta_r$ производится на основе группированных данных, т.е. по таблице интервальных частот

интервал	Δ_1	...	Δ_l
частота	n_1	...	n_l

, (14.3)

где n_i – это число тех наблюдаемых значений переменной X, которые попали в интервал $\Delta_i, i = 1,...,l$.

Таблица (14.3) получается, очевидно, из (13.2) в случае $A_i = \{X \in \Delta_i\}$. Наиболее трудным моментом является построение оценки максимального правдоподобия по группированным данным (14.3). Ограничимся лишь случаем проверки гипотезы о нормальном распределении с неизвестным генеральным средним μ и неизвестной дисперсией σ^2.

Предположим, что $l \geq 4$, а все интервалы кроме Δ_1 и Δ_l имеют одинаковую длину h: $\Delta_i = (x_i, x_{i+1})$, где $x_{i+1} = x_i + h$, $i = 2,...,l-1$. В качестве Δ_1 и Δ_l возьмем бесконечные интервалы: $\Delta_1 = (-\infty, x_2)$, $\Delta_l = (x_l, \infty)$. Выберем в каждом интервале точку $x_i^* \in \Delta_i$, положив

$$x_1^* = x_2 - \frac{h}{2},$$

$$x_i^* = x_i + \frac{h}{2}, \quad i = 2,...,l.$$

Таким образом, все точки $x_1^*,...,x_l^*$ образуют арифметическую прогрессию с шагом h, а точки x_i^* для $i = 2,...,l-1$ являются серединами интервалов Δ_i. Как показано в [6] метод максимального правдоподобия позволяет получить следующие оценки:

$$\mu^* = \frac{1}{n}\sum_{i=1}^{l} x_i^* n_i, \tag{14.4}$$

$$\sigma^{*2} = \frac{1}{n}\sum_{i=1}^{l} (x_i^* - \mu^*)^2 n_i - \frac{1}{12}h^2. \tag{14.5}$$

Отметим, что μ^* обычно мало отличается от выборочного среднего $\bar{x}$, а σ^{*2} – от выборочной дисперсии $\hat{\sigma}^2$, при этом слагаемое $-h^2/12$ в правой части (14.5) в точности соответствует поправке Шеппарда (см. §2.2).

§14.2. Критерий Колмогорова

Для любого $x \in \mathrm{R}^n$ число компонент вектора $\vec{x} = (x_1,...,x_n)$, которые меньше x, обозначим $m(x,\vec{x})$. Для случайного вектора $\vec{X} = (X_1,...,X_n)$ обозначение $m(x,\vec{X})$ имеет тот же смысл, но при этом $m(x,\vec{X})$ является дискретной случайной величиной с возможными значениями $0,1,...,n$. Пусть $\vec{x}$ – реализация случайной выборки $\vec{X}$ объема n из некоторого распределения с функцией $F(x)$. Эмпирическую функцию распределения, соответствующую выборке $\vec{x}$, можно записать в виде

$$\hat{F}(x) = \hat{F}(x,\vec{x}) = \frac{m(x,\vec{x})}{n}.$$

Оценка функции $F(x)$ по случайной выборке $\vec{X}$ записывается аналогично:

$$\hat{F}(x) = \hat{F}(x,\vec{X}) = \frac{m(x,\vec{X})}{n}.$$

Заметим, $\hat{F}(x,\vec{x})$ – числовая функция, тогда как $\hat{F}(x,\vec{X})$ в каждой точке x принимает случайное значение, т.е. является *случайным процессом*.

Определим расстояние между функциями $\hat{F}(x)$ и $F(x)$ формулой

$$d = \sup_x |\hat{F}(x) - F(x)|. \qquad (14.6)$$

Для функции $\hat{F}(x) = \hat{F}(x,\vec{x})$ расстояние $d = d(\vec{x})$ – это просто число, тогда как для $\hat{F}(x) = \hat{F}(x,\vec{X})$ расстояние $d = d(\vec{X})$ является случайной величиной, принимающей значения на отрезке $[0,1]$.

Согласно доказанной А.Н. Колмогоровым теореме в случае непрерывной функции $F(x)$ при любом неотрицательном $u \geq 0$ существует предел

$$\lim_{n\to\infty} P\left(\sqrt{n}d(\vec{X}) < u\right) = K(u), \tag{14.7}$$

где

$$K(u) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2u^2}.$$

Вследствие этой теоремы критерий согласия с критической областью

$$\sqrt{n}d(\vec{x}) > u_\alpha, \tag{14.8}$$

где u_α – корень уравнения $K(u) = 1 - \alpha$, имеет при $n \to \infty$ уровень значимости, стремящийся к α. Другими словами, α – асимптотический уровень значимости. Именно этот критерий и называется *критерием Колмогорова*.

Так же как и рассмотренный в предыдущем параграфе критерий Пирсона, критерий Колмогорова применяется для проверки гипотезы о совпадении истинной функции распределения $F_{\text{ист}}(x)$ с некоторой гипотетической функцией распределения $F(x)$. Поскольку при $n < 20$ фактический уровень значимости заметно отличается от номинального значения α, критерий Колмогорова применяется при $n \geq 20$.

На практике, при вычислении максимального абсолютного отклонения гипотетической функции $F(x)$ от эмпирической функции $\hat{F}(x)$ применяется следующая формула:

$$d(\vec{x}) = \max_{1\leq i\leq n}\left\{\left|\frac{i}{n} - F\left(x_{(i)}\right)\right|, \left|\frac{i-1}{n} - F\left(x_{(i)}\right)\right|\right\}, \tag{14.9}$$

где $x_{(i)}$ – i-й член вариационного ряда

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}.$$

Эквивалентность (14.6) и (14.9) следует из того, что $F(x)$ – неубывающая функция, а $\hat{F}(x)$ принимает постоянное значение на интервалах:

$$(-\infty, x_{(1)});\ (x_{(1)}, x_{(2)}); ...; (x_{(n-1)}, x_{(n)});\ (x_{(n)}, \infty).$$

Критические значения u_α для некоторых α приведены в таблице

α	0,5	0,2	0,1	0,05	0,02	0,01	0,005	0,002	0,001
u_α	0,83	1,07	1,22	1,36	1,52	1,63	1,73	1,86	1,95

.

Лекция 15
Проверка однородности выборок

Пусть $\vec{X}_i = (X_{i1}, X_{i2}, ..., X_{in_i})$ – выборка из распределения $\mathcal{L}_i$ объема n_i, где $i = 1, ..., k$. Независимые выборки $\vec{X}_1, ..., \vec{X}_k$, полученные из одного и того же распределения $\mathcal{L}_1 = ... = \mathcal{L}_k$, называются *однородными*. В зависимости от характера базисного предположения гипотеза однородности нескольких выборок может быть как параметрической, так и непараметрической.

§15.1. Проверка непараметрических гипотез однородности

Для проверки гипотезы однородности нескольких выборок обычно используется критерий Пирсона, а в случае двух выборок применяются также критерии Смирнова и Колмогорова–Смирнова.

1. Проверка однородности по критерию Пирсона

Фиксируем разбиение числовой оси на l попарно непересекающихся интервалов $\Delta_1, ..., \Delta_l$. Число тех значений среди

$X_{i1},...,X_{in_l}$, которые попали в Δ_j обозначим n_{ij}. Нетрудно видеть, что $n_{i1}+...+n_{il}=n_i$ – объем *i*-ой выборки. Суммарную частоту интервала Δ_j относительно всех выборок обозначим $m_j=n_{1j}+...+n_{kj}$. Группированные статистические данные записываются в виде таблицы

№ выборки	Δ_1	Δ_2	...	Δ_l	объем
1	n_{11}	n_{12}	...	n_{1l}	n_1
2	n_{21}	n_{22}	...	n_{2l}	n_2
...	...	...	...	...	...
k	n_{k1}	n_{k2}	...	n_{kl}	n_k
частоты	m_1	m_2	...	m_l	n

,

в которой n – это суммарный объем всех выборок и, одновременно, суммарная частота всех интервалов.

Пусть $\mathcal{L}_i(\Delta_j)$ – вероятность того, что случайная величина, распределенная по закону $\mathcal{L}_i$ попадает в Δ_j. Если выборки $\vec{X}_1,...,\vec{X}_k$ однородны, то для любого Δ_j найдется такое неотрицательное число p_j, что

$$\mathcal{L}_1(\Delta_j)=\mathcal{L}_2(\Delta_j)=...=\mathcal{L}_k(\Delta_j)=p_j.$$

Таким образом, для данного набора интервалов $\Delta_1,...,\Delta_l$ однородность выборки означает справедливость гипотезы

$$H_0:\ \mathcal{L}_i(\Delta_j)=p_j,\ i=1,...,k;\ j=1,...,l;$$

которая, собственно, и проверяется по критерию Пирсона.

Рассмотрим сначала более простой случай, когда константы $p_1,...,p_l$ известны. Для каждой выборки $\vec{X}_i$ составим свою статистику хи-квадрат:

$$\chi_i^2=\sum_{j=1}^{l}\frac{(n_{ij}-n_ip_j)^2}{n_ip_j},\ i=1,...,k.$$

Если верна H_0, то асимптотически (при $n \to \infty$) статистика χ_i^2 распределена по закону хи-квадрат с $l-1$ степенями свободы, $\chi_i^2 \overset{a}{\sim} \chi^2(l-1)$. Следовательно, сумма

$$\chi^2 \equiv \sum_{i=1}^{k} \chi_i^2 \overset{a}{\sim} \chi^2\big(k \cdot (l-1)\big). \qquad (15.1)$$

Таким образом, для проверки гипотезы H_0 с асимптотическим уровнем значимости α можно использовать статистический критерий с критической областью $\chi^2 > \chi_\alpha^2(k \cdot (l-1))$.

Рассмотрим теперь практически более важный случай, когда $p_1,...,p_l$ неизвестны. Заменяя вероятность p_j ее эффективной оценкой $\hat{p}_j = m_j / n$, получим статистику Пирсона вида

$$\hat{\chi}^2 \equiv \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - n_i \hat{p}_j)^2}{n_i \hat{p}_j}.$$

Так как $p_1 + ... + p_l = 1$, то из $p_1,...,p_l$ независимы только $l-1$ величин, поэтому, фактически, на основе выборочных данных оценено лишь $l-1$ параметров распределения. Следовательно, асимптотическое (при верной H_0) хи-квадрат распределение статистики $\hat{\chi}^2$ содержит число степеней свободы на $l-1$ единиц меньше, чем аналогичное распределение (15.1):

$$\hat{\chi}^2 \overset{a}{\sim} \chi^2\big((k-1) \cdot (l-1)\big).$$

Вывод: *для проверки с приближенным уровнем значимости α гипотезы об однородности нескольких выборок с неизвестным распределением применим критерий Пирсона с критической областью $\hat{\chi}^2 > \chi_\alpha^2\big((k-1) \cdot (l-1)\big)$.*

Этот критерий Пирсона используется обычно при условии, что все $n_i \hat{p}_j \geq 5$. Так же как и для других подобных критериев Пирсона данное требование обусловлено тем, что при его нарушении возможно значительное отклонение фактического уровня значимости от α.

2. Проверка однородности по критериям Смирнова и Колмогорова–Смирнова

Предположим, что имеется только две выборки $\vec{X}_1^m$ и $\vec{X}_2^n$, объемы которых равны, соответственно, m и n. Пусть $\vec{x}_1^m$ – реализация $\vec{X}_1^m$ и $\vec{x}_2^n$ – реализация $\vec{X}_2^n$. Эмпирические функции распределения обозначим $\hat{F}_1(x;\vec{X}_1^m), \hat{F}_2(x;\vec{X}_2^n)$ для случайных выборок $\vec{X}_1^m$, $\vec{X}_2^n$ и $\hat{F}_1(x;\vec{x}_1^m), \hat{F}_2(x;\vec{x}_2^n)$ для их реализаций. Определим расстояние между функциями $\hat{F}_1(x) = \hat{F}_1(x;\vec{x}_1^m)$ и $\hat{F}_2(x) = \hat{F}_2(x;\vec{x}_2^n)$ формулой

$$d(\vec{x}_1^m, \vec{x}_2^n) = \sup_{x\in \mathrm{R}} \left|\hat{F}_1(x) - \hat{F}_2(x)\right|. \qquad (15.2)$$

Пусть X – конечное множество, образованное всеми различными компонентами выборок $\vec{x}_1^m$ и $\vec{x}_2^n$. Нетрудно видеть, что

$$d(\vec{x}_1^m, \vec{x}_2^n) = \max_{x\in X} \left|\hat{F}_1(x) - \hat{F}_2(x)\right|, \qquad (15.3)$$

поэтому точная верхняя грань в (15.2) всегда достигается.

