

С.Д. Шапоров

ПРИКЛАДНАЯ СТАТИСТИКА

Учебное пособие

**Санкт-Петербург
2003**

Министерство образования Российской Федерации
Балтийский государственный технический университет «Военмех»

С.Д. Шапоров

ПРИКЛАДНАЯ СТАТИСТИКА

Учебное пособие

Санкт-Петербург
2003

УДК 519.23 (075.8)

Ш 24

Шапорев С.Д.

Ш 24 Прикладная статистика: Учебное пособие / Балт.
гос. техн. ун-т. СПб., 2003. 254 с.

В пособии рассмотрены основные статистические методы, приемы вычислений и программы часто используемые в практике инженерных расчетов по специальностям выпускающих кафедр БГТУ. Содержит наиболее важные разделы математической статистики: методы описательной статистики, метод статистических испытаний, оценивание числовых характеристик и закона распределения случайной величины, проверка гипотез, дисперсионный и корреляционно-регрессионный анализ. Подробно рассмотрены вопросы статистического моделирования случайных величин на ЭВМ. Приведены примеры, их разбор и решения, графические иллюстрации. Используются популярные пакеты STATGRAPHICS и MATHCAD.

Большое внимание уделяется практической работе с описанными алгоритмами, предлагаются лабораторные работы по всем изучаемым темам, написанные в статистическом пакете STATGRAPHICS и математическом пакете MATHCAD. Каждая лабораторная работа включает серию индивидуальных заданий.

Предназначено для студентов дневного и вечернего отделения. Его использование поможет активизировать самостоятельную работу студентов по курсу «Прикладная статистика» и даст возможность преподавателям контролировать индивидуальную работу студентов в течение всего семестра.

УДК 519.23 (075.8)

Р е ц е н з е н т ы: кафедра высшей математики ПГУПС (зав. каф. д-р техн. наук, проф. *В.Г. Десятрев*), д-р техн. наук, проф. *М.С. Попов*

*Утверждено
редакционно-издательским
советом университета*

© БГТУ, СПб., 2003

1. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ И ИХ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

1.1. Законы распределения дискретных случайных величин

Случайной величиной X называется числовая функция $X = X(\omega)$ от элементарного события, определенная на множестве элементарных исходов Ω , и такая, что при любом x множество тех ω , для которых $X(\omega) < x$, принадлежит алгебре событий.

Дискретной случайной величиной называется случайная величина с конечным или счетным множеством возможных значений.

Законом распределения случайной величины называется любое правило, позволяющее находить вероятности всевозможных событий, связанных с этой случайной величиной. Для дискретных случайных величин простейшей формой закона распределения является ряд распределения - это таблица, в одной строке которой перечислены все значения случайной величины, а во второй строке - соответствующие им вероятности. Например,

X	x_1	x_2	x_3	...	x_n
P	p_1	p_2	p_3	...	p_n

Итак, дискретная случайная величина X в результате опыта примет одно из своих возможных значений, то есть произойдет одно из полной группы событий $\omega_1 = (X = x_1), \omega_2 = (X = x_2), \dots, \omega_n = (X = x_n)$, $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$. Вероятности, соответствующие этим событиям, таковы $p_1 = P(X = x_1), p_2 = P(X = x_2), \dots, p_n = P(X = x_n)$. Очевидно, $\sum_{i=1}^n p_i = 1$, так как $x_i, i = \overline{1, n}$ образуют полную группу событий.

Графическое изображение ряда распределения называется многоугольником распределения дискретной случайной величины.

Наиболее общей формой закона распределения является функция распределения. Функцией распределения случайной величины X называется вероятность неравенства $X < x$, рассматриваемая как функция параметра x ,

$$F(x) = P(X < x). \quad (1.1.1)$$

Чаще всего определенную таким образом функцию распределения называют интегральной функцией распределения или интегральным законом распределения. Функция распределения - самая универсальная характеристика, она полностью определяет случайную величину. Функция распределения любой случайной величины обладает следующими свойствами:

- 1) $0 \leq F(x) \leq 1$ для всех x ;
- 2) $F(x_1) \leq F(x_2)$, если $x_1 < x_2$;
- 3) $F(-\infty) = 0$, $F(\infty) = 1$;
- 4) во всех точках области определения функция непрерывна слева, т.е. $F(x-0) = F(x)$ или $\lim_{x \rightarrow x_0-0} F(x) = F(x_0)$.

Можно показать, что любая функция $F(x)$, обладающая этими свойствами, может быть функцией распределения некоторой случайной величины. График $F(x)$ в общем случае представляет собой график неубывающей функции, значения которой начинаются от нуля и достигают единицы, причем в отдельных точках функция может иметь разрывы первого рода. Если известен ряд распределения дискретной случайной величины, то можно легко построить функцию распределения

$$F(x) = P(X < x) = \sum_{x_i < x} P(X = x_i), \quad (1.1.2)$$

где суммирование распространяется на все те значения x_i , которые меньше x .

Пример. На пути движения автомобиля шесть светофоров, каждый из них либо разрешает, либо запрещает дальнейшее движение автомобиля с вероятностью 0.5. Составить ряд распределения и построить функцию распределения числа светофоров, пройденных автомобилем до первой остановки.

Движение автомобиля либо заканчивается на k -м светофоре, если до этого он проходит $k-1$ светофор без задержки, а на k -м будет остановлен, либо автомобиль пройдет все светофоры и остановлен не будет.

Пусть случайная величина X - число светофоров, пройденных автомобилем. Очевидно, что X может принимать значения 0, 1, 2, 3, 4, 5, 6. X подчинена геометрическому закону распределения с дополнительным условием, что опыт будет закончен, если X примет значение шесть. Следовательно,

$$P(X = k) = qp^k, \quad k = 0, 1, 2, 3, 4, 5, \\ P(X = 6) = 1 - \sum_{i=0}^5 P(X = i), \quad (1.1.3)$$

причем, очевидно, что $p = 1/2$, $q = 1 - p = 1/2$. Тогда ряд распределения случайной величины X имеет следующий вид.

X	0	1	2	3	4	5	6
P	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$

Действительно, например, $p_1 = P(X = 0) = qp^0 = q = 1/2$,
 $p_2 = P(X = 1) = qp = 1/4$ и так далее. Зная ряд распределения, легко по-
 строить многоугольник распределения и функцию распределения, пользу-
 ясь формулой $F(x) = P(X < x) = \sum_{x_i < x} P(X = x_i)$ (рис. 1.1).

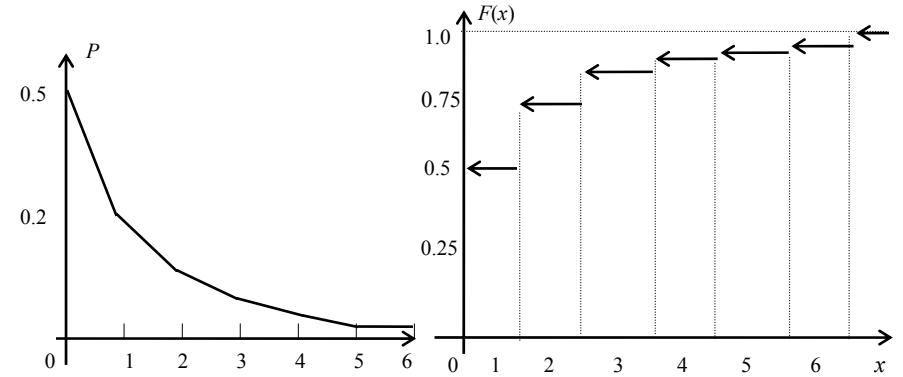


Рис. 1.1. Многоугольник распределения и функция распределения
 дискретной случайной величины

Действительно, $F(0) = P(X < 0) = 0$, $F(1) = P(X < 1) =$
 $= \sum_{x_i < 1} P(X = x_i) = p_1 = \frac{1}{2}$ и так далее. Тогда функция распределения мо-
 жет быть выражена в следующем виде:

$$F(x) = \begin{cases} 0, & x \leq 0, \\ \frac{1}{2} = 0.5, & 0 < x \leq 1, \\ \frac{3}{4} = 0.75, & 1 < x \leq 2, \\ \frac{7}{8} = 0.875, & 2 < x \leq 3, \\ \frac{15}{16} = 0.9375, & 3 < x \leq 4, \\ \frac{31}{32} = 0.96875, & 4 < x \leq 5, \\ \frac{63}{64} = 0.984375, & 5 < x \leq 6, \\ 1, & x > 6. \end{cases} \quad (1.1.4)$$

1.2. Числовые характеристики дискретных случайных величин, их свойства

Ряд распределения или функция распределения дискретной случайной величины являются ее исчерпывающими характеристиками, однако они достаточно громоздки, поэтому возникает необходимость в менее «объемных» характеристиках. Таковыми являются характеристики положения и рассеивания. Характеристики положения дают некоторое среднее положение случайной величины, вокруг которого она группируется, а характеристики рассеивания указывают степень рассеивания случайной величины вокруг ее среднего положения.

Наиболее употребительная характеристика положения - математическое ожидание - среднее взвешенное из значений x_i , причем каждое x_i при осреднении должно учитываться с весом p_i . Таким образом, математическое ожидание дискретной случайной величины равно

$$m_X = M(X) = \frac{x_1 p_1 + x_2 p_2 + \dots + x_n p_n}{p_1 + p_2 + \dots + p_n} = \frac{\sum_{i=1}^n x_i p_i}{\sum_{i=1}^n p_i} = \sum_{i=1}^n x_i p_i. \quad (1.2.1)$$

Если в правой части формулы (1.2.1) стоит ряд, то $M(X) = \sum_{i=1}^{\infty} x_i p_i = \lim_{n \rightarrow \infty} \sum_{i=1}^n x_i p_i$, причем ряд должен сходиться абсолютно.

Математическое ожидание у данного конкретного распределения может и не существовать.

Математическое ожидание случайной величины X связано своеобразной зависимостью со средним арифметическим наблюдаемых значений случайной величины X при большом числе опытов. Эта зависимость того же типа, что между частотой и вероятностью, а именно, при большом числе опытов среднее арифметическое значений X сходится по вероятности к своему математическому ожиданию.

Свойства математического ожидания:

1. $M(C) = C$, $C = \text{const}$. Постоянную величину можно рассматривать как случайную, принимающую только одно значение с вероятностью равной единице, т.е. $M(C) = \sum_{i=1}^1 C \cdot 1 = C$.

2. Константу можно выносить за знак математического ожидания,

т.е. $M(C \cdot X) = C \cdot M(X)$. Действительно, $M(CX) = \sum_{i=1}^n Cx_i p_i =$
 $= C \sum_{i=1}^n x_i p_i = CM(X)$.

3. Свойство аддитивности: $M(X + Y) = M(X) + M(Y)$, так как

$$M(X + Y) = \sum_{i=1}^n (x_i + y_i) p_i = \sum_{i=1}^n x_i p_i + \sum_{i=1}^n y_i p_i = M(X) + M(Y).$$

Совокупность второго и третьего свойств называется свойством линейности и выражается следующим равенством $M(C_1 X_1 + C_2 X_2 + \dots + C_n X_n) = C_1 M(X_1) + C_2 M(X_2) + \dots + C_n M(X_n)$. В частности, если $Y = kX + b$, то $M(Y) = M(kX + b) = kM(X) + b$.

4. Свойство монотонности: если $X \geq Y$, то $M(X) \geq M(Y)$.

5. Мультипликативное свойство: для независимых случайных величин X и Y справедливо $M(X \cdot Y) = M(X) \cdot M(Y)$.

Кроме математического ожидания в качестве характеристик положения случайной величины часто используются мода и медиана.

Модой дискретной случайной величины X называется такое значение x_k , $k = \overline{1, n}$, для которого

$$P(X = d_X) = \max_k P(X = x_k), \quad (1.2.2)$$

т.е. мода есть наиболее вероятное значение дискретной случайной величины, если это значение единственно. Мода может быть и не единственной, т.е. распределение может иметь несколько мод (мультимодальное распределение).

Медианой дискретной случайной величины X называется число h_X , удовлетворяющее условию

$$P(X < h_X) = P(X \geq h_X) = 1/2. \quad (1.2.3)$$

Так как данное уравнение в общем случае может иметь несколько корней, то значение медианы может быть не единственным.

Перейдем теперь к определению характеристик рассеивания случайной величины около своего математического ожидания.

Начальным моментом k -го порядка дискретной случайной величины X называется математическое ожидание k -й степени случайной величины

$$\alpha_k = M(X^k) = \sum_{i=1}^n x_i^k p_i. \quad (1.2.4)$$

Это определение совпадает с определением начального момента в ме-

ханике, если вероятности p_i интерпретировать как массы точек x_i . В частности из формулы (1.2.4) следует, что первый начальный момент есть математическое ожидание, т.е. $\mu_1 = m_X$.

Центральным моментом k -го порядка дискретной случайной величины X называется математическое ожидание k -й степени соответствующей центрированной случайной величины

$$\mu_k = M[(X - m_X)^k] = \sum_{i=1}^n (x_i - m_X)^k p_i. \quad (1.2.5)$$

Дисперсией случайной величины X называется математическое ожидания квадрата соответствующей центрированной величины, т.е. ее второй центральный момент,

$$D(X) = D_X = \mu_2 = M[(X - m_X)^2] = \sum_{i=1}^n (x_i - m_X)^2 p_i. \quad (1.2.6)$$

Средним квадратическим отклонением или стандартным отклонением (стандартом) случайной величины X называется величина

$$\sigma_X = \sqrt{D_X}. \quad (1.2.7)$$

Для дисперсии из формулы (1.2.6) легко выводится следующая часто употребляемая формула:

$$D_X = \sum_{i=1}^n x_i^2 p_i - m_X^2. \quad (1.2.8)$$

Свойства дисперсии:

1. Дисперсия любой случайной величины X неотрицательна, причем $D_X = 0$ тогда и только тогда, когда X - постоянная, т.е. $D(X) \geq 0$, $D(C) = 0$.

2. Если $Y = X + C$, где $C = \text{const}$, то $D(Y) = D(X + C) = D(X)$.

3. Если $C = \text{const}$, то $D(C \cdot X) = C^2 D(X)$.

4. Если случайные величины X и Y независимы, то $D(X + Y) = D(X) + D(Y)$.

Коэффициентом асимметрии называется число A , определяемое формулой

$$A = \frac{\mu_3}{\sigma_X^3} = \frac{\sum_{i=1}^n (x_i - m_X)^3 p_i}{\sigma_X^3}. \quad (1.2.9)$$

Коэффициент асимметрии служит для характеристики асимметрии многоугольника распределения. В случае отрицательного коэффициента

асимметрии более пологий склон многоугольника распределения наблюдается слева, в противном случае – справа. В первом случае асимметрию называют левосторонней, во втором – правосторонней.

Эксцессом или коэффициентом крутости называется число

$$E = \frac{\mu_4}{\sigma_X^4} - 3. \quad (1.2.10)$$

Эта характеристика служит для сравнения на «крутость» данного и нормального распределения. Эксцесс для случайной величины, распределенной нормально, равен нулю. Если распределению соответствует отрицательный эксцесс, то соответствующий многоугольник распределения имеет более пологую вершину по сравнению с нормальной кривой. В случае положительного эксцесса многоугольник более крутой по сравнению с нормальной кривой.

1.3. Законы распределения непрерывных случайных величин

Непрерывной случайной величиной называется такая случайная величина, вероятность попадания которой в любую бесконечно малую область бесконечно мала и для которой при каждом x существует конечный или бесконечный предел

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X < x + \Delta x)}{\Delta x}. \quad (1.3.1)$$

Все основные определения относительно закона распределения здесь остаются в силе. Для непрерывной случайной величины невозможно задать ряд распределения. Функция же распределения для нее существует и представляет собой непрерывную кривую.

Функцией распределения непрерывной случайной величины X называется вероятность следующего неравенства:

$$F(x) = P(X < x) = \int_{-\infty}^x f(t) dt \quad (1.3.2)$$

при условии, что существует такая неотрицательная функция $f(x)$ (рис. 1.2), интегрируемая в бесконечных пределах. Эта функция называется плотностью распределения вероятностей. Справедливы следующие соотношения:

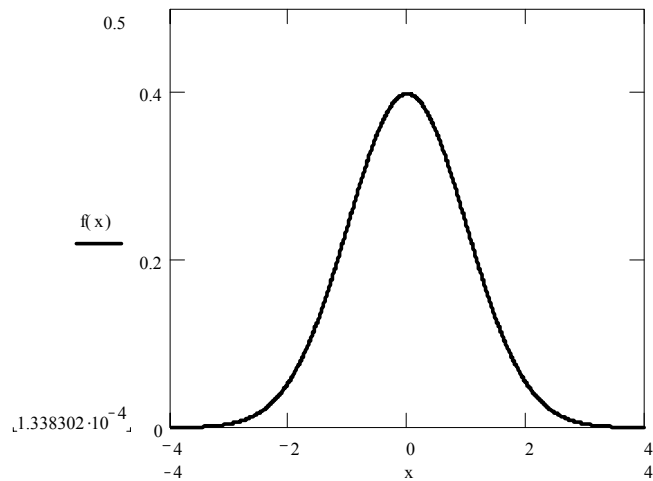


Рис. 1.2. Функция плотности вероятности непрерывной случайной величины

$$\begin{cases} F(x) = \int_{-\infty}^x f(t) dt, \\ f(x) = \frac{dF(x)}{dx}. \end{cases} \quad (1.3.3)$$

Функции плотности вероятностей соответствует кривая плотности распределения, или кривая плотности вероятности. Она является одной из форм закона распределения, но не универсальной, ибо существует только для непрерывной случайной величины. Ее некоторой аналогией для дискретных случайных величин является многоугольник распределения.

Свойства функции плотности распределения: 1. $f(x) \geq 0$,

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

1.4. Числовые характеристики непрерывных случайных величин

Математическим ожиданием непрерывной случайной величины X с плотностью вероятности $f(x)$ называется

$$M(X) = m_X = \int_{-\infty}^{\infty} xf(x)dx. \quad (1.4.1)$$

Все свойства математического ожидания, приведенные в предыдущих подразделах, остаются справедливыми и для этого определения. Еще две характеристики положения, а именно, мода и медиана остаются в силе для непрерывной случайной величины и даже определяются в этом случае наиболее естественным образом, если пользоваться понятием функции плотности распределения.

Модой непрерывной случайной величины X называется число d_X , определяемое как точка максимума функции плотности вероятности $f(x)$. Итак,

$$f'(d_X) = 0, \quad f'(x < d_X) > 0 \quad \text{и} \quad f'(x > d_X) < 0. \quad (1.4.2)$$

Медианой непрерывной случайной величины X называется число h_X , удовлетворяющее условию

$$\int_{-\infty}^{h_X} f(x)dx = \int_{h_X}^{\infty} f(x)dx = \frac{1}{2}. \quad (1.4.3)$$

Все определения для начальных и центральных моментов остаются в силе, только суммы заменяются интегралами.

Дисперсией непрерывной случайной величины называется ее второй центральный момент, т.е.

$$D_X = D(X) = \int_{-\infty}^{\infty} (x - m_X)^2 f(x)dx. \quad (1.4.4)$$

Квантилью, или квантилем, порядка p распределения непрерывной случайной величины X называется число t_p , удовлетворяющее условию

$$P(X < t_p) = p \quad \text{или} \quad \int_{-\infty}^{t_p} f(x)dx = p. \quad (1.4.5)$$

Очевидно что, например, $h_X = t_{0.5}$.