Для случайных выборок расстояние $d(\vec{X}_1^m, \vec{X}_2^n)$ между $\hat{F}_1(x) = \hat{F}_1(x;\vec{X}_1^m)$ и $\hat{F}_2(x) = \hat{F}_2(x;\vec{X}_2^n)$ определяется аналогично (15.2).

Предположим, что выборки $\vec{X}_1^m$ и $\vec{X}_2^n$ однородны и, кроме того, извлечены из распределения с непрерывной функцией $F(x)$. Оказывается, что в этом случае распределение случайной величины $d(\vec{X}_1^m, \vec{X}_2^n)$ не зависит от $F(x)$. Чтобы доказать это определим случайные векторы

$$\vec{Y}_1^m = \left(F(X_{1,1}),\ldots,F(X_{1,m})\right),$$
$$\vec{Y}_2^n = \left(F(X_{2,1}),\ldots,F(X_{2,n})\right).$$

Компоненты векторов $\vec{Y}_1^m$ и $\vec{Y}_2^n$ независимы и равномерно распределены на отрезке [0,1]. Следовательно, $\vec{Y}_1^m$ и $\vec{Y}_2^n$ можно рассматривать как выборки объемов m и n из равномерного рас-

пределения на отрезке [0,1]. Независимость распределения $d(\vec{X}_1^m, \vec{X}_2^n)$ от $F(x)$ вытекает из того, что $d(\vec{X}_1^m, \vec{X}_2^n) = d(\vec{Y}_1^m, \vec{Y}_2^n)$.

Процентные точки $d_\alpha(m,n)$ распределения $d(\vec{Y}_1^m, \vec{Y}_2^n)$, как обычно, определяются соотношением $P\left(d(\vec{Y}_1^m, \vec{Y}_2^n) > d_\alpha(m,n)\right) = \alpha$. Для проверки гипотезы однородности двух выборок, полученных из непрерывных распределений, применяется *критерий Н.В. Смирнова*, критическая область которого задается неравенством $d(\vec{x}_1^m, \vec{x}_2^n) > d_\alpha(m,n)$.

Таблицы критических значений $d_\alpha(m,n)$ при небольших m и n можно найти в некоторых справочниках, например, в [3]. При больших m и n можно воспользоваться результатом Н.В. Смирнова, который доказал, что при $m,n \to \infty$ (и непрерывной $F(x)$)

$$P\left\{\sqrt{\frac{mn}{m+n}}\, d(\vec{X}_1^m, \vec{X}_2^n) < u\right\} \to K(u), \qquad (15.4)$$

где $K(u)$ – функция предельного распределения Колмогорова (см. формулу (14.7) из предыдущей лекции). Соответствующий критерий называется *критерием Колмогорова–Смирнова* и задается критической областью

$$\sqrt{\frac{mn}{m+n}}\, d(\vec{x}_1^m, \vec{x}_2^n) > u_\alpha,$$

где u_α – корень уравнения $K(u) = 1 - \alpha$. Как следует из (15.4) критерий Колмогорова–Смирнова имеет асимптотический уровень значимости α, т.е. его фактическое значение достигает α лишь в пределе при $m,n \to \infty$. На практике критерий Колмогорова–Смирнова применяется, если объемы обеих выборок не меньше 50.

§15.2. Проверка гипотезы о совпадении нескольких генеральных средних методом дисперсионного анализа

Пусть $\vec{X}_i = (X_{i1},...,X_{in_i})$ – выборка объема n_i из $N(\mu_i,\sigma^2)$, где $i = 1,...,k$. Предположим также, что $n = n_1 + ... + n_k$ случайных величин

$$X_{11},...,X_{1n_1}; X_{21},...,X_{2n_2}; ...; X_{k1},...,X_{kn_k}$$

независимы в совокупности. Таким образом, выборки $\vec{X}_1,...,\vec{X}_k$ независимы и получены из нормальных распределений с одинаковой дисперсией σ^2 и, возможно, различными средними $\mu_1,...,\mu_k$. Гипотеза о равенстве всех средних одновременно записывается как

H_0: $\mu_1 = ... = \mu_k$,

а альтернативная гипотеза – как

H_1: $(\exists ij)\, \mu_i \neq \mu_j$.

Заметим, что при верной H_0 выборки $\vec{X}_1,...,\vec{X}_k$ являются однородными, поскольку в этом случае генеральные распределения совпадают:

$$N(\mu_1,\sigma^2) = ... = N(\mu_k,\sigma^2).$$

Следовательно, гипотеза H_0 при сделанных базисных предположениях может рассматриваться как параметрический аналог рассмотренных ранее непараметрических гипотез однородности.

Рассмотрим объединенную выборку объема $n = n_1 + ... + n_k$:

$$\vec{X} = (X_{11},...,X_{1n_1}; ...; X_{k1},...,X_{kn_k}).$$

Интерпретируя выборки $\vec{X}_1,...,\vec{X}_k$ как группы, на которые разбита совокупность $\vec{X}$, введем обозначения, аналогичные тем, что использовались в лекции 2 при изучении межгрупповой дисперсии:

$$\vec{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_i$$

– выборочное среднее в *i*-й совокупности;

$$\hat{\sigma}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$$

– выборочная дисперсия в той же выборке;

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k}\overline{X}_i n_i = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}$$

– выборочное среднее в объединенной выборке $\vec{X}$;

$$\overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{k}\hat{\sigma}_i^2 n_i$$

– средняя групповая дисперсия;

$$\delta^2 = \frac{1}{n}\sum_{i=1}^{k}(\overline{X}_i - \overline{X})^2 n_i$$

– межгрупповая дисперсия;

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2$$

– выборочная дисперсия признака в объединенной выборке $\vec{X}$.

Из результатов лекции 2 следует, что выборочную дисперсию $\hat{\sigma}^2$ можно представить в виде суммы $\hat{\sigma}^2 = \overline{\sigma}^2 + \delta^2$, где первое слагаемое $\overline{\sigma}^2$ характеризует среднюю изменчивость признака в каждой выборке $\vec{X}_1,...,\vec{X}_k$, а второе слагаемое δ^2 характеризует разброс выборочных средних $\overline{X}_1,...,\overline{X}_k$.

Критерий проверки H_0 против H_1 основан на следующей теореме, которую приводим без доказательства.

Теорема 15.1 *Предположим, что верна гипотеза* H_0. *Тогда*:

1) $\frac{n\delta^2}{\sigma^2} \sim \chi^2(k-1)$;

2) $\frac{n\overline{\sigma}^2}{\sigma^2} \sim \chi^2(n-k)$;

3) δ^2 *и* $\overline{\sigma}^2$ *независимы*.

Определим так называемое F-отношение:

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{k}(\overline{X}_i - \overline{X})^2 n_i}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2}. \qquad (15.5)$$

Статистику F можно представить также в виде

$$F = \frac{(n\delta^2)/(k-1)}{(n\overline{\sigma}^2)/(n-k)} = \frac{(n\delta^2/\sigma^2)/(k-1)}{(n\overline{\sigma}^2/\sigma^2)/(n-k)}. \qquad (15.6)$$

Из (15.6) и теоремы 15.1 вытекает, что

$$F \sim F(k-1, n-k),$$

где $F(k-1, n-k)$ – распределение Фишера с $k-1$ и $n-k$ степенями свободы. Для проверки H_0 с уровнем значимости α применяется критерий с критической областью

$$F > F_\alpha(k-1, n-k),$$

где $F_\alpha(k-1, n-k)$ – 100α-процентная точка распределения Фишера с $k-1$ и $n-k$ степенями свободы.

Лекция 16
Метод наименьших квадратов и парная регрессия

§16.1. Приближенное решение системы линейных уравнений

Рассмотрим линейную систему

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1, \\ a_{21}x_1 + a_{22}x_2 = b_2, \\ a_{31}x_1 + a_{32}x_2 = b_3. \end{cases} \tag{16.1}$$

Система (16.1), вообще говоря, несовместна. После подстановки в нее произвольной пары чисел (x_1, x_2) одно или несколько уравнений будут нарушены. *Отклонением* (или *невязкой*) i-го уравнения системы (16.1) называется разность между его левой и правой частями:

$$e_i = a_{i1}x_{11} + a_{i2}x_2 - b_i.$$

Сумма квадратов отклонений во всех уравнениях далее обозначается

$$S(x_1, x_2) = e_1^2 + e_2^2 + e_3^2.$$

Метод наименьших квадратов (МНК) состоит в том, что приближенное решение системы (16.1) ищется как точное решение оптимизационной задачи

$$S(x_1, x_2) \to \min. \tag{16.2}$$

Способ построения решения (16.2), фактически, не зависит от числа неизвестных и уравнений в исходной системе линейных уравнений. Однако ограничимся пока подробным рассмотрением простейшей системы (16.1), чтобы максимально упростить геометрическую интерпретацию МНК.

Пусть $\vec{a}_1, \vec{a}_2$ – столбцы коэффициентов при x_1, x_2 в системе (16.1). Предположим, что $\vec{a}_1$ и $\vec{a}_2$ линейно независимы. Тогда множество всех линейных комбинаций $\Pi = \{x_1\vec{a}_1 + x_2\vec{a}_2\}$ векторов $\vec{a}_1$ и $\vec{a}_2$ является плоскостью в R^3. Нетрудно видеть, что $S(x_1, x_2)$ – это квадрат расстояния от точки $x_1\vec{a}_1 + x_2\vec{a}_2$ до точки $\vec{b} = (b_1, b_2, b_3)^T$. Пусть $\vec{b}^* = x_1^*\vec{a}_1 + x_2^*\vec{a}_2$ – ортогональная проекция $\vec{b}$ на плоскость Π. Так как $\vec{b}^*$ – ближайшая к $\vec{b}$ точка плоскости Π, ее координаты x_1^*, x_2^* на плоскости Π являются решением задачи (16.2) и одновременно приближенным решением системы (16.1).

Чтобы найти x_1^*, x_2^* заметим, что вектор $\vec{b}^* - \vec{b}$ ортогонален плоскости Π. Следовательно, для $i = 1,2$ имеем

$$(\vec{a}_i, \vec{b}^* - \vec{b}) = 0,$$

что эквивалентно системе

$$\begin{cases} (\vec{a}_1, \vec{b}^*) = (\vec{a}_1, \vec{b}), \\ (\vec{a}_2, \vec{b}^*) = (\vec{a}_2, \vec{b}). \end{cases} \tag{16.3}$$

Поскольку $\vec{b}^* = x_1^*\vec{a}_1 + x_2^*\vec{a}_2$, из (16.3) следует, что вектор (x_1^*, x_2^*) является решением системы

$$\begin{cases} (\vec{a}_1, \vec{a}_1)x_1 + (\vec{a}_1, \vec{a}_2)x_2 = (\vec{a}_1, \vec{b}), \\ (\vec{a}_2, \vec{a}_1)x_1 + (\vec{a}_2, \vec{a}_2)x_2 = (\vec{a}_2, \vec{b}), \end{cases}$$

которую удобно записать в матричном виде:

$$A^T A\vec{x} = A^T\vec{b}, \tag{16.4}$$

где $A = (a_{ij})$ – матрица коэффициентов, а $\vec{x}$ – столбец неизвестных в (16.1). Из (16.4) находим МНК-решение системы (16.1):

$$\vec{x} = (A^T A)^{-1} A^T\vec{b}. \tag{16.5}$$

Чуть более общие рассуждения показывают, что формула (16.5) задает МНК-решение записанной в матричном виде $A\vec{x} = \vec{b}$ линейной системы с произвольным числом неизвестных и уравнений. Единственное ограничение состоит в том, чтобы столбцы мат-

рицы A были линейно независимы. Несложно доказывается (см., например, [10, часть 1]), что при этом условии обратная матрица $(A^T A)^{-1}$ существует, что обеспечивает существование и единственность решения системы (16.4).

Количество приложений метода наименьших квадратов столь велико, что не представляется возможным даже простое перечисление всех областей, в которых МНК успешно применялся. В Приложениях 2 и 3 приводятся программы, использующие МНК для исследования статистики финансовых временных рядов.