Критической точкой порядка p распределения непрерывной случайной величины X называется число κ_p , удовлетворяющее уравнению

$$P(X \geq \kappa_p) = p \text{ или } \int_{\kappa_p}^{\infty} f(x) dx = p. \quad (1.4.6)$$

Квантили и критические точки одного и того же распределения связаны между собой простым соотношением $\kappa_p = t_{1-p}$.

Асимметрия и эксцесс для непрерывных случайных величин определяются аналогично формулам (1.2.9) и (1.2.10).

Пример. Случайная величина X подчинена закону арксинуса (рис. 1.3) с плотностью распределения вероятностей

$$f(x) = \begin{cases} 0, & |x| \geq a, \\ \frac{1}{\pi\sqrt{a^2 - x^2}}, & |x| < a. \end{cases} \text{ Найти функцию распределения } F(x) \text{ и вычис-}$$

лить m_X , D_X , d_X , h_X , $\kappa_{0.75}$.

Найдем сначала $F(x)$. По определению $F(x) = \int_{-\infty}^x f(t) dt =$

$$= \int_{-a}^x \frac{dt}{\pi\sqrt{a^2 - t^2}} = \frac{1}{\pi} \left(\arcsin \frac{t}{a} \right) \Big|_{-a}^x = \frac{1}{\pi} \left(\arcsin \frac{x}{a} + \arcsin 1 \right) = \frac{1}{2} + \frac{1}{\pi} \arcsin \frac{x}{a}.$$

Графики функции плотности вероятности и функции распределения приведены ниже.

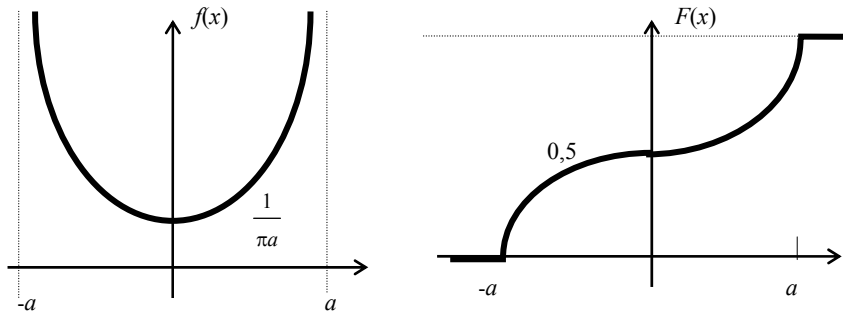


Рис. 1.3. Графики функций плотности вероятности и распределения закона арксинуса

Определим теперь все числовые характеристики, необходимые по условию задачи.

$$m_X = \alpha_1 = \int_{-a}^a xf(x)dx = \int_{-a}^a x \frac{dx}{\pi\sqrt{a^2 - x^2}} = -\frac{1}{2\pi} \int_{-a}^a \frac{d(a^2 - x^2)}{\sqrt{a^2 - x^2}} =$$

$$= -\frac{1}{2\pi} 2\sqrt{a^2 - x^2} \Big|_{-a}^a = 0. \text{ Этот результат очевиден и из рисунка функции}$$

плотности вероятности. Найдем моду. $f'(x) = \frac{x}{\pi(a^2 - x^2)^{3/2}} = 0, x = 0$, но

$x = 0$ - это точка минимума, а не максимума. Следовательно, моды данное распределение не имеет. Медиану также найдем по определению

$$\int_{-a}^{h_X} f(x)dx = \int_{-a}^{h_X} \frac{dx}{\pi\sqrt{a^2 - x^2}} = \frac{1}{\pi} \arcsin \frac{x}{a} \Big|_{-a}^{h_X} = \frac{1}{2}. \text{ Отсюда } \arcsin \frac{h_X}{a} = 0,$$

$h_X = 0$. В силу симметричности кривой функции плотности вероятности этот результат тоже очевиден из рисунка $f(x)$.

$$D_X = \mu_2 = \int_{-a}^a (x-0)^2 f(x)dx = \int_{-a}^a \frac{x^2 dx}{\pi\sqrt{a^2 - x^2}} =$$

$$= \left\langle \begin{array}{l} x = a \sin t, dx = a \cos t dt, \\ x^2 = a^2 \sin^2 t, \sqrt{a^2 - x^2} = a \cos t, \\ x = \pm a, t = \pm \pi/2 \end{array} \right\rangle = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} a^2 \sin^2 t dt =$$

$$= \frac{a^2}{\pi} \int_{-\pi/2}^{\pi/2} \frac{1 - \cos 2t}{2} dt = \frac{a^2}{\pi} \left[\frac{t}{2} \Big|_{-\pi/2}^{\pi/2} - \frac{1}{4} \sin 2t \Big|_{-\pi/2}^{\pi/2} \right] = \frac{a^2}{2}.$$

Наконец, найдем требуемую критическую точку.

$$\int_{\kappa_{0.75}}^a \frac{dx}{\pi\sqrt{a^2 - x^2}} = 0.75 = \frac{1}{\pi} \int_{\arcsin(\kappa/a)}^{\pi/2} \frac{a \cos t}{a \cos t} dt = \frac{1}{\pi} t \Big|_{\arcsin(\kappa/a)}^{\pi/2} = \frac{1}{\pi} \left(\frac{\pi}{2} - \arcsin \frac{\kappa}{a} \right) = 0.75.$$

$$\text{Отсюда } \frac{1}{\pi} \arcsin \frac{\kappa}{a} = -0.25, \frac{\kappa}{a} = \sin \left(-\frac{\pi}{4} \right) = -\frac{\sqrt{2}}{2}, \kappa_{0.75} = -\frac{\sqrt{2}a}{2}.$$

1.5. Выборочные аналоги интегральной и дифференциальной функций распределения

Предметом математической статистики является изучение случайных величин по результатам наблюдений. В ней развиваются методы обработки результатов опытов, анализа полученной из опытов статистической информации, получения числовых оценок параметров распределений.

Центральное понятие математической статистики - понятие выборки. Выборка понимается следующим образом. Пусть проводится некоторый эксперимент, связанный со случайной величиной X функцией распределения $F(x)$.

Выборкой объема n из генеральной совокупности с функцией распределения $F(x)$ называется последовательность x_1, x_2, \dots, x_n наблюдаемых значений случайной величины X , соответствующих n независимым повторениям данного эксперимента. Таким образом, выборка или выборочная совокупность – это множество случайно отобранных объектов или наблюдений над некоторой случайной величиной, а генеральная совокупность – это совокупность всех объектов или всех возможных мыслимых значений случайной величины, из которых производится выборка. Каждый элемент выборки представляет собой конкретную реализацию одной и той же случайной величины с функцией распределения $F(x)$. Можно, поэтому уточнить понятие выборки следующим образом.

Выборкой объема n называется n независимых случайных величин X_1, X_2, \dots, X_n , каждая из которых распределена так же, как некоторая случайная величина X с функцией распределения $P(X \leq x) = F(x)$. Выборка называется репрезентативной или представительной, если она достаточно хорошо представляет количественные соотношения генеральной совокупности. Репрезентативность выборки обеспечивается случайностью отбора. Это означает, что любой объект выборки отобран случайно, при этом все объекты имеют одинаковую вероятность попасть в выборку.

Как известно, существуют четыре схемы выбора элементов множеств (выборки): схемы с возвращением элемента или без возвращения и с последующим упорядочиванием или без упорядочивания выбранных элементов. Все эти схемы реализуются в конкретных инженерных задачах.

Выборка, упорядоченная по возрастанию наблюдаемых значений случайной величины, называется вариационным рядом.

Пусть теперь имеется выборка x_1, x_2, \dots, x_k объема n . Среди элементов x_i могут быть и одинаковые. Пусть в выборке элемент x_i встречается

n_i раз. Число n_i называется частотой. Очевидно, что $\sum_{i=1}^k n_i = n$. Отноше-

ние частоты n_i к объему выборки n называется относительной частотой

значения x_i и обозначается $w_i = \frac{n_i}{n}$. $\sum_{i=1}^k w_i = \sum_{i=1}^k \frac{n_i}{n} = \frac{1}{n} \cdot n = 1$.

Совокупность пар (x_i, n_i) называется статистическим рядом или статистическим распределением и обычно записывается в виде таблицы:

X	x_1	x_2	\dots	x_k
n_i	n_1	n_2	\dots	n_k

Если X - дискретная случайная величина, то статистический ряд, записанный в виде

X	x_1	x_2	\dots	x_k
w_i	w_1	w_2	\dots	w_k

является аналогом ряда распределения. Если же X - непрерывная случайная величина, то статистический ряд записывается в виде

X	$[x_0, x_1]$	$[x_1, x_2]$	\dots	$[x_{k-1}, x_k]$
w_i	w_1	w_2	\dots	w_k

где w_i - относительные частоты попадания случайной величины X в интервал $[x_{i-1}, x_i]$, $i = 1, 2, \dots, k$.

При большом объеме выборки n ее элементы объединяются в группы и получается группированный статистический ряд. Для этого все интервалы выборки разделяются на l разрядов (от 6 до 20). Следует помнить, что группировка всегда вносит некоторую погрешность в вычисления. Эта погрешность растет с уменьшением числа разрядов. Графическим представлением выборки являются полигон частот и гистограмма. Полигон частот строится для дискретной случайной величины. Это график, точки которого имеют координаты (x_i, n_i) или (x_i, w_i) . Таким образом, полигон частот для выборки является налогом многоугольника распределения дискретной случайной величины. Для иллюстрации распределения непрерывной случайной величины строят гистограмму (рис. 1.4). Гисто-

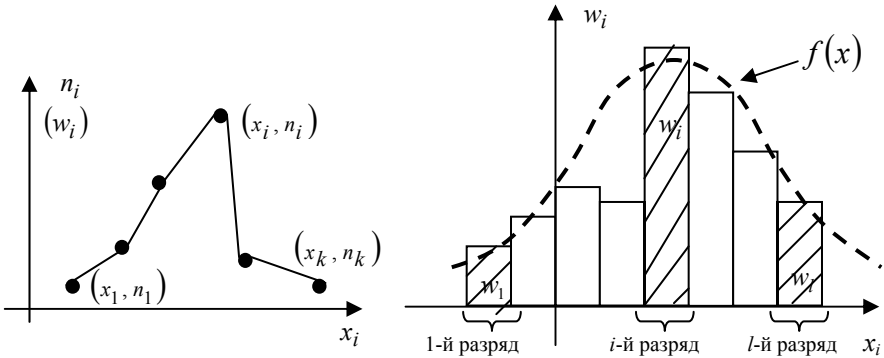


Рис. 1.4. Полигон частот и гистограмма выборки

граммой частот группированной выборки называется ступенчатая фигура, составленная из прямоугольников, построенных на интервалах группировки (разрядах) так, что площадь каждого прямоугольника равна или пропорциональна частоте n_i или относительной частоте w_i . Очевидно, что при увеличении числа опытов длину разряда можно неограниченно уменьшать, и тогда гистограмма будет все более и более приближаться к некоторой кривой, ограничивающей единичную площадь. Ясно, что эта кривая – график функции плотности вероятности непрерывной случайной величины X . Таким образом, гистограмма – аналог кривой плотности вероятности.

Введем, наконец, понятие выборочной функции распределения. Пусть имеется выборка объема n , x – некоторое действительное число, а n_x – число выборочных значений случайной величины X , меньших x . Тогда число n_x/n является относительной частотой наблюдаемых в выборке значений X , меньших x , т.е. относительной частотой появления события $X < x$. Ясно, что при изменении x будет меняться и величина n_x/n . Это означает, что относительная частота n_x/n – функция аргумента x . А так как эта функция находится по выборочным опытным данным, то ее называют выборочной, статистической или эмпирической.

Статистической или эмпирической функцией распределения называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X < x$, т.е.

$$F^*(x) = \begin{cases} 0, & x \leq x_1, \\ \sum_{i=1}^k w_i, & x_k < x \leq x_{k+1}, \quad k = 1, 2, \dots, n-1, \\ 1, & x > x_n. \end{cases} \quad (1.5.1)$$

Формально эмпирическая функция распределения обладает всеми свойствами интегральной функции распределения (подразд. 1.1). Имея статистический ряд, очень легко получить статистическую функцию распределения. Действительно,

$$F^*(x_1) = 0, \quad F^*(x_2) = w_1,$$

$$F^*(x_3) = w_1 + w_2, \dots,$$

$$F^*(x_k) = \sum_{i=1}^{k-1} w_i,$$

$$F^*(x_{n+1}) = \sum_{i=1}^n w_i = 1.$$

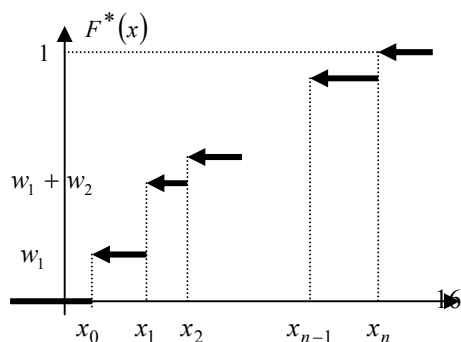


Рис. 1.5. График эмпирической функции распределения

На графике этой функции (рис. 1.5) видны все основные особенности эмпирической функции распределения. Она не убывает, а ее значения находятся в интервале $[0, 1]$. Резкие скачки графика функции $F^*(x)$, придающие ей ступенчатый вид, имеют место в тех точках, которым соответствуют концы разрядов, а величина скачка равна относительной частоте разряда. Часто график $F^*(x)$ строят в виде непрерывной кривой, соединяя

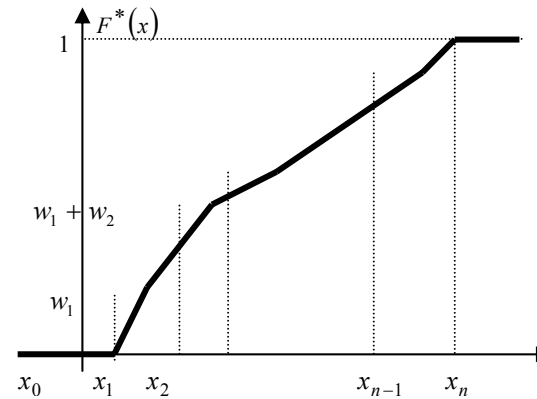


Рис. 1.6. Кумулятивная кривая

точки графика, соответствующие концам или серединам разрядов, отрезками прямой (рис. 1.6).

Отметим, что подобный график эмпирической функции распределения, дающий приближенное представление о графике теоретической функции $F(x)$, часто называют кумулятивной кривой (от англ. accumulation – накопление).

Так как по теореме Бернулли^{*} относительные частоты w_i при $n \rightarrow \infty$ сходятся по вероятности к соответствующим вероятностям событий, то при $n \rightarrow \infty$ $F^*(x)$ приближается к интегральной функции распределения. О сходимости $F^*(x)$ к $F(x)$ доказана теорема, носящая имя авторов.

Теорема 1.1 (Гливенко-Кантелли^{}).** Эмпирическая функция распределения $F^*(x)$ равномерно по x с вероятностью 1 сходится при $n \rightarrow \infty$ к теоретическому распределению $F(x)$, т.е.

$$P\left\{\lim_{n \rightarrow \infty} \sup_{-\infty < x < +\infty} |F^*(x) - F(x)| = 0\right\} = 1. \quad (1.5.2)$$

Смысл этой теоремы в том, что при увеличении объема выборки n у эмпирической функции распределения исчезают свойства случайности и

^{*} Яков Бернулли (1654 – 1705) – швейцарский математик.

^{**} Валерий Иванович Гливенко (1896-1940) – советский математик, Франческо Паоло Кантелли (1875-1966) – итальянский математик.

она приближается к теоретической функции распределения. Аналогично, если n велико, то значение гистограммы $w_n(x)$ в точке x приближенно равно $[F(x_{i+1}) - F(x_i)] / \Delta_i$, где $\Delta_i = x_{i+1} - x_i$, а x_i, x_{i+1} - концы интервала, в котором находится x . Если Δ_i - мало, то гистограмма $w_n(x)$ достаточно хорошо воспроизводит функцию плотности $f(x)$.

1.6. Лабораторная работа № 1. Методы описательной статистики в пакете STATGRAPHICS

Существуют разнообразнейшие методы обработки данных, имеющие различную сложность и нередко требующие больших вычислительных ресурсов. Это огромный мир, созданный многолетними трудами профессиональных математиков и прикладных научных специалистов.

Вместе с тем, следует отметить, что значительная часть методов и статистических пакетов быстро устаревает. Это связано со стремительными темпами развития отрасли. В таких условиях выигрывает программное обеспечение, обладающее очень высокими потребительскими качествами. Пакет STATGRAPHICS (STATistical GRAPHICS System) выгодно отличается от других статистических пакетов удобством пользовательского интерфейса и объемом используемых методов обработки, принадлежит к классу универсальных пакетов. Этот пакет, созданный американской корпорацией Manugistics, является одним из наиболее эффективных систем статистического анализа данных.

С момента выхода первой версии STATGRAPHICS эволюцию претерпели почти все основные составляющие пакета. Сейчас STATGRAPHICS Plus for Windows включает более 250 статистических и системных процедур, применяющихся в бизнесе, экономике, маркетинге, биологии, социологии, на производстве и в других областях. Весь пакет в целом имеет модульную структуру. Каждая статистическая процедура в STATGRAPHICS Plus for Windows сопровождается интегрированной в систему отличной графикой. Все элементы графических отображений могут быть подвергнуты коррекции и преобразованию. Для этого нужно выбрать требуемый элемент, щелкнув на нем левой кнопкой мыши, затем щелкнуть правой кнопкой. Тогда на экране появиться диалоговое окно, в котором можно выполнить необходимые изменения.

Собственный вариант программы расчета можно сохранить в виде файла StatFolio. Если возникнет потребность в обработке другого множества данных по составленной схеме анализа, нужно в этот вариант просто загрузить новый файл данных. Результаты расчетов, таблицы и графики будут выданы автоматически.

В пакете реализовано средство помощи пользователю – статконсультант (StatAdvisor), которое представляет интерпретацию результатов. Для вызова статконсультанта нужно щелкнуть левой кнопкой мыши на графическом или табличном окне пакета, а затем на пиктограмме StatAdvisor. Появится консультационное окно, содержащее исчерпывающие советы, разъяснения и рекомендации.