§16.2. Факторная дисперсия и коэффициент детерминации

Предположим, что на плоскости задано n точек $(x_1, y_1), \ldots, (x_n, y_n)$ и необходимо подобрать прямую $y = \alpha + \beta x,$ проходящую как можно ближе к этим точкам. Если бы все точки лежали на прямой $y = \alpha + \beta x,$ то коэффициенты α и β были бы точным решением системы

$$\begin{cases} \alpha + x_1 \beta = y_1, \\ \ldots\ldots\ldots\ldots \\ \alpha + x_n \beta = y_n. \end{cases} \tag{16.6}$$

На самом деле точки $(x_1, y_1), \ldots, (x_n, y_n)$ обычно не лежат на одной прямой и система (16.6) является несовместной. Тем не менее, коэффициенты α и β искомой прямой легко находятся как МНК-решения системы (16.6).

Действительно, записав (16.6) в матричном виде

$$A\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \vec{b}, \tag{16.7}$$

где

$$A = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

и применяя формулу (16.5), получим МНК-решение системы (16.6):

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix}=\left\{\begin{pmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_3\end{pmatrix}\begin{pmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_n\end{pmatrix}\right\}^{-1}\begin{pmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_3\end{pmatrix}\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}=$$

$$=\left\{n\begin{pmatrix}1 & \overline{x}\\ \overline{x} & \overline{x^2}\end{pmatrix}\right\}^{-1}\left\{n\begin{pmatrix}\overline{y}\\ \overline{xy}\end{pmatrix}\right\}=$$

$$=\frac{1}{\overline{x^2}-\overline{x}^2}\begin{pmatrix}\overline{x^2} & -\overline{x}\\ -\overline{x} & 1\end{pmatrix}\begin{pmatrix}\overline{y}\\ \overline{xy}\end{pmatrix}=$$

$$=\frac{1}{\overline{x^2}-\overline{x}^2}\begin{pmatrix}\overline{x^2}\cdot\overline{y}-\overline{x}\cdot\overline{xy}\\ \overline{xy}-\overline{x}\cdot\overline{y}\end{pmatrix}. \qquad (16.8)$$

Исходные данные $x_1,...,x_n;y_1,...,y_n$ далее интерпретируются как значения некоторых признаков X и Y в совокупности $\hat{\Omega}=\{1,...,n\}$:

$$x_i=X(i),\quad y_i=Y(i),\quad i\in\hat{\Omega}.$$

Используя дисперсию $\hat{D}$ и ковариацию $\hat{\mathrm{Cov}}$ признаков X и Y в совокупности $\hat{\Omega}$, представим $\hat{\beta}$ в виде:

$$\hat{\beta}=\frac{\hat{\mathrm{Cov}}(X,Y)}{\hat{D}(X)}. \qquad (16.9)$$

Определим также на $\hat{\Omega}$ признаки $\hat{Y}$ и E, положив

$$\hat{Y}(i)=\hat{y}_i=\hat{\alpha}+\hat{\beta}x_i,$$
$$E(i)=e_i=y_i-\hat{y}_i,\quad i\in\hat{\Omega}.$$

Предположим, что необходимо объяснить «изменчивость» переменной Y за счет приближенной линейной связи $Y\approx\alpha+\beta X$.

При таком подходе дисперсию $\hat{D}(\hat{Y})$ естественно назвать дисперсией, *объясненной* зависимостью Y от фактора X, или *факторной* дисперсией. Далее будет доказано, что $\hat{D}(Y) = \hat{D}(E) + \hat{D}(\hat{Y})$, поэтому *остаточную* дисперсию $\hat{D}(E)$ можно трактовать как часть дисперсии переменной Y, которая осталась необъясненной. Также заметим, что разности $e_1,...,e_n$ называются *остатками*, поэтому остаточная дисперсия $\hat{D}(E)$ – это еще и эмпирическая дисперсия совокупности остатков.

Теорема 16.1 *Эмпирическая дисперсия переменной Y равна сумме остаточной и факторной дисперсий*,

$$\hat{D}(Y) = \hat{D}(E) + \hat{D}(\hat{Y}).$$

Доказательство. Имеем

$$\hat{D}(Y) = \hat{D}(Y - \hat{Y} + Y) = \hat{D}(E) + \hat{D}(\hat{Y}) + 2\hat{\mathrm{Cov}}(E, \hat{Y}).$$

Далее

$$\hat{\mathrm{Cov}}(E, \hat{Y}) = \hat{\mathrm{Cov}}(Y - \hat{Y}, \hat{Y}) = \hat{\mathrm{Cov}}(Y, \hat{Y}) - \hat{\mathrm{Cov}}(\hat{Y}, \hat{Y}) =$$
$$= \hat{\mathrm{Cov}}(Y, \hat{\alpha} + \hat{\beta}X) - \hat{D}(\hat{\alpha} + \hat{\beta}X) = \hat{\beta}\,\hat{\mathrm{Cov}}(Y, X) - \hat{\beta}^2\hat{D}(X) =$$
$$= \hat{\beta}\left\{\hat{\mathrm{Cov}}(Y, X) - \hat{\mathrm{Cov}}(Y, X)\right\} = 0.$$

Определение. *Доля R^2 дисперсии переменной Y, объясненной приближенной линейной зависимостью от переменной X, называется* ***коэффициентом детерминации***:

$$R^2 = R^2_{YX} = \frac{\hat{D}(\hat{Y})}{\hat{D}(Y)}. \qquad (16.10)$$

Из теоремы 16.1 немедленно следует, что $R^2 \in [0,1]$, при этом значение $R^2 = 1$ возможно только в случае, если остаточная дисперсия $\hat{D}(E) = 0$, т.е. когда все точки (x_i, y_i) лежат на одной прямой.

Необходимо также отметить, что коэффициент детерминации R^2, фактически, совпадает с квадратом эмпирического коэффициента корреляции r_{XY}:

$$R^2_{YX} = \frac{\hat{D}(\hat{Y})}{\hat{D}(Y)} = \frac{\hat{\beta}^2 \hat{D}(X)}{\hat{D}(Y)} = \frac{\hat{\mathrm{Cov}}^2(X,Y)}{\hat{D}(X)\hat{D}(Y)} = r^2_{XY}.$$

Отсюда, между прочим, вытекает, что коэффициент детерминации обладает свойством симметрии $R^2_{XY} = R^2_{YX}$, совершенно неочевидным при исходном определении (16.10).

§16.3. Линейная модель парной регрессии с детерминированным регрессором

Пусть имеется n неслучайных величин $x_1,...,x_n$ и n случайных величин $Y_1,...,Y_n$. Рассматриваемая в этой лекции модель парной регрессии основана на следующих предположениях.

1. $Y_i = \alpha + \beta x_i + \varepsilon_i$, где α и β – неслучайные, а ε_i – случайные величины.

2. Вектор $(x_1,...,x_n)$ не коллинеарен вектору $(1,...,1)$.

3. Для *ошибок* ε_i выполняются соотношения:

а) $E(\varepsilon_i) = 0, \quad i = 1,...,n;$

б) $D(\varepsilon_i) = \sigma^2, \quad i = 1,...,n;$

в) $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i, j = 1,...,n$ и $i \neq j$.

Из 3а следует, что средним значением Y_i будет

$$E(Y_i) = \alpha + \beta x_i,$$

поэтому величина $\alpha + \beta x_i$ интерпретируется (при известных α и β) как прогноз Y_i:

$$Y_i \approx \alpha + \beta x_i. \qquad (16.11)$$

Используя условие 3б, нетрудно проверить, что σ^2 – это среднеквадратичная ошибка прогноза (16.11).

Теорема (*Гаусса–Маркова*)16.2 *Если выполняются условия* 1–3, *то оценки параметров* α *и* β, *полученные методом наименьших квадратов*, *являются несмещенными и имеют наименьшую дисперсию в классе всех несмещенных и линейных по* Y_i *оценок.*

Доказательство проведем с применением матричных обозначений, благодаря чему оно легко обобщается на случай нескольких детерминированных регрессоров. Определим векторы

$$\vec{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \vec{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

и матрицу

$$A = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

С учетом данных обозначений модель парной регрессии примет вид:

$$\vec{Y} = A\vec{\beta} + \vec{\varepsilon}. \tag{16.12}$$

Применяя МНК к (16.12), получим вектор оценок параметров модели:

$$\vec{\beta}_{МНК} = (\hat{\alpha}_{МНК}, \hat{\beta}_{МНК})^T = (A^T A)^{-1} A^T \vec{Y}. \tag{16.13}$$

Формула (16.13) аналогична полученной ранее формуле (16.8), отличие – в том, что роль детерминированных величин y_i теперь играют случайные величины Y_i. Несмещенность оценок $\hat{\alpha}_{МНК}, \hat{\beta}_{МНК}$ означает справедливость следующих равенств:

$$E(\hat{\alpha}_{МНК}) = \alpha, \quad E(\hat{\beta}_{МНК}) = \beta. \tag{16.14}$$

Вместо того чтобы по отдельности находить математические ожидания (16.14), найдем математическое ожидание вектора оценок:

$$E\left(\vec{\beta}_{МНК}\right) = E\left((A^T A)^{-1} A^T \vec{Y}\right) =$$
$$= (A^T A)^{-1} A^T E\left(\vec{Y}\right) =$$
$$= (A^T A)^{-1} A^T A\vec{\beta} = \vec{\beta}.$$

Итак, несмещенность МНК-оценок доказана. Докажем минимальность дисперсии. Пусть a и b – какие-либо линейные оценки α и β:

$$a = c_{11}Y_1 + ... + c_{1n}Y_n,$$
$$b = c_{21}Y_1 + ... + c_{2n}Y_n.$$

Произвольную линейную по Y_i оценку вектора параметров $\vec{\beta}$ можно, очевидно, представить как $C\vec{Y}$, где $C = (c_{ij})$. Несмещенность (при любых α и β) оценки $C\vec{Y}$ означает, что

$$E\left(C\vec{Y}\right) = CE\left(\vec{Y}\right) = CA\vec{\beta} = \vec{\beta}.$$

Отсюда следует, что $CA = I_2$ – единичная (2×2)-матрица.

Пусть $B = A^T A, \quad M = C - B^{-1}A^T$. Поскольку

$$(B^{-1})^T = ((A^T A)^{-1})^T = (A^T A^{TT})^{-1} = B^{-1},$$

с учетом $CA = I_2$ найдем

$$MM^T = (C - B^{-1}A^T)(C - B^{-1}A^T)^T =$$
$$= CC^T - C(B^{-1}A^T)^T - B^{-1}A^T C^T + B^{-1}A^T(B^{-1}A^T)^T =$$
$$= CC^T - CA\cdot(B^{-1})^T - B^{-1}\cdot(CA)^T + B^{-1}\cdot A^T A\cdot(B^{-1})^T =$$
$$= CC^T - B^{-1}.$$

Следовательно, CC^T можно представить как $CC^T = MM^T + B^T$. Очевидно, что диагональные элементы любой матрицы вида MM^T неотрицательны, поэтому матрица CC^T имеет минимальные диагональные элементы лишь при условии, что $C = B^{-1}A^T$.

Отсюда уже легко получить минимальность дисперсии МНК-оценок параметров α и β. Действительно, $D(a)$ и $D(b)$ – это диагональные элементы ковариационной матрицы $\mathrm{Cov}(C\vec{Y}, C\vec{Y})$. Вычислив

$$\begin{aligned}
\mathrm{Cov}(C\vec{Y}, C\vec{Y}) &= E([C\vec{Y} - E(C\vec{Y})][C\vec{Y} - E(C\vec{Y})]^T) = \\
&= E([C\vec{Y} - CA\vec{\beta})][C\vec{Y} - CA\vec{\beta}]^T) = \\
&= E(C[\vec{Y} - A\vec{\beta})][\vec{Y} - A\vec{\beta}]^T C^T) = \\
&= E(C\vec{\varepsilon}\vec{\varepsilon}^T C^T) = CE(\vec{\varepsilon}\vec{\varepsilon}^T)C^T = \\
&= C \cdot \sigma^2 I_n \cdot C^T = \sigma^2 CC^T,
\end{aligned}$$

убеждаемся в том, что оценки a и b имеют минимальную дисперсию только в случае $C = B^{-1}A^T$, т.е. когда они получены по методу наименьших квадратов.