После запуска STATGRAPHICS Plus for Windows на экране монитора появляется следующая заставка (рис. 1.7). Многие пункты головного меню выполняют те же действия, что и в большинстве прочих Windows – при-

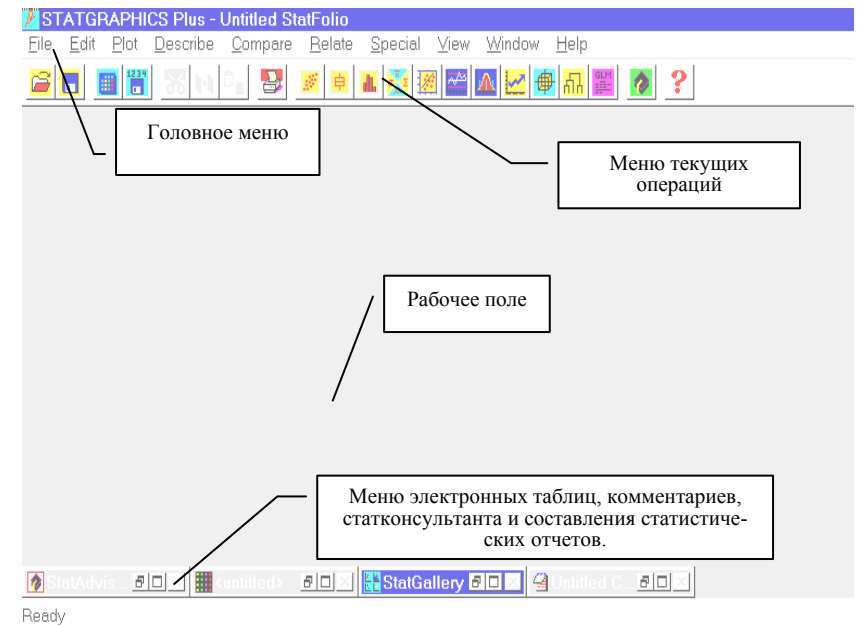


Рис. 1.7. Окно STATGRAPHICS

ложений. Большинство статистических процедур сгруппированы в пунктах Describe, Compare и Relate.

Меню Describe содержит статистические методы анализа по одной и множеству переменных, процедуры подбора распределений, средства табуляции данных.

Меню Compare включает методы сравнения двух и более выборок данных, процедуры одно- и многофакторного дисперсионного анализа.

Меню Relate содержит процедуры простого, полиномиального и множественного регрессионного анализа. В этой и следующих лабораторных работах будет подробно разобрано содержимое этих пунктов меню.

Решим простейшую задачу описательной статистики. Найдем размах выборки, число и длину интервалов, составим таблицу частот, построим гистограмму частот, а также вычислим все числовые характеристики следующей выборки.

Числа выборки представляют собой продолжительность работы электронных ламп одного типа в часах:

13.4	14.7	15.2	15.1	13.0	8.8	14.0	17.9	15.1	16.5
16.6	14.2	16.3	14.6	11.7	16.4	15.1	17.6	14.1	18.8
11.6	13.9	18.0	12.4	17.2	14.5	16.3	13.7	15.5	16.2
8.4	14.7	15.4	11.3	10.7	16.9	15.8	16.1	12.3	14.0
17.7	14.7	16.2	17.1	10.1	15.8	18.3	17.5	12.7	20.7
13.5	14.0	15.7	21.9	14.3	17.7	15.4	10.9	18.2	17.3
15.2	16.7	17.3	12.1	19.2					

Воспользуемся пунктом меню Describe (описание данных). Нам необходимо выполнить анализ одной переменной. Этот анализ в пакете STATGRAPHICS содержит процедуры вычисления следующих характеристик.

1. Суммарные статистики: среднее, медиана, мода, среднее геометрическое, дисперсия, стандартное отклонение, минимум, максимум, размах, нижний и верхний квартиль, межквартильный размах, коэффициенты асимметрии и эксцесса.
2. Процентили и табуляции частот.
3. Гистограммы и график плотности.
4. Доверительные интервалы.
5. Проверка гипотез о среднем и медиане, знаковый и ранговый тест.
6. Графики «дерево с листьями», «ящик с усами», квантильный график, график нормального распределения, симметричный график и диаграмма рассеивания.

Введем исходные данные в новую электронную таблицу, для чего щелкнем по пиктограмме Untitled в левом нижнем углу рабочего поля (рис. 1.8). Необходимо наименовать переменную, которую мы будем вводить в первый столбец. Щелкнем правой кнопкой мыши по заголовку Col_1, появится контекстное меню, в котором следует выбрать команду Modify Column (Определить столбец).

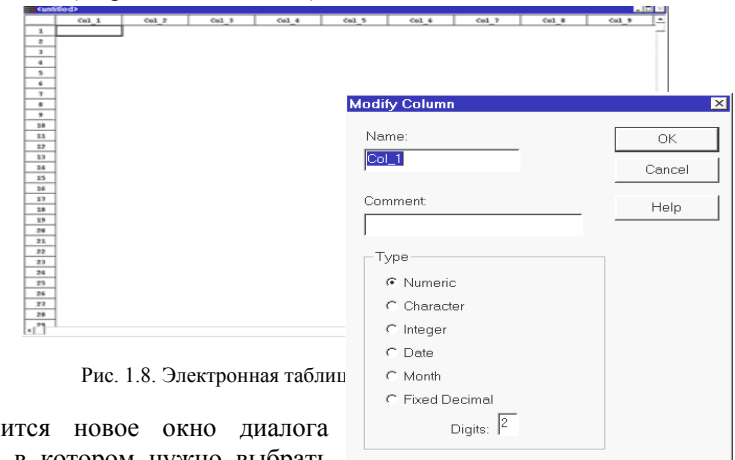


Рис. 1.8. Электронная таблица

Появится новое окно диалога (рис. 1.9), в котором нужно выбрать соответствующий тип вводимых данных (Numeric) и ввести имя переменной (Lamp).

Рис. 1.9. Панель модификации колонки

Следует иметь в виду, что разделителем целой и дробной части чисел в пакете STATGRAPHICS является запятая, а все имена вводятся латинскими буквами. После ввода данных в первую колонку их можно преобразовывать, вызвав из контекстного меню пункт Generate Date (Генерировать данные). Допускается более сотни манипуляций с переменными с помощью предоставляемых операторов. Наконец, надо сохранить файл данных командой File→Save Data File As, ввести имя файла и нажать ОК. В заголовке таблицы вместо <untitled> появится указанное имя.

Проанализируем теперь статистические данные. Выберем Describe→Numeric Data→One Variable Analysis (Анализ одной переменной). Появится окно для задания анализируемой переменной. В нашем случае это Lamp. После нажатия на кнопку ОК возникнет поле анализа одной

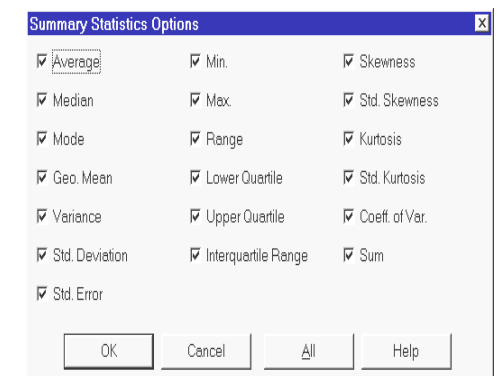


Рис. 1.10. Диалоговое окно задания общих статистик

переменной с первоначальной сводкой о введенных данных. В этой сводке указано имя Lamp, число наблюдений 65 и их пределы от 8.4 до 21.9. В верхней части рабочего поля расположены кнопки меню текущих операций, с помощью которых можно изменять входные данные, выбирать табличные и графические опции и сохранять результаты анализа в файле данных. Окна, в которых отображаются табличные и графические результаты, раскрываются на все рабочее поле двумя щелчками левой кнопки мыши. Щелчок правой кнопки мыши открывает доступ к специальному меню, задающему параметры графических изображений или изменения в текущем анализе данных. Например, при щелчке правой кнопки мыши на окне общих статистик на экране возникнет следующее диалоговое окно (рис. 1.10).

Для вычисления нужных статистик следует поставить галочку напротив соответствующих названий. Зададим вычисление всех суммарных статистик, а также гистограмму и график плотности, диаграмму рассеивания и квантильный график. В результате на рабочее поле будут выданы следующие табличные и графические изображения (рис. 1.11). На приведен-

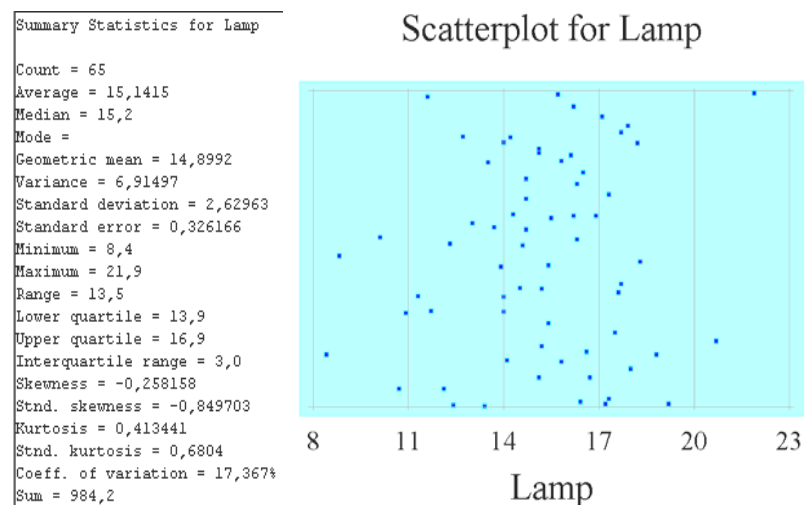


Рис. 1.11. Диаграмма рассеивания переменной Lamp

ных графиках и таблицах по порядку первой расположена таблица со значениями всех вычисленных статистик выборки. Справа от таблицы изображена диаграмма рассеивания элементов выборки (см. рис. 1.11). Далее идет таблица частот, включающая в себя значения верхней (Lower Limit) и нижней (Upper Limit) границ интервала группировки, его середину

(Midpoint), число (Frequency) и относительную частоту (Relative Frequency) попаданий в интервал группировки, а также их накопленные показатели. Ниже расположена сама гистограмма, построенная по данным предыдущей таблицы (рис. 1.12).

Наконец, последними приведены таблица первоначальных сведений об элементах выборки и таблица процентилей (рис. 1.13). Процентиля уровней 0.5, 0.25 и 0.75 соответствуют медиане, нижней и верхней квартили выборки. Справа от таблиц приведен процентильный график.

Меню Describe позволяет анализировать множества переменных, подбирать распределения и производить табуляцию данных. Эти возможности, не использованные в данном задании, частично будут задействованы в следующих лабораторных работах.