Литература

1. *Бабайцев В.А.*, *Браилов А.В.*, *Солодовников А.С.* Математика в экономике. Теория вероятностей: Курс лекций. — М.: ФА, 2002. — 232 с.
2. *Беляев Ю.К.*, *Носко В.П.* Основные понятия и задачи математической статистики. — М.: Изд-во МГУ, ЧеРо,1998. — 192 с.
3. *Большев Л.Н.*, *Смирнов Н.В.* Таблицы математической статистики. — М.: Наука, 1983. — 416 с.
4. *Горяинов В.Б.*, *Павлов И.В.*, *Цветкова Г.М.*, *Тескин И.О.* Математическая статистика. — М.: Изд-во МГТУ им. Н.Э. Баумана, 2002. — 424 с.
5. *Колемаев В.А.*, *Калинина В.Н.* Теория вероятностей и математическая статистика. — М.: ИНФРА-М, 1999. — 302 с.
6. *Крамер Г.* Математические методы статистики. — Москва–Ижевск: НИЦ «Регулярная и хаотическая динамика», 2003. — 648 с.
7. *Кремер Н.Ш.* Теория вероятностей и математическая статистика. — М.: ЮНИТИ-ДАНА, 2000. — 543 с.
8. *Пугачев В.С.* Теория вероятностей и математическая статистика. — М.: ФИЗМАТЛИТ, 2002. — 496 с.
9. *Соколов Г.А.*, *Гладких И.М.* Математическая статистика. — М.: Изд-во «Экзамен», 2004. — 432 с.
10. *Солодовников А.С.*, *Бабайцев В.А.*, *Браилов А.В.* Математика в экономике. — М.: Финансы и статистика, 2003. Часть 1 — 384 с. Часть 2 — 560 с.
11. *Ширяев А.Н.* Основы стохастической финансовой математики. — М.: Фазис, 1998. Т.1. — 489 с.

Приложение 1

Матричный калькулятор

Матричный калькулятор – это интерпретатор языка матричных вычислений, содержащий базу данных по мировым индексам, курсам валют и акциям в РТС. Текущая версия: 2.7.0.

Краткое описание языка

Комментарий – текст, начинающийся с // и заканчивающийся концом строки. Комментарии не участвуют в вычислениях и используются для улучшения читаемости программы.

Литералы существуют двух видов: десятичные и строковые.

Примеры

0.05 – десятичный литерал (число).
"LKOH" – строковый литерал (строка символов).

Имя – последовательность букв, цифр и символов подчеркивания, начинающаяся с буквы или символа подчеркивания.

Знаки операций

Стандартная арифметика

+ , –, *, / – плюс, минус, умножить, разделить;
% – остаток от деления.

Матричные операции

&* – матричное умножение;
&^ – матричное возведение в степень;

Операции сравнения

>, < – больше, меньше;
>=, <= – больше или равно, меньше или равно;
==, != – равно, не равно.

Логические операции

&, |, ! – и, или, не.

Выражение – комбинация литералов, имен, знаков операций и скобок.

Примеры

Выражение	Значение
1+2^3	9
(1+2)^3	27
8/4*2	4
2^2^3	256
3>2	1 (истина)
3<2	0 (ложь)
8 == 2^3	1 (истина)
2<3 & 3<4	1 (истина)

Замечание. Входящие в выражение имена должны быть определены до начала вычисления выражения. Так, имена чисел и матриц определяются оператором присваивания, имена функций – оператором function и т.д. В то же время, существует ряд имен, определенных до начала выполнения любой программы: pi, sin, green, ...

Программа – это ряд операторов произвольных типов. Выполнение программы состоит в последовательном выполнении операторов. После выполнения всех операторов на экран, как правило, выводится последнее вычисленное значение. Программа может содержать следующие операторы:

оператор-выражение,

оператор присваивания,

операторы function, law, if-else и for,

а также составные операторы.

Матрицы

Существует несколько способов построения матриц.

Некоторые матрицы можно получить с помощью матричных функций. Например, функция E(n) создает единичную матрицу размера (n×n).

Любую матрицу A=(a_{ij}) можно задать выражением

[a_{11},a_{12},...,a_{1n}; a_{21},...,a_{2n}; ...; a_{m1},...,a_{mn}],

где a_{ij} – скалярные выражения. Элементы строк разделяются запятой, строки – точкой с запятой. Однако выражения a_{ij} могут быть и матричными. Например, выражение

```
[E(n),E(n)]
```

задает (n×2n)-матрицу, составленную из двух единичных матриц.

Векторы-столбцы, элементы которых образуют арифметическую последовательность, задаются выражением:

```
a:b:c,
```

где a – начальный член последовательности, b – конечный, c – шаг. Выражение a:b:1 допускает сокращенную запись

```
a:b
```

Так, матрицу [2;3;4;5;6;7;8;9] можно задать выражением 2:9.

Если f(x) – скалярная функция от скалярного аргумента, а M – произвольная матрица, выражение f(M) задает матрицу значений функции f(x) на всех элементах матрицы M.

Аналогичная конструкция применяется и для функции от нескольких аргументов. Например, выражение log(2:9,11) задает столбец логарифмов числа 11 по основаниям 2,3, ..., 9, а выражение log(11,2:9) задает столбец логарифмов чисел 2,3, ..., 9 по основанию 11.

Если замене подвергаются несколько аргументов, то необходимо, чтобы совпадали размеры замещающих матриц. Эти же размеры имеет и матрица значений. Например, значением выражения log(2:3,[32;81]) будет [5;4].

Для всех бинарных операций замена числового операнда матричным приводит к образованию матрицы значений подобно тому, как это происходит в случае скалярной функции от двух скалярных аргументов.

Если, например, A – некоторая целочисленная матрица, то A%7 – матрица остатков от деления элементов A на 7, а 7%A – матрица остатков от деления 7 на элементы A.

Суперматрицы

Элементами суперматрицы являются объекты произвольного типа. Суперматрицы создаются с помощью операции **[,]** , а также с помощью функций cols и rows путем разбиения матрицы на столбцы и строки.

Если f(M) – скалярная функция от матричного аргумента M, а S – матрица матриц, то выражение f(S) имеет смысл матрицы значений функции f на элементах S. Аналогичное правило действует и для функций от произвольного количества аргументов.

Если f(X,Y) – скалярная функция от двух матричных аргументов, то выражение

```
matrix(f,SX,SY)
```

для суперматриц SX и SY задает матрицу значений f на элементах SX и SY. Например, ковариационная матрица столбцов матрицы A может быть задана выражением

```
matrix(cov,cols(A),cols(A))
```

Законы

Законы (вероятностные распределения) создаются с помощью оператора law или с помощью специальных функций, таких как: ulaw, nlaw и др. Например, оператор

```
X = ulaw(0,1);
```

определяет имя X как равномерное распределение на отрезке [0,1]. Отметим, что X – не случайная величина, а закон. Соответствующая случайная величина задается выражением X().

Различие между X и X() проявляется еще и в том, что при нескольких запусках программы значение X не меняется в отличие от значения X().

Круглые скобки после имени

Если f – функция, то

```
f(a,b,...)
```

– значение f на аргументах a, b, ... Допустимое количество аргументов зависит от определения функции.

Если M – матрица или суперматрица, то

```
M(i,j)
```

– элемент, расположенный в i-ой строке и j-м столбце матрицы M. Допускается также и один аргумент. Например, для матрицы M, заданной оператором

```
M = [a,b,c; x,y,z];
```

значением M(4) будет x.

Если L – закон, то

```
L(m,n)
```

– матрица размера m×n, элементы которой – независимые случайные величины, распределенные по закону L. Для столбца L(m,1) применяется запись L(m). Допускается также отсутствие аргументов, однако, в этом случае значением L() будет не матрица, а число.

Если x – число, то

```
x(m,n)
```

– матрица размера m×n, все элементы которой равны x. Аналогично предыдущему, x(m) – столбец чисел x высоты m. Например, 0(5) – столбец из пяти нулей.

Операции

Арифметические операции

-x ..... - число с противоположным знаком.
x + y .....- сумма.
x - y ..... - разность.
x * y .....- произведение.
x / y .....- частное.
x % y .....- остаток от деления.
x ^ y .....- степень.

Матричные операции

#X .....- число элементов матрицы X.
~X .....- обратная матрица.
'X.....
.....- транспонированная матрица.
X &* Y .....
.....- произведение матриц X и Y.
X &^ n .....
..... - степень матрицы X, n - целое.
X \ Y .....
.....= ~('X &* X) &* 'X &* Y - левое деление.
x : y .....
.....- столбец чисел от x до y с шагом 1.
x : y : z - столбец чисел от x до y с шагом z.
[A, B, …; …] .....- объединение матриц (чисел) A, B ...

Суперматричные операции

#S .....
.....- число элементов суперматрицы S.
[A, B, …; …]- объединение суперматриц (чисел, строк) A, B ...
@S .....
.....- см. "операция @".
$X .....
..... - (1×1)–суперматрица, содержащая матрицу X.

Логические операции

!x..... - логическое отрицание.
x & y .....- логическое **И**.
x | y .....- логическое **ИЛИ**.

Операции сравнения

x > y - больше.
x < y .....- меньше.
x >= y - больше или равно
x <= y.....- меньше или равно.
x == y - равно

x != y..... - не равно.

Строковые операции

s1 + s2 - соединение строк s1 и s2.

s + x - соединение строки s и записи числа x.

x + s - соединение записи числа x и строки s.

Операции с законами

L1 o L2

– распределение случайной величины, полученной в результате арифметической операции 'o' над независимыми случайными величинами, распределенными по законам L1 и L2. При этом, один из законов L1 или L2 может быть числом, что соответствует распределению константы.

L ^+ n

– то же, что и L+L+...+L, где закон L повторяется n раз.

Операция @

Результат операции @S зависит от типа элементов суперматрицы S: если все элементы суперматрицы S – числа, то @S – обычная матрица, составленная из этих чисел; если все элементы суперматрицы S – матрицы (суперматрицы), то @S – суперматрица, составленная из элементов элементов S.

Операторы

Оператор-выражение

выражение;

Действие: вычисляется *выражение*.

Оператор присваивания

имя = *выражение*;

Действие: вычисляется *выражение*, *имя* становится определенным и получает значение, равное вычисленному.

Индексная форма оператора

имя(*индекс*) = *выражение*;

имя(*индекс*$_1$, *индекс*$_2$) = *выражение*;

Действие: вычисляется *выражение*, элемент ранее определенной матрицы *имя* получает новое значение.

Составной оператор

- это последовательность операторов любых типов, исключая **law** и **function**, окруженная фигурными скобками.

Пример

```
{
  x = x + 1;
  y = y + ln(x);
}
```

Оператор цикла for

for(*переменная* **in** *матрица*) *тело*

Тело оператора **for** – это простой или составной оператор. *Тело* выполняется столько раз, сколько элементов содержит *матрица*, при этом *переменная* поочередно принимает все значения элементов.

Тело оператора **for** может содержать операторы **break** и **continue**, нарушающие стандартную последовательность выполнения оператора.

Оператор **continue** прерывает выполнение тела оператора **for** и вызывает переход к следующему элементу матрицы.

Оператор **break** немедленно завершает выполнение всего оператора **for**.

Переменная цикла **for** может содержать поля **col**, **row** и **num**, характеризующие положение текущего элемента *матрицы*.

Оператор law

law *имя*() = *выражение*;

Случайные величины, распределенные по закону *имя*, генерируются путем вычисления *выражения*.

Оператор function

Описание этого оператора см. раздел «пользовательские функции».

Функции

Стандартные функции

abs(x) = |x|.
arccos(x) - арккосинус.
arcsin(x) - арксинус.
arctg(x) - арктангенс.
ceil(x) - ближайшее к x сверху целое.
cos(x) - косинус.
exp(x) = e^x.
fa(n) = n! - факториал, n от 0 до 170.
floor(x) - ближайшее к x снизу целое.
gamma(x) - гамма-функция для целых и полуцелых x>0.
inv(x) = 0, если x=0; 1/x, если x≠0.
ln(x) - натуральный логарифм.
lg(x) - десятичный логарифм.
plus(x) – (x + |x|)/2.
round(x) – ближайшее к x целое.
sign(x) - знак x, т.е. 1, 0 или -1.
sin(x) - синус.
tg(x) - тангенс.

cbin(n,k) - число сочетаний из n по k.
log(x,y) - логарифм y по основанию x.
mul(x,y) = x * y.
mod(x,y) = x % y.
fmod(x,y) = x - floor(x/y)*y, y>0.
pow(x,y) = x ^ y.
powmod(x,y,z) = (x ^ y) % z.
round(x,k) – округляет x до k знаков после запятой.

iff(c,a,b) = a, если c=1, и b, если c=0.