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		7,0		0	0,0000	0	0,0000
1	7,0	9,0	8,0	2	0,0308	2	0,0308
2	9,0	11,0	10,0	3	0,0462	5	0,0769
3	11,0	13,0	12,0	8	0,1231	13	0,2000
4	13,0	15,0	14,0	15	0,2308	28	0,4308
5	15,0	17,0	16,0	21	0,3231	49	0,7538
6	17,0	19,0	18,0	13	0,2000	62	0,9538
7	19,0	21,0	20,0	2	0,0308	64	0,9846
8	21,0	23,0	22,0	1	0,0154	65	1,0000
above	23,0			0	0,0000	65	1,0000

Mean = 15,1415 Standard deviation = 2,62963

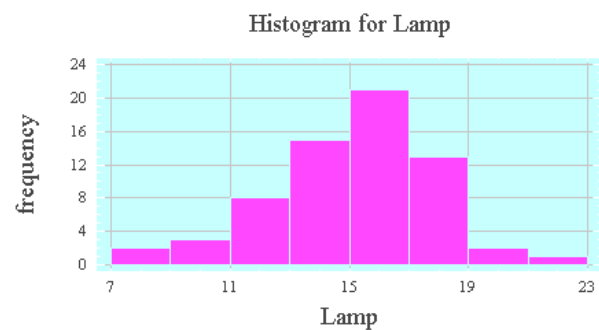


Рис. 1.12. Результаты анализа и гистограмма переменной Lamp

Analysis Summary
Data variable: Lamp
65 values ranging from 8,4 to 21,9
Percentiles for Lamp
1,0% = 8,4
5,0% = 10,7
10,0% = 11,6
25,0% = 13,9
50,0% = 15,2
75,0% = 16,9
90,0% = 18,0
95,0% = 18,8
99,0% = 21,9

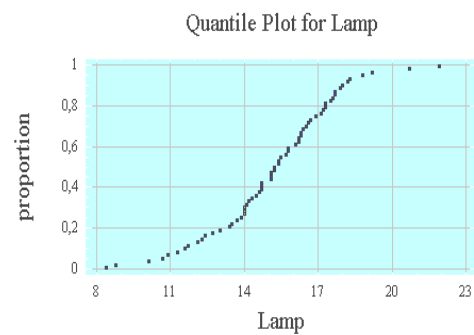


Рис. 1.13. Квантильный график переменной Lamp

Задание № 1. По данным выборкам с помощью пакета STATGRAPHICS вычислить все суммарные статистики, построить гистограмму, квантильный график и диаграмму рассеивания.

1. Урожайность зерновых культур в СССР с 1945 по 1989 гг. в центнерах с гектара:

5.6, 4.6, 7.3, 6.7, 6.9, 7.9, 7.4, 8.6, 7.8, 7.7, 8.4, 9.9, 8.4, 11.1, 10.4, 10.9, 10.7, 10.9, 8.3, 11.4, 9.5, 13.7, 12.1, 14.0, 13.2, 15.6, 15.4, 14.0, 17.6, 15.4, 10.9, 17.5, 15.0, 18.5, 14.2, 14.9, 12.6, 15.2, 15.9, 14.4, 16.2, 18.0, 18.3, 17.0, 18.8.

2. Время решения контрольной задачи учениками четвертого класса в секундах:

38, 60, 41, 51, 33, 42, 45, 21, 53, 60, 68, 52, 47, 46, 49, 49, 14, 57, 54, 59, 77, 47, 28, 48, 58, 32, 42, 58, 61, 30, 61, 35, 47, 72, 41, 45, 44, 55, 30, 40, 67, 65, 39, 48, 43, 60, 54, 42, 59, 50.

3. Измерения емкости затвора–стока у 80 полевых транзисторов дали следующие результаты:

1.9, 3.1, 1.3, 0.7, 3.2, 1.1, 2.9, 2.7, 2.7, 4.0, 1.7, 3.2, 0.9, 0.8, 3.1, 1.2, 2.6, 1.9, 2.3, 3.2, 4.1, 1.3, 2.4, 4.5, 2.5, 0.9, 1.4, 1.6, 2.2, 3.1, 1.5, 1.1, 2.3, 4.3, 2.1, 0.7, 1.2, 1.5, 1.8, 2.9, 0.8, 0.9, 1.7, 4.1, 4.3, 2.6, 0.9, 0.8, 1.2, 2.1, 3.2, 2.9, 1.1, 3.2, 4.5, 2.1, 3.1, 5.1, 1.1, 1.9, 0.9, 3.1, 0.9, 3.1, 3.3, 2.8, 2.5, 4.0, 4.3, 1.1, 2.1, 3.8, 4.6, 3.8, 2.3, 3.9, 2.4, 4.1, 4.2, 0.9.

4. Положительные отклонения от номинального размера у партии деталей в миллиметрах:

17, 21, 8, 20, 23, 18, 22, 20, 17, 12, 20, 11, 9, 19, 20, 9, 19, 17, 21, 13, 17, 22, 22, 10, 20, 20, 15, 19, 20, 20, 13, 21, 21, 9, 14, 11, 19, 18, 23, 19.

5. Время восстановления диодов из одной партии в наносекундах:

69, 73, 70, 68, 61, 73, 70, 72, 67, 70, 66, 70, 76, 68, 71, 71, 68, 70, 64, 65, 72, 70, 70, 69, 66, 70, 77, 69, 71, 74, 72, 72, 72, 68, 70, 67, 71, 67, 72, 69, 66, 75, 76, 69, 71, 67, 70, 73, 71, 74.

6. Время реакции в секундах:

8.5, 7.1, 6.7, 6.2, 2.9, 4.4, 6.0, 5.8, 5.4, 8.2, 6.9, 6.5, 6.1, 3.8, 6.0, 6.0, 5.6, 5.3, 7.7, 6.8, 6.5, 6.1, 4.2, 4.7, 5.6, 5.4, 5.3, 7.4, 6.7, 6.4, 6.1, 4.5, 6.0, 5.8, 5.6, 5.1.

7. Диаметры головок заклепок в миллиметрах:

13.39, 13.42, 13.38, 13.53, 13.51, 13.20, 13.40, 13.40, 13.28, 13.43, 13.46, 13.53, 13.55, 13.29, 13.24, 13.34, 13.54, 13.66, 13.43, 13.42, 13.38, 13.34, 13.57, 13.26, 13.33, 13.43, 13.50, 13.44, 13.53, 13.48, 13.48, 13.34, 13.36, 13.59, 13.36, 13.44, 13.34, 13.33, 13.25, 13.28, 13.49, 13.33, 13.26, 13.26, 13.55, 13.54, 13.37, 13.31, 13.37, 13.33.

8. Максимальная емкость двадцати подстроечных конденсаторов в пикофарадах:

4.45, 4.40, 4.42, 4.45, 4.38, 4.42, 4.36, 4.35, 4.40, 4.45, 4.42, 4.44, 4.36, 4.42, 4.44, 4.38, 4.39, 4.40, 4.42, 4.45.

9. Максимальные расходы воды реки Сыр-Дарьи у горы Беговат за 1910-1953 гг. в кубометрах в секунду:

2.46, 1.69, 1.34, 2.22, 2.18, 1.22, 1.22, 0.75, 1.26, 1.73, 1.74, 3.09, 1.57, 1.97, 2.23, 2.03, 1.58, 0.90, 2.40, 1.65, 1.96, 2.30, 1.79, 1.48, 3.44, 1.91, 3.06, 2.08, 1.06, 1.56, 1.88, 2.10, 2.02, 1.74, 1.18, 2.12, 1.38, 0.90, 1.45, 1.78, 1.97, 2.27, 2.34, 2.44.

10. Отклонения длины валиков от номинального размера в миллиметрах, отобранных из текущей продукции прецизионного токарного автомата:

1.0, 1.5, -2.5, 0.0, -1.5, 1.0, 1.0, 15.0, -1.0, 2.0, 2.0, 3.0, 11.0, -1.0, 5.0, 4.5, 0.5, 3.5, 8.0, 5.0, 4.5, 3.5, 9.5, 12.5, 7.5, 7.5, 10.0, 8.5, 10.0, -3.0, 5.0, 3.5, -3.0, -14.0, 17.0, -9.0, -13.0, -12.5, 8.5, 12.5, 6.0, 8.5, 0.0, 7.0, -1.0, -3.0, 0.5, 0.0, -2.0, -4.5, 2.0, -10.0, -8.5, -3.5, -11.5, -11.5, -7.5, -11.5, -6.5, 2.0.

11. Пробы железа имели следующие точки плавления (в градусах Цельсия):

1493, 1519, 1518, 1512, 1512, 1514, 1489, 1508, 1508, 1494, 1509, 1506, 1512, 1483, 1507, 1491, 1490, 1501, 1516, 1492, 1503, 1511, 1515, 1499, 1505.

12. Приведены данные по содержанию хрома (в весовых процентах) в образцах нержавеющей стали:

17.4, 17.9, 17.6, 18.1, 18.0, 17.6, 18.9, 18.2, 16.9, 17.5, 18.4, 17.8, 17.4, 18.5, 24.6, 20.8, 18.1, 26.0, 21.8, 17.7, 16.7, 18.8, 21.4, 19.5, 18.8.

13. Дано содержание железистой сыворотки (в микрограммах на 100 мл) в 40 образцах:

111, 107, 100, 99, 102, 106, 109, 108, 104, 99, 107, 108, 106, 98, 105, 103, 110, 105, 104, 100, 101, 96, 97, 102, 107, 113, 116, 113, 110, 98, 96, 108, 103, 104, 114, 114, 113, 108, 106, 99.