Статистические функции

X,Y – **матрицы**, L – **закон**.

valfr(X) – таблица значений и частот.

calpha(Y,X) – коэффициент альфа Y относительно X.

cbeta(Y,X) – коэффициент бета Y относительно X.

cor(X,Y) – коэффициент корреляции X и Y.

cov(X,Y) – ковариация X и Y.

mean(X) – среднее арифметическое элементов X.

mean(L) – математическое ожидание.

stdev(L),**stdev**(X) – стандартное отклонение.

disp(L),**disp**(X) – дисперсия.

disp(X,k) = disp(X) * #X/ (#X– k) – эмпирическая дисперсия, k – число связей.

stdev(X,k) = $\text{disp}(X,k)^{1/2}$ – эмпирическое стандартное отклонение.

cmoment(X,k) – центральный момент порядка k.

pvKolm(*u*) – P-значение критерия Колмогорова, для $u = n^{0.5}d$, где n – число наблюдений, d – расстояние между функциями распределения.

crKolm(*alpha*) – критическое значение u_{alpha} статистики u с уровнем значимости *alpha*.

corf(X,Y,L) − вычисляет вектор K корреляций между рядами X и Y с лагами L. Размерность K равна размерности L. Размерности X и Y должны совпадать. Элемент K(i) представляет корреляцию между максимальными подвекторами векторов X и Y, сдвинутыми относительно друг друга на L(i) позиций.

acf(X,L) − автокорреляционная функция. Вычисляется как corf(X,X,L).

Классы law и matrix также содержат статистические методы.

Матричные функции

setmaxdim(n) – задает максимальную размерность.

dim(X) – размерность, количество элементов в X.

cdim(X) – количество столбцов в X.

rdim(X) – количество строк в X.

det(X) – определитель X.

sum(X) – сумма элементов X.

sums(X) – нарастающие суммы элементов X.

prod(X) – произведение элементов X.

prods(X) – нарастающие произведения элементов X.

len(X) – длина вектора X.

max(X) – максимум элементов X.

min(X) – минимум элементов X.

diag(X) – диагональная матрица.

E(n) = **ident**(n) – единичная матрица.

dif(X) – столбец первых разностей вектора X. Размерность вектора уменьшается на 1.

cdif(X) – матрица столбцовых первых разностей матрицы X. Число строк уменьшается на 1.

ret(X) – столбец доходностей вектора X. Размерность вектора уменьшается на 1.

cret(X) – матрица столбцовых доходностей матрицы X. Число строк уменьшается на 1.

intnum(T,x) – номер n полуинтервала $[T_n,T_{n+1})$, содержащего x.

matrix(f, X, Y) – матрица значений функции f на векторах X и Y. Функция f зависит от двух скалярных или матричных аргументов. Если первый аргумент f матричный, то X – матрица матриц. Если второй аргумент f матричный, то Y – матрица матриц.

Суперматричные функции

cols(X) – суперматрица, элементами которой являются столбцы матрицы X.

rows(X) – суперматрица, элементами которой являются строки матрицы X.

resize(X,nc) - суперматрица, состоящая из тех же элементов, что и X, но содержащая nc столбцов. Если размерность X не делится на nc, последняя строка новой суперматрицы дополняется элементами "" ("пустыми" строками).

super(m,n) – суперматрица размера (m×n), все элементы которой являются "пустыми" строками символов, т.е. "".

scdim(X) – число столбцов в суперматрице X.

srdim(X) – число строк в суперматрице X.

См. Методы класса supermatrix.

Функции сортировки

sort(X) - сортировка всех элементов X по возрастанию.

sort(X,k) - сортировка строк матрицы X по возрастанию элементов k-го столбца.

dsort(X) - сортировка всех элементов X по убыванию.

dsort(X,k) - сортировка строк матрицы X по убыванию элементов k-го столбца.

Функции отбора

Функция **select**(X, C) создает матрицу из строк матрицы X, для которых элемент столбца C с тем же номером равен 1 (логическое значение **"истина"**). Здесь C – вектор-столбец, размерность которого совпадает с числом строк в X. Аналогично, **sselect**(X, C) создает суперматрицу из строк суперматрицы X.

Пример 1

Создается (3×7)-матрица из строк матрицы E(7) с номерами 2,3 и 7:

```
select(E(7), [0;1;1;0;0;0;1]);
```

Пример 2

Создается матрица из строк матрицы X с положительными элементами во втором столбце:

```
select(X, X.c(2)>0);
```

Пример 3

После выполнения строк:

```
S = ["a"; "bb"; "ccc"];
R = sselect(S,dim(code(S))>1);
```

значением переменной R будет суперматрица ["bb"; "ccc"].

Пользовательские функции

Пользователь может задать новую скалярную функцию оператором **function** вида

function *имя*(T_1 x_1, T_2 x_2, ..., T_n x_n) = *выражение*;

Здесь T_1, ..., T_n – типы соответствующих аргументов.

Для аргумента функции допускаются следующие типы:

scalar – число,
matrix – матрица,
function – числовая функция,
law – закон (распределение),
string – строка символов.

Отсутствие типа означает, что аргумент имеет тип **scalar**.

Выражение при любых допустимых значениях аргументов должно иметь числовое значение.

Пример 1

Вычисляется сумма корней целых чисел от 1 до 100.

```
function x(k) = k^0.5;
sum(x(1:100));
```

Пример 2

Вычисляется определенный интеграл от функции $y=1/(1+x^2)$ на отрезке [–1,1] по формуле Симпсона.

```
function y(x) = 2/(1+x^2);
sint(y,-1,1,20);  // почти pi
```

Пример 3

Функция estd(L,n) получает оценку дисперсии закона L по выборке объема n. Программа вычисляет среднее арифметическое оценок, полученных по 1000 выборок из X объема 5, где X – стандартная нормальная совокупность.

```
function estd(law L, n) = cmoment(L(n),2);
X = nlaw(0,1);
mean(estd(X,5(1000)));
```

Множественный вызов функции

Следующие условия являются необходимыми для множественного вызова.

- Все фактические аргументы, типы которых не соответствуют описанию функции, должны быть матрицами. Далее такие матрицы называются **замещающими**.
- Все замещающие матрицы должны иметь одинаковое число строк m и одинаковое число столбцов n.

Множественный вызов состоит в том, что для всех элементов замещающих матриц (m×n раз) производится стандартный вызов функции. Результатом множественного вызова является (m×n)-матрица, составленная из результатов стандартных вызовов.

Пример

Пусть f(x,y,z) – числовая функция от числовых аргументов, A и B – матрицы, c – число. Тогда оператор

```
R = f(A, B, c);
```

эквивалентен следующей последовательности операторов:

```
assert(cdim(A) == cdim(B));
assert(rdim(A) == rdim(B));
R = 0(rdim(A), cdim(A));
for(k in 1:#A)
    R(k) = f(A(k), B(k), c);
```

Вызов-закон

Данный тип вызова осуществляется в тех случаях, когда на месте скалярных аргументов скалярной функции находятся дискретные распределения с конечным числом значений.

Пример

Пусть f(x,y,z) – числовая функция от числовых аргументов, A и B – дискретные законы, c – число. Тогда после выполнения оператора

```
L = f(A, B, c);
```

значением переменной L становится распределение случайной величины f(X,Y,c), где X, Y – независимые случайные величины, распределенные по законам A и B соответственно.

Аналогичным образом выполняются вызовы-законы и для арифметических операций. Если, например, A и B – дискретные законы с конечным числом значений, то A+B будет сверткой этих распределений.

Методы классов

Методы класса law

Далее X – случайная величина, распределенная по закону L.

L.**valefr**(n) - см. Таблицы частот значений.

L.**rvalefr**(n) - см. Таблицы частот значений.

L.**intefr**(T,n) - таблица интервальных ожидаемых частот.

L.**intefrgel**(T,n) - аналогично предыдущему, только интервалы имеют вид [a,b).

L.**intefrgle**(T,n) - аналогично предыдущему, только интервалы имеют вид (a,b].

L.**pe**(a) - вероятность {X == a}.

L.**pl**(a) - вероятность {X < a}.

L.**ple**(a) - вероятность {X <= a}.

L.**pg**(a) - вероятность {X > a}.

L.**pge**(a) - вероятность {X >= a}.

L.**invpl**(p) - функция, обратная к L.pl(a).

L.**invpg**(p) - функция, обратная к L.pg(a).

L.**den**(a) - плотность вероятности в точке a.

L.**pgl**(a,b) - вероятность {a < X < b}.

L.**pgle**(a,b) - вероятность {a < X <= b}.

L.**pgel**(a,b) - вероятность {a <= X < b}.

L.**pgele**(a,b) - вероятность {a <= X <= b}.

intefrgel

Метод **L.intefrgel(T,n)** применим ко всем распределениям L, кроме пользовательских. Аргумент T задает вектор, элементы которого служат границами интервалов. Метод создает матрицу, первый столбец которой содержит левые границы интервалов, второй – правые, третий – вероятности интервалов, умноженные на n. Для целого положительного n эта матрица является таблицей ожидаемых интервальных частот выборки объема n из распределения L. Все интервалы, кроме крайних, задаются неравенствами "больше или равно" и "меньше", т.е. являются полуинтервалами вида [a,b). Вероятности крайних интервалов находятся по особому правилу: считается, что крайние точки (вопреки их реальному положению)

находятся в бесконечности. Например, L.intefrgel(1:5,n) – это матрица вида

1	2	n*L.pl(2)
2	3	n*L.pgel(2, 3)
3	4	n*L.pgel(3, 4)
4	5	n*L.pge(4)

intefrgle

Аналогично предыдущему, только интервалы задаются неравенствами "больше" и "меньше или равно", т.е. являются полуинтервалами вида (a,b]. Например, L.intefrgle(1:5,n) – это матрица вида

1	2	n*L.ple(2)
2	3	n*L.pgle(2, 3)
3	4	n*L.pgle(3, 4)
4	5	n*L.pg(4)

intefr

Так же как и две предыдущие функции **intefr** создает таблицу ожидаемых частот, однако ожидаемая частота интервала от a до b определяется как полусумма ожидаемых частот полуинтервалов (a,b] и [a,b). Заметим, что если вероятности попадания в границы интервалов равны 0, то intefr, intefrgle и intefrgel создают одинаковые таблицы. В частности, это так, если распределение L непрерывно.

См. также **Таблицы интервальных частот**.

Методы класса matrix

M.**c**(j) – j-ый столбец матрицы M (j – число).

M.**c**(J) – матрица, составленная из столбцов матрицы M, номера которых заданы вектором J.

M.**r**(j) – j-ая строка матрицы M (j – число).

M.**r**(J) – матрица, составленная из строк матрицы M, номера которых заданы вектором J.

M.**minor**(i,j) – минор матрицы M.

M.**algc**(i,j) = (–1)^(i+j)*M.minor(i,j) – алгебраическое дополнение (complement).

M.**repl0**(n) – замена нулевых элементов вектора M на ближайшие ненулевые значения. Если n>0, ненулевые значения ищутся среди ближайших n элементов с *большими* номерами. Если же n<0, ненулевые значения ищутся среди ближайших (–n) элементов с *меньшими* номерами.

M**.intfr**(T) – см. Таблицы интервальных частот.

M**.rvalfr**(a,b) – см. Таблицы частот значений.

M.**frl**(a) – количество (частота) элементов X матрицы M, для которых X < a.

M.**frle**(a) – количество (частота) элементов X матрицы M, для которых X ≤ a.

M.**frg**(a) – количество (частота) элементов X матрицы M, для которых X > a.

M.**frge**(a) – количество (частота) элементов X матрицы M, для которых X ≥ a.

M.**frgl**(a,b) – количество (частота) элементов X матрицы M, для которых a < X < b.

M.**frgel**(a,b) – количество (частота) элементов X матрицы M, для которых a ≤ X < b.

M.**frgle**(a,b) – количество (частота) элементов X матрицы M, для которых a < X ≤ b.

M.**frgele**(a,b) – количество (частота) элементов X матрицы M, для которых a ≤ X ≤ b.

Методы класса supermatrix

M.**c**(j) – j-ый столбец суперматрицы M (j – число).
M.**c**(J) – суперматрица, составленная из столбцов суперматрицы M, номера которых заданы вектором J.

M.**r**(j) – j-ая строка суперматрицы M (j – число).
M.**r**(J) – суперматрица, составленная из строк суперматрицы M, номера которых заданы вектором J.

См. Методы класса matrix, Суперматричные функции.

Таблицы интервальных частот

Методы **intfr**(T), **intfrgel**(T) и **intfrgle**(T) класса **matrix** создают различные варианты таблиц интервальных частот. Единственным аргументом этих методов является вектор разбиения T, содержащий граничные точки интервалов. Каждый элемент матрицы, попавший внутрь интервала разбиения, увеличивает его частоту на 1. Элемент матрицы, попавший вне всех интервалов, увеличивает на 1 частоту ближайшего к нему интервала.

Методы отличаются способом учета значений, попавших на границы интервалов.

M.**intfr**(T) – элемент матрицы M, попавший на границу двух интервалов, увеличивает их частоты на ½.

M**.intfrgel**(T) – элемент a матрицы M, попавший на границу двух (полу)интервалов, увеличивает на 1 частоту полуинтервала [a,b). Название метода связано с видом неравенств, задающих полуинтервалы: "больше или равно" и "меньше".

M**.intfrgle**(T) – элемент b матрицы M, попавший на границу двух (полу)интервалов, увеличивает на 1 частоту полуинтервала (a,b]. Название метода происходит от вида неравенств, задающих полуинтервалы: "больше " и "меньше или равно".

В любом случае таблица интервальных частот состоит из трех столбцов:

1-ый столбец – левые границы интервалов,

2-ой столбец – правые границы интервалов,

3-ий столбец – частоты интервалов.

Например, M.intfrgel(1:5) – это матрица вида

1	2	n2
2	3	n3 – n2
3	4	n4 – n3
4	5	N – n4

где

n2 – число элементов матрицы M, меньших 2;

n3 – число элементов матрицы M, меньших 3;

n4 – число элементов матрицы M, меньших 4;

N – число всех элементов матрицы M.

Методы **intfr**, **intfrgel** и **intfrgle** подобны методам intefr, intefrgel и intefrgle.

Таблицы частот значений

Таблицы частот элементов матрицы M создают функция **valfr**(M) и метод M.**rvalfr**(a,b).

Таблицы ожидаемых частот выборки объема n из распределения L создают методы L.**valefr**(n) и L.**rvalefr**(a,b,n).

В любом случае таблица частот состоит из двух столбцов:
1-ый столбец – значения,
2-ой столбец – (ожидаемые) частоты этих значений.

Например, для матрицы M = [1,2,3,5,2,3] значением **valfr**(M) будет таблица

1	1
2	2
3	2
5	1

Методы M.**rvalfr**(a,b) и L.**rvalefr**(a,b,n) создают таблицы (ожидаемых) частот значений, принадлежащих отрезку [a,b]. При этом частоты значений, не попавших в [a,b] прибавляются к частоте ближайшего конца отрезка.

Например, для той же матрицы M значением M.**rvalfr**(2,4) будет

2	3
3	2
4	1

Методы L.**valefr**(n) и L.**rvalefr**(a,b,n) создают таблицы математических ожиданий частот выборки объема n из распределения L. Ожидаемая частота значения V совпадает с n*L.pe(V).

Примеры

```
binlaw(3, 0.5).valefr(8)
```

создает

0	1
1	3
2	3
3	1

binlaw(5, 0.5).rvalefr(1,4,32) создает

1	6
2	10
3	10
4	6

Законы

Стандартные законы

ulaw(a,b) – равномерное (uniform) непрерывное распределение на отрезке [a,b].

elaw(V) – эмпирический закон, т.е. распределение относительных частот элементов V.

binlaw(n,p) – биномиальный закон с параметрами n и p.

explaw(lambda) – показательный закон с параметром lambda.

nlaw(m,sigma) – нормальный закон с параметрами m, sigma.

tlaw(k) – распределение Стьюдента, k – число степеней свободы.

x2law(k) – распределение хи-квадрат, k – число степеней свободы.

Flaw(k1,k2) – F-распределение с параметрами k1 и k2 .

dlaw(D) – дискретное распределение. Первый столбец матрицы D – значения, второй – вероятности.

Нестандартные законы

задаются оператором law.

Выборки из распределения

Если L – закон, то: L() – случайная величина, распределенная по закону L. L(n) – столбец n независимых случайных величин, распределенных по закону L. L(n,k) – (n×k)-матрица, элементы которой – случайные величины, распределенные по закону L.

Пример 1

```
nlaw(0,1)(3);
```

– разыгрывается трехмерный случайный вектор, компоненты которого распределены по нормальному закону с параметрами 0 и 1.

Пример 2

Закон NL является приближенно нормальным:

```
law NL() = sum( ulaw(0,1)(10) );
```

Действительно, ulaw(0,1)(10) – столбец из 10 независимых равномерно распределенных на отрезке [0,1] случайных чисел; их сумма распределена приближенно по нормальному закону.

Выборки без возвращения

sample(A,n) – бесповторная выборка-столбец n элементов из матрицы A.

sample(A,n,k) – бесповторная выборка-матрица n×k элементов из A.

samples(A,n,k) – (n×k)-матрица, столбцы которой являются независимыми бесповторными выборками объема n из A.

initrand(x)

Функция **initrand**(x) – инициализирует базисный генератор псевдослучайных чисел целым значением x.

Базисный генератор используется в функциях sample и samples и при вычислении выражений вида L(), L(n), L(m,n), где L – некоторый закон.

Если программа не содержит **initrand**(), инициализация генератора псевдослучайных чисел производится на основе момента запуска программы.

Ряды данных

Время

Момент времени представляется действительным числом. При этом единицей измерения времени служат сутки, а началом отсчета является полночь 30 декабря 1899 года.

time(Year,Month,Day,Hour,Min) – число, представляющее Hour часов и Min минут Day числа Month месяца Year года.

Например,
time(1899,12,30,6,0) = 0.25,
time(1900,1,1,0,0) = 2,
time(1901,1,1,0,0) = 367 и т.д.

date(Year,Month,Day) – число, представляющее 0 часов 0 минут Day числа Month месяца Year года.

currtime() – текущее время.

year(t) – год, содержащий момент времени t. См. formatdate, "%Y", "%y".

month(t) – месяц (1 – 12) года, содержащий момент времени t. См. formatdate, "%B", "%b", "%m".

week(t) – неделя (0 – 53) года, содержащая момент времени t. См. formatdate, "%W".

weekday(t) – день недели, содержащий момент времени t. Понедельник = 1, ... , воскресенье = 7. См. formatdate, "%a", "%A".

monthday(t), **mday**(t) – день (1 – 31) месяца, содержащий момент времени t.

lastmday(t) – последний день (28 – 31) месяца, содержащего момент времени t.

yearday(t), **yday**(t) – день (1 – 366) года, содержащего момент времени t.

lastyday(t) – последний день (365 – 366) года, содержащего момент времени t.

hour(t) – час (0 – 23) суток, содержащий момент времени t.

minute(t) – минута (0–59) часа, содержащая момент времени t.

timer(Seconds) – задает интервал времени, по истечении которого предлагается прервать выполнение программы. Если Seconds=0, проверка длительности работы программы отключается.

setftime(f), **setfdate**(f), **setfmonth**(f) – задают функцию, преобразующую абсциссу курсора в момент времени. Когда такая функция задана, момент времени непрерывно выводится в левом верхнем углу окна рисунка.

Если использовалась функция:

setftime - отображаются дата и время;

setfdate - отображается только дата;

setfmonth - отображаются год и месяц.

indexdaily

indexdaily(d1,d2,ticker,param) – ряд ежедневных значений индекса ticker от даты d1 до d2.

Допустимые значения param:
"OPEN", "HIGH", "LOW", "CLOSE", "VOL", "WAPRICE".

Допустимые значения ticker:
"CAC", "DAX", "DJA", "DJI", "DJT", "DJU", "FTSE", "NASD", "NIKKEI", "SPX", "OEX", "MID", "SML", "IPX", "FUX", "TNX", "TYX", "MTMS".

forexdaily

forexdaily(d1,d2,ticker,param) – ряд ежедневных значений курса ticker от даты d1 до d2.

Допустимые значения param:
"OPEN", "HIGH", "LOW", "CLOSE", "VOL", "WAPRICE".

Допустимые значения ticker:
"EUR/USD", "EUR/JPY", "EUR/GBP", "USD/CAD", "USD/CHF", "USD/JPY", "USD/RUR", "GBP/CHF", "GBP/JPY", "GBP/USD".

rts1daily

rts1daily(d1,d2,ticker,param) – ряд ежедневных значений курса ticker от даты d1 до d2.

Допустимые значения param:
"OPEN", "HIGH", "LOW", "CLOSE", "VOL", "WAPRICE".

Допустимые значения ticker:
"AVIA", "AFLT", "ARHE", "CHMF", "CHNG", "DGEN", "EESR", "EESRP", "ENCO", "ESMO", "GAZP", "GMKN", "IRGZ", "KUBN", "KOEN", "LKOH", "LSNG", "MSNG", "NKNC", "NNSI", "NTMK",

"PEGS", "RBCI", "RITK", "RTKM", "RTKMP", "RTSC", "RTSI", "RTSIF", "RTST", "RUIX", "RUIXOIL", "SBER", "SBERP", "SIBN", "SNGS", "SNGSP", "SPTL", "TATN", "TLEN", "URKA", "URSI", "USBN", "YAEN", "YUKO".

Замечание. Функции **forexdaily, rts1daily, indexdaily** отсутствующие данные доопределяют нулевым значением. Источник данных: http://www.rbc.ru.

Скользящее среднее

sma(X,n) – вычисляет простое скользящее среднее Y вектора данных X. Элементы Y являются средними арифметическими n соседних элементов X. Размерность Y на n–1 меньше размерности X.

wma(X,W) – вычисляет взвешенное скользящее среднее Y вектора данных X. Элементы Y являются взвешенными средними n соседних элементов X, где n – размерность вектора весов W. Размерность Y на n–1 меньше размерности X.

ema(X,a) – вычисляет экспоненциальное скользящее среднее Y вектора данных X. Размерность Y совпадает с размерностью X. Элемент Y(1) = X(1), остальные элементы Y определяются рекурсивно:

$Y(i) = aX(i) + (1–a)Y(i–1)$.

Trade

Функция Trade(B,P,O) вычисляет траекторию портфеля с начальным состоянием B в соответствии с операциями O и ценами P. Начальное состояние задается исходной суммой денег B(1) и количеством актива B(2). Если O(t) больше (меньше) 0, в момент t производится покупка (продажа) актива по цене P(t).

Графика

Цвет

Цвет задается функцией **rgb** с целыми аргументами от 0 до 255. Например,

```
boloto = rgb(189,183,107);
white = rgb(255,255,255);
```

Без предварительного определения можно использовать:

aqua, aquamarine, bisque, black, blue, brown, chartreuse, chocolate, coral, cornflower, crimson, cyan, deeppink, gold, gray, green, hotpink, magenta, navy, olive, orange, orchid, pink, plum, purple, red, rosybrown, salmon, silver, skyblue, steelblue, snow, tan, tomato, violet, wheat, white, whitesmoke, yellow.

Метод w() создает светлые оттенки цветов вида colour.w(q), которые хорошо подходят для закрашивания областей. Цвет **colour.w(q)** – это смесь **colour** и **white**, в которой доля **white** составляет q, $0 < q < 1$.

Линия

Функция **line(X,Y,...)** рисует набор отрезков, образующих непрерывную линию. Координаты концов отрезков задаются векторами X и Y:

X(1), Y(1) и X(2), Y(2) – координаты концов первого отрезка;
X(2), Y(2) и X(3), Y(3) – координаты концов второго отрезка;
X(3), Y(3) и X(4), Y(4) – третьего и т.д.

Функция **line** может иметь от 3 до 6 аргументов.

line(X, Y, C)

– непрерывная линия единичной толщины цвета C.

line(X, Y, C, w)

– непрерывная линия цвета C, w – ее толщина.

line(X, Y, C, w, D)

– разрывная линия. "Короткие" отрезки рисуются цветом C, w – их толщина.

Аргумент D задает условия, при которых отрезок считается "коротким": Если D – число, отрезок "короткий", когда его высота не превосходит D. Если D – вектор, отрезок "короткий", когда его ширина не превосходит D(1), а высота – D(2). "Длинные" отрезки вообще не рисуются.

line(X, Y, C_1, k, D, C_2)

– двуцветная линия. "Короткие" отрезки рисуются цветом C_1, "длинные" – C_2. Толщина "коротких" отрезков – w, "длинных" – 1.

Линия с узлами

Функция **sqline(X, Y, C1, w, C2, s)** рисует ломаную с квадратными узлами в вершинах.