14. На телефонной станции проводились наблюдения над числом неправильных соединений в минуту. Наблюдения в течение часа дали следующие результаты:

3, 1, 3, 1, 4, 2, 2, 4, 0, 3, 0, 2, 2, 0, 2, 1, 4, 3, 3, 1, 4, 2, 2, 1, 1, 2, 1, 0, 3, 4, 1, 3, 2, 7, 2, 0, 0, 1, 3, 3, 1, 2, 4, 2, 0, 2, 3, 1, 2, 5, 1, 1, 0, 1, 1, 2, 2, 1, 1, 5.

15. При измерении диаметров валиков после шлифовки получены следующие результаты (в миллиметрах):

6.75, 6.77, 6.77, 6.73, 6.76, 6.74, 6.70, 6.75, 6.71, 6.77, 6.79, 6.73, 6.70, 6.74, 6.75, 6.71, 6.70, 6.78, 6.81, 6.69, 6.80, 6.68, 6.74, 6.83, 6.76, 6.82, 6.71, 6.77, 6.75, 6.82, 6.80, 6.72, 6.69, 6.81, 6.74, 6.80, 6.76, 6.77, 6.81, 6.82, 6.73, 6.72, 6.77, 6.78, 6.75, 6.68, 6.72, 6.69, 6.76, 6.70.

16. Октановое число бензина:

84.0, 83.5, 84.0, 85.0, 83.1, 83.5, 81.7, 85.4, 84.1, 83.0, 85.8, 82.4, 82.4, 83.4, 83.3, 83.1, 83.3, 82.4, 83.3, 82.6, 82.0, 83.2, 84.0, 84.2, 82.2, 83.6, 84.9, 83.2, 82.8, 83.4, 80.2, 82.7, 83.0, 85.0, 83.0, 85.0, 83.7, 83.6, 83.1, 82.5.

17. Замеры количества осадков (в сантиметрах), выпавших во время нескольких ураганов:

1.05, 1.40, 0.69, 1.41, 0.51, 1.49, 1.38, 2.00, 0.96, 1.31, 2.07, 1.02, 0.89, 1.51, 0.66, 1.16, 0.64, 1.07, 0.33, 1.59, 1.11, 1.33, 0.96, 1.40, 1.71, 0.75, 0.75, 0.92, 1.03, 0.78.

18. Результаты измерений некоторой физической характеристики пластикового материала, полученного из нескольких партий:

55, 42, 45, 41, 43, 53, 41, 43, 34, 50, 42, 41, 43, 46, 42, 44, 43, 45, 34, 48, 47, 46, 48, 41, 38, 49, 41, 44, 40, 48, 52, 50, 45, 30, 35, 52, 35, 46, 40, 48.

19. Время приготовления кофе, выраженное с точностью до сотых долей минуты, для нескольких типов электрических кофеварок:

1.38, 9.69, 0.39, 1.42, 0.54, 5.94, 0.59, 1.42, 0.39, 1.46, 0.55, 6.15, 0.61, 2.63, 2.44, 0.56, 0.69, 0.71, 0.95, 0.50, 2.69, 2.68, 0.53, 0.72, 0.74, 0.93, 0.53, 5.37, 2.18, 0.97.

20. Потери металла в сотнях тонн за период от установки оборудования до момента разрушения некоторой его части:

84, 60, 40, 47, 34, 46, 67, 92, 95, 40, 98, 60, 59, 108, 86, 117, 46, 93, 100, 92, 93, 79, 66, 82, 68.

21. Результаты измерения роста (в сантиметрах) случайно отобранных 50 студентов:

155.0, 159.1, 167.5, 181.7, 175.0, 164.8, 165.2, 171.6, 180.3, 170.0, 173.9, 168.3, 169.5, 169.0, 162.8, 165.1, 159.0, 161.5, 155.5, 160.8, 161.2, 175.0, 176.1, 167.2, 170.8, 165.2, 168.4, 157.3, 178.0, 182.0, 181.5, 175.0, 177.3, 171.6, 169.0, 165.3, 163.4, 166.0, 172.8, 159.3, 161.2, 157.1, 165.7, 160.4, 174.7, 165.4, 169.3, 173.8, 177.2, 179.6.

22. По данным 40 опытов получены следующие экспериментальные значения случайной величины X :

8, 14, 42, 22, -40, 18, -16, 38, -4, 2, -16, 34, 6, -11, 54, 8, 20, 74, -26, 0, 4, -28, 16, -22, 36, 44, 10, -13, 16, 24, -19, 46, 5, -7, 17, 23, 47, -21, 6, 14.

23. Годовое количество осадков (в дюймах) в Лондоне с 1863 по 1912 год:

21.59, 16.93, 29.48, 31.60, 26.25, 23.40, 25.42, 21.32, 25.02, 33.86, 22.67, 18.82, 28.44, 26.16, 28.17, 34.08, 33.82, 30.28, 27.92, 27.14, 24.40, 20.35, 26.64, 27.01, 19.21, 27.74, 23.85, 21.23, 28.15, 22.61, 19.80, 27.94, 21.47, 23.52, 22.86, 17.69, 22.54, 23.28, 22.17, 20.84, 38.10, 20.65, 22.97, 24.26, 23.01, 23.67, 26.75, 25.36, 24.79, 27.88.

24. Годовая урожайность ячменя (в центнерах на 1 акр) в Англии и Уэльсе с 1890 по 1939 год:

16.7, 16.3, 16.5, 13.3, 16.5, 15.0, 15.9, 15.5, 16.9, 16.4, 14.9, 14.5, 16.6, 15.1, 14.6, 16.0, 16.8, 16.8, 15.5, 17.3, 15.5, 15.5, 14.2, 15.8, 15.7, 14.1, 14.8, 14.4, 15.6, 13.9, 14.7, 14.3, 14.0, 14.5, 15.4, 15.3, 16.0, 16.4, 17.2, 17.8, 14.4, 15.0, 16.0, 16.8, 16.9, 16.6, 16.2, 14.0, 18.1, 17.5.

25. Ряд из 60 равномерных случайных чисел, принимающих целые значения от 0 до 19:

3, 15, 15, 8, 19, 1, 3, 12, 19, 13, 16, 4, 17, 8, 6, 15, 3, 3, 7, 4, 5, 14, 15, 10, 3, 10, 13, 14, 15, 8, 10, 1, 18, 17, 4, 10, 16, 2, 13, 3, 14, 7, 16, 3, 10, 12, 0, 3, 2, 3, 10, 5, 10, 3, 2, 11, 14, 18, 8, 14.

26. Поквартальные индексы розничной цены на овощи в Англии в 1951–1958 годах:

295.0, 317.5, 314.9, 321.4, 324.7, 323.7, 322.5, 332.9, 372.9, 380.9, 353.0, 348.9, 354.0, 345.7, 319.5, 317.6, 333.7, 323.9, 312.8, 310.2, 323.2, 342.9, 300.3, 309.8, 304.3, 285.9, 292.3, 298.7, 312.5, 336.1, 295.5, 318.4.

27. В результате измерения контролируемого размера отобранных изделий получены следующие наблюдения:

1.08, 1.10, 1.12, 1.38, 1.18, 1.12, 1.36, 1.25, 1.15, 1.14, 1.40, 1.42, 1.11, 1.22, 1.36, 1.33, 1.35, 1.35, 1.41, 1.21, 1.37, 1.13, 1.15, 1.29, 1.31, 1.17, 1.45, 1.34, 1.17, 1.23, 1.39, 1.06, 1.26, 1.31, 1.37.

28. В результате взвешивания 30 проб химического вещества получены следующие данные (в миллиграммах):

25, 28, 30, 31, 28, 26, 50, 52, 20, 24, 26, 23, 40, 36, 28, 31, 32, 33, 36, 35, 29, 42, 42, 45, 38, 40, 41, 29, 25, 37.

29. Получены следующие результаты анализов на содержание углерода (в процентах) в пробах нелегированной стали: 0.18, 0.12, 0.12, 0.08, 0.08, 0.12, 0.19, 0.32, 0.27, 0.11, 0.14, 0.23, 0.16, 0.09, 0.08, 0.05, 0.13, 0.17, 0.10, 0.14, 0.30, 0.27, 0.31, 0.24, 0.22, 0.34, 0.14, 0.46, 0.39, 0.24, 0.28, 0.11, 0.42, 0.29, 0.11.

30. Результаты лабораторных анализов 60 образцов сланцевых пород на содержание кремния (SiO_2) в процентах:

57.8, 54.6, 54.8, 51.7, 61.1, 62.3, 52.2, 49.2, 53.9, 60.0, 56.2, 55.2, 53.3, 57.9, 54.0, 52.6, 53.8, 53.6, 51.5, 54.0, 50.4, 53.0, 53.3, 51.6, 50.9, 49.6, 52.2, 50.5, 51.1, 52.2, 49.2, 49.3, 48.8, 53.5, 52.8, 52.9, 52.1, 47.3, 49.8, 49.3, 50.1, 54.4, 49.0, 48.9, 51.3, 51.6, 46.2, 50.4, 50.7, 53.1, 52.9, 51.3, 52.7, 46.6, 46.5, 51.3, 51.0, 47.5, 47.7, 44.9.

1.7. Нормальное распределение и его числовые характеристики

В теории вероятностей нормальный закон занимает особое место, так как является предельным законом для многих других при выполнении некоторых весьма нежестких ограничений. Именно, распределение суммы случайных величин следует приближенно нормальному закону, если среди этих случайных величин нет резко выделяющихся, сами же случайные величины в отдельности могут быть подчинены любому закону.

Случайная величина X имеет нормальное распределение, если ее функция плотности вероятности имеет вид

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}. \quad (1.6.1)$$

Закон имеет два параметра m и σ , т. е. относится к классу двухпараметрических законов. Графики функций плотности вероятности и функции распределения приведены на рис. 1.14.