Аргументы

X, Y задают координаты вершин ломаной (см. line);

C1 - цвет линии;

w - толщина линии;

C2 - цвет квадрата;

s - сторона квадрата.

Вертикальные линии

Функция **vlines(M,...)** рисует набор вертикальных отрезков, "стоящих" на оси абсцисс. Каждый отрезок задается соответствующей строкой матрицы M:

M(1,1) - абсцисса, M(1,2) - высота 1-го отрезка;
M(2,1) - абсцисса, M(2,2) - высота 2-го отрезка;
M(3,1) - абсцисса, M(3,2) - высота 3-го отрезка и т.д.

Функция **vlines(M,...)** может иметь 2 или 3 аргумента.

vlines(M, C)

– рисует набор вертикальных отрезков толщиной в 1 пиксель цветом C.

vlines(M, C, k)

- рисует набор вертикальных отрезков толщиной в k пикселей цветом C.

Оси координат

Далее dx,dy – числа, str – строка символов.

axes(dx,dy) – рисует оси координат; dx,dy – расстояния между делениями.

axes(dx) – dy определяется автоматически.

axes() – dx и dy определяются автоматически.

axes(str) – подписи под делениями – символы str, dx =1.

Точки

Функция **points(X,Y,C,s)** рисует точки цвета **C** и размера **s**. Координаты точек задаются векторами X и Y:

X(1), Y(1) - координаты первой точки;
X(2), Y(2) - координаты второй точки;
X(3), Y(3) - третьей и т.д.

Гистограмма

Если матрица M содержит 3 столбца, функции **hist(M,...)** и **dhist(M,...)** рисуют гистограмму – набор прямоугольников, "стоящих" на оси абсцисс. Каждый прямоугольник задается соответствующей строкой матрицы M.

В случае **hist**:
M(1,1) – левая граница, M(1,2) – правая граница, M(1,3) – высота 1-го прямоугольника;
M(2,1) – левая граница, M(2,2) – правая граница, M(2,3) – высота 2-го прямоугольника и т.д.

В случае **dhist**:
M(1,1) – левая граница, M(1,2) – правая граница, M(1,3) – площадь 1-го прямоугольника;
M(2,1) – левая граница, M(2,2) – правая граница, M(2,3) – площадь 2-го прямоугольника и т.д.

Как правило, M – таблица интервальных частот. В этом случае **hist(M,...)** строит гистограмму частот, а **dhist(M,...)** – гистограмму плотности (density) интервальных частот.

Функции **hist()** и **dhist()** могут иметь от 2 до 4 аргументов.

hist(M, C, w, h), dhist(M, C, w, h)

– строят гистограмму, ширина прямоугольников которой составляет w·100% от их нормальной ширины. Внутренность прямоугольников закрашивается цветом C, а их границы рисуются пером черного цвета толщины h.

hist(M, C, w), dhist(M, C, w)

– аналогично предыдущему для h=1.

hist(M, C), dhist(M, C)

– аналогично предыдущему для w=0.5 и h=1.

Если матрица M содержит 2 столбца, функция **hist(M,...)** рисует столбиковую диаграмму.

Столбиковая диаграмма

Если матрица M содержит 2 столбца, функция **hist(M,...)** рисует столбиковую диаграмму – набор прямоугольников, "стоящих" на оси абсцисс. Каждый прямоугольник задается соответствующей строкой матрицы M:

M(1,1) – центр основания, M(1,2) – высота 1-го прямоугольника;
M(2,1) – центр основания, M(2,2) – высота 2-го прямоугольника и т.д.

Как правило, M – таблица частот. В этом случае **hist(M,...)** строит столбиковую диаграмму частот.

Функция **hist()** может иметь от 2 до 4 аргументов.

hist(M, C, w, h)

– строит столбиковую диаграмму, ширина прямоугольников которой составляет w·100% от среднего расстояния между центрами оснований. Внутренность прямоугольников закрашивается цветом C, а их границы рисуются пером черного цвета толщины h.

hist(M, C, w)

– аналогично предыдущему для h=1.

hist(M, C)

– то же, что и hist(M,C,0.5,1).

Если матрица M содержит 3 столбца, функция **hist(M,...)** рисует гистограмму.

Функции show и erase

Функция **show()** показывает рисунок, не дожидаясь окончания работы программы.

Функция **erase()** уничтожает все построенные ранее графические элементы.

В следующей программе эти функции применяются для последовательного создания трех рисунков.

```
d1 = date(2000,1,1);
d2 = date(2005,12,31);
Tickers = ["LKOH", "EESR","MSNG"];
for(Ticker in Tickers) {
      wintitle(Ticker);
      Hi0 = rts1daily(d1,d2,Ticker,"HIGH");
      Vol = rts1daily(d1,d2,Ticker,"VOL");
      Cond = (Hi0 > 0 & Vol > 0);
      Hi = select(Hi0, Cond);
      V = select(Vol, Cond);
      D = select(d1:d2, Cond);
      T = 1 : #D;
      c = max(Hi) / max(V);
      vlines([T, c*V(T)], violet);
      line(T, Hi(T), blue);
      axes();
      show();
      erase();
}
```

Разное

Прерывание

Функция **assert**(C) проверяет условие C. Если C верно (C=1), то выполнение программы продолжается. В противном случае программа прерывается.

Прерывание программы создает также функция message.

Печать

При печати очень больших или малых чисел применяется экспоненциальная форма записи. Например, значение выражения 25*10^(-7) будет записано как 2.5e-006. Запись eXXX обозначает десятичную экспоненту, т.е. "десять в степени XXX".

print

Функция print(X, Y, ...) печатает значения выражений X, Y, ... в файл, заданный функцией setout. Например, программа

```
setout("out.txt");
print(2*pi);
```

печатает 6.2831853 в файл "out.txt".

println

Функция println() работает аналогично print(). Отличие состоит в том, что после печати последнего значения происходит переход на новую строку.

cprint

Функция cprint(X, Y, ...) предлагает напечатать значения выражений X, Y, ... в буфер обмена (Clipboard).

message

Функция message(X, Y, ...) прерывает выполнение программы и печатает значения выражений X, Y, ... в окне сообщения. После "OK" работа программы продолжается.

savetable

Функция **savetable**(T, fname, lang) сохраняет таблицу T в файл fname, используя разделители, характерные для программ с языковой локализацией lang. Например, оператор

```
savetable(E(8),"out.txt", "eng");
```

печатает единичную матрицу в файл "out.txt", разделяя матричные элементы запятыми.

Параметр lang может принимать только два значения: "rus" и "eng". Если он не указан, предполагается значение "rus".

Таблица T должна быть матрицей или суперматрицей, все элементы которой – числа и строки.

Пример

Следующая программа создает таблицу индексов в csv-формате.

```
ticker = ["DJA", "SPX", "DAX", "NIKKEI"];
d1 = date(2001,9,1);
d2 = date(2001,10,3);
D = d1:d2;
C = weekday(D)<6;
I = indexdaily(d1,d2,ticker,"CLOSE");
SI = select(@I, C);
SD = select(D, C);
Tbl = ["день","дата",ticker;
formatdate("%a",SD),
      formatdate("%Y-%m-%d",SD), SI];
savetable(Tbl, "Table.csv");
```

См. date, weekday, indexdaily, select, formatdate.

formatdate

Функции **strdate**(date) и **formatdate**(format,date) преобразуют число-дату date в строку символов. Способ преобразования определяется строкой format, которая может содержать следующие спецификации:

%a – аббревиатура дня недели,
%A – полное название дня недели,
%b – аббревиатура месяца,
%B – полное название месяца,
%d – день месяца (01 – 31),
%j – день года (001 – 366),
%W – неделя (00 – 53),
%m – месяц (01 – 12),
%x – стандартное представление даты,
%y – год без столетия (00 – 99),
%Y – год полностью.

Функция **strdate**(date) дает тот же результат, что и **formatdate**("%x",date).

Приложение 2
Квантильный метод оценки параметров

Предположим, что непрерывная функция распределения $F(x;\theta_1,...,\theta_k)$ зависит от k параметров $\theta_1,...,\theta_k$. Пусть x_p – корень уравнения

$$F(x;\theta_1,...,\theta_k)=p.$$

Предположим, что все квантили x_p линейно зависят от $\theta_1,...,\theta_k$:

$$x_p=\beta_{1,p}\theta_1+...+\beta_{k,p}\theta_k.$$

В этом случае для эмпирических квантилей $\hat{x}_p$ будут выполняться аналогичные соотношения, но уже приближенно. Так для набора эмпирических квантилей $\hat{x}_{p_1},...,\hat{x}_{p_n}$ имеем систему приближенных соотношений

$$\begin{cases}\beta_{1,p_1}\theta_1+...+\beta_{k,p_1}\theta_k\approx\hat{x}_{p_1},\\ .....\\ \beta_{1,p_n}\theta_1+...+\beta_{k,p_n}\theta_k\approx\hat{x}_{p_n}.\end{cases} \qquad (П2.1)$$

Решая (П2.1) методом наименьших квадратов, получим оценки $\hat{\theta}_1,...,\hat{\theta}_k$ *квантильного метода*. Эти оценки обладают важным свойством устойчивости относительно «засорения» выборочной совокупности сколь угодно большими, но редкими «выбросами».

В качестве примера найдем оценки параметров μ,σ нормального распределения $N(\mu,\sigma^2)$. Пусть $Z\sim N(0,1)$, $X\sim N(\mu,\sigma^2)$, $X_p(\mu,\sigma)$ – квантиль уровня p распределения $N(\mu,\sigma^2)$. Вследствие того, что

$$\mu+\sigma Z\sim N(\mu,\sigma^2),$$

для любого $p \in (0,1)$ имеем

$$X_p(\mu,\sigma) = \mu + \sigma X_p(0,1).$$

Таким образом, система (П2.1) приобретает вид

$$\begin{cases} \mu + X_{p_1}(0,1)\sigma \approx \hat{x}_{p_1}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \mu + X_{p_n}(0,1)\sigma \approx \hat{x}_{p_n}. \end{cases} \qquad (П2.2)$$

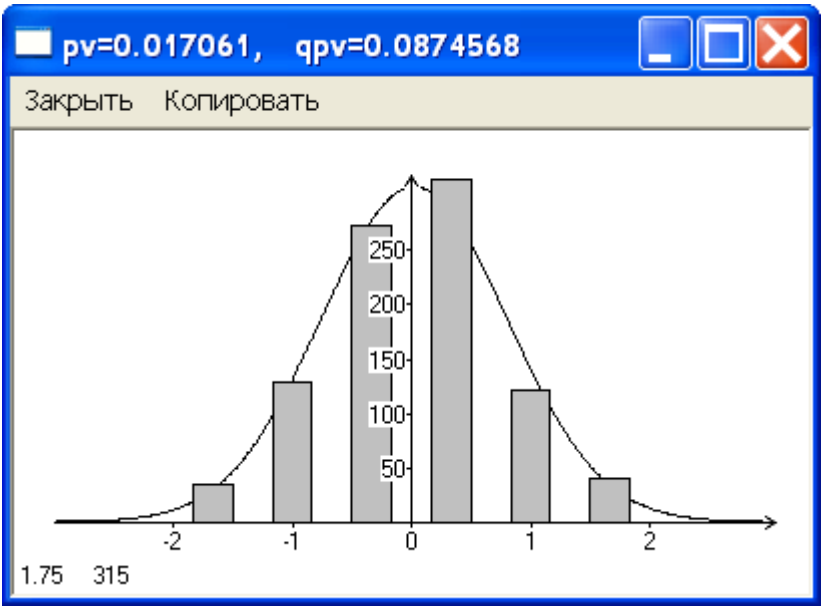
Поскольку распределение зависит лишь от двух неизвестных параметров для решения системы (П2.2) вместо общей формулы (16.5) можно использовать ее частный случай (16.8). Отметим, что теперь μ играет роль коэффициента α, а σ – роль коэффициента β, поэтому в приведенной ниже программе для расчета μ применяется функция calpha(), а для расчета σ – функция cbeta().

```
d1 = date(2003,1,1);      // начальная дата
d2 = date(2006,8,1);      // конечная дата
I0 = indexdaily(d1,d2,"DJI","CLOSE");
// индекс с нулями
I = select(I0, I0 > 0);
// индекс только по рабочим дням
R = 100*dif(ln(I));
// дневная логдоходность (ЛД) в процентах
n = #R;                   // число наблюдений
NI = 6;                   // число интервалов
rmax = 2;
// верхняя граница ЛД в процентах
rmin = -rmax;             // нижняя граница ЛД
h = (rmax - rmin)/NI;     // длина интервала
Pt = rmin:rmax:h;         // все граничные точки
TN = R.intfr(Pt);
// таблица интервальных частот
N = TN.c(3);
// наблюдаемые интервальные частоты
```

```
// Оценка параметров квантильным методом
Q = 0.05:0.95:0.05;       // уровни квантилей
XQ = nlaw(0,1).invpl(Q); // стандартные квантили
EL = elaw(R);
// эмпирическое распределение
YQ = EL.invpl(Q);         // эмпирические квантили
qm = calpha(YQ,XQ);
// квантильная оценка среднего
qsigma = cbeta(YQ,XQ);
// квантильная оценка ст. отклонения
L = nlaw(qm,qsigma);
// предполагаемое распределение
E = L.intefr(Pt,n).c(3); // ожидаемые частоты
x2 = sum((N-E)^2 / E);   // статистика хи-квадрат
qpv = x2law(NI-3).pg(x2);// p-значение
// еще раз для стандартной оценки парметров
m = mean(R);
sigma = stdev(R);
L = nlaw(m,sigma);
E = L.intefr(Pt,n).c(3);
x2 = sum((N-E)^2 / E);
pv = x2law(NI-3).pg(x2);
// Вывод p-значений и построение гистограммы
wintitle("pv=" + pv + ",   qpv=" + qpv);
X = rmin-1 : rmax+1 : 0.01;
Y = #R * h * L.den(X);
line(X, Y, black);
hist(TN, silver);
axes();
```

Данная программа для дневной логарифмической доходности индустриального индекса Доу-Джонса за период с 01.01.2003 по 01.08.2006 находит оценки ее генерального среднего и среднего

квадратичного отклонения как квантильным методом (переменные **qm** и **qsigma**), так и традиционным образом (переменные **m** и **sigma**). В обоих случаях производится проверка нормальности распределения логарифмической доходности по критерию Пирсона путем вычисления *P*-значений (**qpv** и **pv**). Для уровня значимости $\alpha = 0{,}05$ в первом случае гипотеза принимается (**qpv** > 0,05), во втором отвергается (**pv** < 0,05). В результате выполнения программы появляется окно, содержащее гистограмму и *P*-значения критерия Пирсона.



Приложение 3. Статистические таблицы

Таблица значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0,1	0,3970	0,3965	0,3961	0,3956	0,3951	0,3945	0,3939	0,3932	0,3925	0,3918
0,2	0,3910	0,3902	0,3894	0,3885	0,3876	0,3867	0,3857	0,3847	0,3836	0,3825
0,3	0,3814	0,3802	0,3790	0,3778	0,3765	0,3752	0,3739	0,3725	0,3712	0,3697
0,4	0,3683	0,3668	0,3653	0,3637	0,3621	0,3605	0,3589	0,3572	0,3555	0,3538
0,5	0,3521	0,3503	0,3485	0,3467	0,3448	0,3429	0,3410	0,3391	0,3372	0,3352
0,6	0,3332	0,3312	0,3292	0,3271	0,3251	0,3230	0,3209	0,3187	0,3166	0,3144
0,7	0,3123	0,3101	0,3079	0,3056	0,3034	0,3011	0,2989	0,2966	0,2943	0,2920
0,8	0,2897	0,2874	0,2850	0,2827	0,2803	0,2780	0,2756	0,2732	0,2709	0,2685
0,9	0,2661	0,2637	0,2613	0,2589	0,2565	0,2541	0,2516	0,2492	0,2468	0,2444
1,0	0,2420	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0,2203
1,1	0,2179	0,2155	0,2131	0,2107	0,2083	0,2059	0,2036	0,2012	0,1989	0,1965
1,2	0,1942	0,1919	0,1895	0,1872	0,1849	0,1826	0,1804	0,1781	0,1758	0,1736
1,3	0,1714	0,1691	0,1669	0,1647	0,1626	0,1604	0,1582	0,1561	0,1539	0,1518
1,4	0,1497	0,1476	0,1456	0,1435	0,1415	0,1394	0,1374	0,1354	0,1334	0,1315
1,5	0,1295	0,1276	0,1257	0,1238	0,1219	0,1200	0,1182	0,1163	0,1145	0,1127
1,6	0,1109	0,1092	0,1074	0,1057	0,1040	0,1023	0,1006	0,0989	0,0973	0,0957
1,7	0,0940	0,0925	0,0909	0,0893	0,0878	0,0863	0,0848	0,0833	0,0818	0,0804
1,8	0,0790	0,0775	0,0761	0,0748	0,0734	0,0721	0,0707	0,0694	0,0681	0,0669
1,9	0,0656	0,0644	0,0632	0,0620	0,0608	0,0596	0,0584	0,0573	0,0562	0,0551
2,0	0,0540	0,0529	0,0519	0,0508	0,0498	0,0488	0,0478	0,0468	0,0459	0,0449
2,1	0,0440	0,0431	0,0422	0,0413	0,0404	0,0396	0,0387	0,0379	0,0371	0,0363
2,2	0,0355	0,0347	0,0339	0,0332	0,0325	0,0317	0,0310	0,0303	0,0297	0,0290
2,3	0,0283	0,0277	0,0270	0,0264	0,0258	0,0252	0,0246	0,0241	0,0235	0,0229
2,4	0,0224	0,0219	0,0213	0,0208	0,0203	0,0198	0,0194	0,0189	0,0184	0,0180
2,5	0,0175	0,0171	0,0167	0,0163	0,0158	0,0154	0,0151	0,0147	0,0143	0,0139
2,6	0,0136	0,0132	0,0129	0,0126	0,0122	0,0119	0,0116	0,0113	0,0110	0,0107
2,7	0,0104	0,0101	0,0099	0,0096	0,0093	0,0091	0,0088	0,0086	0,0084	0,0081
2,8	0,0079	0,0077	0,0075	0,0073	0,0071	0,0069	0,0067	0,0065	0,0063	0,0061
2,9	0,0060	0,0058	0,0056	0,0055	0,0053	0,0051	0,0050	0,0048	0,0047	0,0046
3,0	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036	0,0035	0,0034
3,1	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026	0,0025	0,0025
3,2	0,0024	0,0023	0,0022	0,0022	0,0021	0,0020	0,0020	0,0019	0,0018	0,0018
3,3	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014	0,0013	0,0013
3,4	0,0012	0,0012	0,0012	0,0011	0,0011	0,0010	0,0010	0,0010	0,0009	0,0009
3,5	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007	0,0007	0,0007	0,0006

Таблица значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}t^2} dt$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999

Таблица процентных точек $t_\alpha(k)$ распределения Стьюдента с k степенями свободы

k	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}01$	$\alpha = 0{,}005$	$\alpha = 0{,}0025$	$\alpha = 0{,}001$	$\alpha = 0{,}0005$
1	6,314	12,706	31,821	63,656	127,321	318,289	636,578
2	2,920	4,303	6,965	9,925	14,089	22,328	31,600
3	2,353	3,182	4,541	5,841	7,453	10,214	12,924
4	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	1,703	2,052	2,473	2,771	3,057	3,421	3,689
28	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	1,699	2,045	2,462	2,756	3,038	3,396	3,660
30	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	1,684	2,021	2,423	2,704	2,971	3,307	3,551
60	1,671	2,000	2,390	2,660	2,915	3,232	3,460
80	1,664	1,990	2,374	2,639	2,887	3,195	3,416
100	1,660	1,984	2,364	2,626	2,871	3,174	3,390
∞	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Пример. $X \sim t(100) \Rightarrow P(X > 1{,}66) = 0{,}05.$

Таблица процентных точек $\chi^2_\alpha(k)$ распределения хи-квадрат с k степенями свободы

k	$\alpha = 0{,}99$	$\alpha = 0{,}975$	$\alpha = 0{,}95$	$\alpha = 0{,}05$	$\alpha = 0{,}025$	$\alpha = 0{,}01$
1	0,00016	0,00098	0,00393	3,841	5,024	6,635
2	0,020	0,051	0,103	5,991	7,378	9,210
3	0,115	0,216	0,352	7,815	9,348	11,345
4	0,297	0,484	0,711	9,488	11,143	13,277
5	0,554	0,831	1,145	11,070	12,832	15,086
6	0,872	1,237	1,635	12,592	14,449	16,812
7	1,239	1,690	2,167	14,067	16,013	18,475
8	1,647	2,180	2,733	15,507	17,535	20,090
9	2,088	2,700	3,325	16,919	19,023	21,666
10	2,558	3,247	3,940	18,307	20,483	23,209
11	3,053	3,816	4,575	19,675	21,920	24,725
12	3,571	4,404	5,226	21,026	23,337	26,217
13	4,107	5,009	5,892	22,362	24,736	27,688
14	4,660	5,629	6,571	23,685	26,119	29,141
15	5,229	6,262	7,261	24,996	27,488	30,578
16	5,812	6,908	7,962	26,296	28,845	32,000
17	6,408	7,564	8,672	27,587	30,191	33,409
18	7,015	8,231	9,390	28,869	31,526	34,805
19	7,633	8,907	10,117	30,144	32,852	36,191
20	8,260	9,591	10,851	31,410	34,170	37,566
21	8,897	10,283	11,591	32,671	35,479	38,932
22	9,542	10,982	12,338	33,924	36,781	40,289
23	10,196	11,689	13,091	35,172	38,076	41,638
24	10,856	12,401	13,848	36,415	39,364	42,980
25	11,524	13,120	14,611	37,652	40,646	44,314
26	12,198	13,844	15,379	38,885	41,923	45,642
27	12,878	14,573	16,151	40,113	43,195	46,963
28	13,565	15,308	16,928	41,337	44,461	48,278
29	14,256	16,047	17,708	42,557	45,722	49,588
30	14,953	16,791	18,493	43,773	46,979	50,892
40	22,164	24,433	26,509	55,758	59,342	63,691
60	37,485	40,482	43,188	79,082	83,298	88,379
80	53,540	57,153	60,391	101,879	106,629	112,329
100	70,065	74,222	77,929	124,342	129,561	135,807

Пример. $X \sim \chi^2(16) \Rightarrow P(X > 32) = 0{,}01$.

Таблица процентных точек $F_{0,05}(k,l)$ распределения Фишера с k и l степенями свободы, $k = 1, 2, \ldots, 10$.

l	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
1	161	199	216	225	230	234	237	239	241	242
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

Пример. $X \sim F(2{,}10) \Rightarrow P(X > 4{,}1) = 0{,}05$.

Таблица процентных точек $F_{0,05}(k,l)$ распределения Фишера с k и l степенями свободы, $k>10$.

l	k=11	k=12	k=15	k=20	k=24	k=30	k=40	k=60	k=120	k= ∞
1	243	244	246	248	249	250	251	252	253	254
2	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,76	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,94	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,70	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	4,03	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,60	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,31	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,10	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,94	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,82	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,72	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,63	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,57	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,51	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,46	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,41	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,37	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,34	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,31	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,28	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,26	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,24	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,22	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,20	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,18	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,17	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,15	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,14	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,13	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,04	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,95	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,87	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,79	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	–

Пример. $X \sim F(12,40) \Rightarrow P(X > 2) = 0,05$.

Таблица процентных точек $F_{0,01}(k,l)$ распределения Фишера с k и l степенями свободы, $k = 1, 2, \ldots, 10$.

l	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32

Пример. $X \sim F(5,30) \Rightarrow P(X > 3,7) = 0,01$.

Таблица процентных точек $F_{0,01}(k,l)$ распределения Фишера с k и l степенями свободы, $k>10$.

l	k=11	k=12	k=15	k=20	k=24	k=30	k=40	k=60	k=120	k= ∞
1	6083	6107	6157	6209	6234	6260	6286	6313	6340	6366
2	99,41	99,42	99,43	99,45	99,46	99,47	99,48	99,48	99,49	99,50
3	27,13	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	14,45	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	9,96	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	7,79	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	6,54	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	5,73	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	5,18	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	4,77	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	4,46	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	4,22	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	4,02	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	3,86	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	3,73	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	3,62	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	3,52	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	3,43	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	3,36	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	3,29	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	3,24	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	3,18	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	3,14	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	3,09	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	3,06	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	3,02	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	2,99	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	2,96	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	2,93	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	2,91	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	2,73	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	2,56	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	2,40	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	2,25	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	–

Пример. $X \sim F(20,60) \Rightarrow P(X > 2,2) = 0,01$.