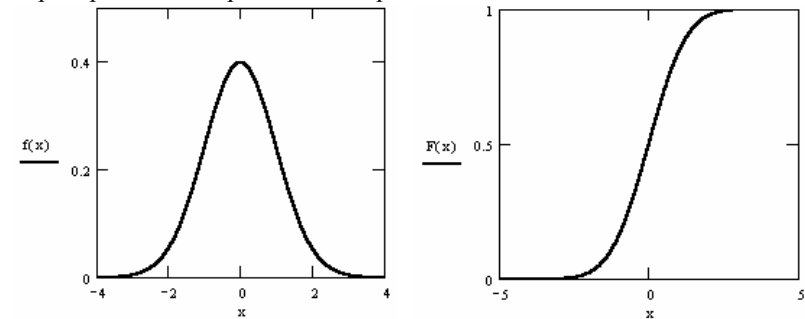


Рис. 1.14. Графики функции плотности вероятности и функции распределения стандартного нормального закона

Найдем как всегда сначала функцию распределения

$$F(x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-m)^2}{2\sigma^2}} dt = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt = \left\langle \frac{t-m}{\sigma} = u, \frac{dt}{dt} = \sigma du \right\rangle =$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\frac{x-m}{\sigma}} e^{-\frac{u^2}{2}} \sigma du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-m}{\sigma}} e^{-\frac{t^2}{2}} dt.$$

Последний интеграл не выражается через элементарные функции. Он называется функцией Лапласа* и обозначается

$$\Phi\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-m}{\sigma}} e^{-\frac{t^2}{2}} dt. \quad (1.6.2)$$

Исторически различают несколько разновидностей функции Лапласа. Формула (1.6.2) дает обычную функцию Лапласа, называемую функцией Лапласа; функция

$$\Phi^*\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_0^{\frac{x-m}{\sigma}} e^{-\frac{t^2}{2}} dt \quad (1.6.3)$$

называется нормированной функцией Лапласа; наконец, в артиллерии широко применяется формула

$$\bar{\Phi}\left(\frac{x-m}{\sigma}\right) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{x-m}{\sigma}} e^{-t^2} dt, \quad (1.6.4)$$

которая называется приведенной функцией Лапласа. Связь между всеми этими функциями легко устанавливается по общему правилу замены переменных в определенном интеграле. Например,

$$\Phi(x) = \frac{1}{2} + \Phi^*\left(\frac{x}{\sigma}\right), \quad \bar{\Phi}(x) = 2\Phi\left(\sqrt{2}x\right) - 1.$$

Определим параметры нормального закона. Ограничимся двумя точечными характеристиками: математическим ожиданием и дисперсией

$$m_X = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \left\langle \frac{x-m}{\sigma\sqrt{2}} = t, \frac{dx}{dx} = \sqrt{2}\sigma dt \right\rangle = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + m) e^{-t^2} \sqrt{2}\sigma dt =$$

* Пьер Симон Лаплас (1749-1827) - французский математик, механик и астроном.

$$= \frac{\sigma\sqrt{2}}{\sqrt{\pi}} \int_{-\infty}^{\infty} te^{-t^2} dt + \frac{m}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt = \frac{\sigma\sqrt{2}}{2\sqrt{\pi}} \left(-e^{-t^2} \right) \Big|_{-\infty}^{\infty} + \frac{m}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt = \frac{m}{\sqrt{\pi}} \sqrt{\pi} = m.$$

Здесь использован интеграл Эйлера* -Пуассона**

$$\int_{-\infty}^{\infty} e^{-t^2} dt = 2 \int_0^{\infty} e^{-t^2} dt = \sqrt{\pi}. \text{ Аналогично}$$

$$D_X = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - m_X)^2 e^{-\frac{(x-m_X)^2}{2\sigma^2}} dx = \left\langle \frac{x - m_X}{\sqrt{2\sigma}} = t, \right\rangle = \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^2 e^{-t^2} dt =$$

$$= \left\langle \begin{matrix} t = u, dt = du, \\ te^{-t^2} dt = dv, \\ v = -\frac{1}{2} e^{-t^2}. \end{matrix} \right\rangle = \left[-\frac{1}{2} te^{-t^2} \right]_{-\infty}^{\infty} + \frac{1}{2} \int_{-\infty}^{\infty} e^{-t^2} dt \left] \frac{2\sigma^2}{\sqrt{\pi}} = \frac{2\sigma^2}{\sqrt{\pi}} \cdot \frac{1}{2} \sqrt{\pi} = \sigma^2.$$

Итак, $D_X = \sigma^2$, $\sigma_X = \sigma$. Ясен и смысл параметров m и σ нормального распределения: m - центр рассеивания - является и центром симметрии распределения. Это хорошо видно из графика функции плотности вероятности. Размерность центра рассеивания m равна размерности случайной величины X . Несколько сложнее обстоит дело с параметром σ . Этот параметр характеризует форму кривой распределения. При увеличении σ график все более «размазывается» по оси OX , т.е. случайная величина X имеет большее рассеивание около центра симметрии. Чем меньше σ , тем более «островершинен» график функции плотности вероятности. Размерность D_X совпадает с размерностью X^2 . Легко показать, что для нормального распределения $d_X = h_X = m_X$. Полезна формула, выражающая любой центральный момент нормального распределения через дисперсию $\mu_k = (k-1)!! D_X$.

* Леонард Эйлер (1707-1783) – швейцарский математик.

** Симон Дениз Пуассон (1781-1840) – французский математик.

2. РАСПРЕДЕЛЕНИЯ, СВЯЗАННЫЕ С НОРМАЛЬНЫМ РАСПРЕДЕЛЕНИЕМ

Данные распределения играют в статистической методологии исключительно важную роль. Они широко используются наряду с нормальным, когда рассматривается распределение выбранных статистик.

2.1. χ^2 -распределение

Пусть X_1, X_2, \dots, X_n - независимые случайные величины, каждая из которых имеет нормальное распределение $N(0,1)$. Обозначим сумму их квадратов через $\chi^2_n = X_1^2 + X_2^2 + \dots + X_n^2 = \sum_{i=1}^n X_i^2$. Очевидно,

$\chi^2_n \geq 0$ и $P(\chi^2_n < 0) = 0$. Эта сумма квадратов имеет плотность распределения

$$k_n(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (2.1.1)$$

Здесь $\Gamma\left(\frac{n}{2}\right)$ - гамма-функция (Эйлеров интеграл второго рода);

$\Gamma\left(\frac{n}{2}\right) = \int_0^\infty e^{-x} x^{\frac{n}{2}-1} dx$. Интегральная функция распределения имеет вид

$$\begin{cases} K_n(x) = P(\chi^2_n < x) = \int_0^x k_n(t) dt \text{ или} \\ 1 - K_n(x) = P(\chi^2_n \geq x) = \int_x^\infty k_n(t) dt. \end{cases} \quad (2.1.2)$$

Число n называется числом степеней свободы данного распределения. Формула (2.1.1) выводится методом математической индукции.

$$\begin{aligned} & \text{Если } X^2 \leq n, \text{ то } -\sqrt{n} \leq X \leq \sqrt{n} \text{ и } P\{X^2 \leq n\} = P\{-\sqrt{n} \leq X \leq \sqrt{n}\} = \\ & = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{n}}^{\sqrt{n}} e^{-\frac{z^2}{2}} dz = \sqrt{\frac{2}{\pi}} \int_0^{\sqrt{n}} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_0^n e^{-\frac{t}{2}} \frac{1}{2} dt, \text{ здесь } z = \sqrt{t}. \text{ Отсю-} \end{aligned}$$

да функция распределения случайной величины X_1^2 равна

$$P(X_1^2 < x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t}{2}} \frac{1}{2} dt, \text{ а ее плотность вероятности по теореме Бар-}$$

роу*

$$k_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}} x_1^{-\frac{1}{2}} e^{-\frac{x}{2}}, \quad x_1 \geq 0. \quad (2.1.3)$$

Это и есть распределение χ^2 с одной степенью свободы. Теперь получим общую формулу для распределения $X_n = \chi^2 = X_1^2 + X_2^2 + \dots + X_n^2$, где n - любое число. Воспользуемся для этого методом математической индукции. Допустим, что $k(x_n)$ описывается неким выражением, и покажем, что $k(x_{n+1})$ имеет аналогичное выражение. Пусть

$$k(x_n) = C_n x_n^{\frac{n}{2}-1} e^{-\frac{x_n}{2}}, \quad x_n \geq 0, \quad (2.1.4)$$

где C_n - константа, причем такая, что $C_n \int_0^\infty x_n^{\frac{n}{2}-1} e^{-\frac{x_n}{2}} dx_n = 1$ (условие

нормировки). Полагая в (2.1.4) $n = 1$, приходим к формуле (2.1.3). Действительно,

$$\begin{aligned} C_1 &= \left(\int_0^\infty x^{-\frac{1}{2}} e^{-\frac{x}{2}} dx \right)^{-1} \cdot \int_0^\infty x^{\frac{1}{2}-1} e^{-\frac{x}{2}} dx = \left\langle \begin{array}{l} \frac{x}{2} = y, \quad x = 2y, \\ dx = 2dy, \\ x^{-\frac{1}{2}} = \frac{1}{\sqrt{2}} y^{-\frac{1}{2}} \end{array} \right\rangle = \\ &= \frac{2}{\sqrt{2}} \int_0^\infty y^{-\frac{1}{2}} e^{-y} dy = \sqrt{2} \Gamma\left(\frac{1}{2}\right) = \sqrt{2\pi}. \end{aligned}$$

* Исаак Барроу (1630-1677) - английский математик.

Тогда $C_1 = 1/\sqrt{2\pi}$, что соответствует формуле (2.1.3). Теперь положим $Y_{n+1} = X_{n+1}^2$, где $X_{n+1} \in N(0, 1)$ и не зависит от X_n . Совместная плотность $p_{X_n, Y_{n+1}}(x, y) = k(x_n) \cdot p_{Y_{n+1}}(y)$. Но распределение Y_{n+1} такое же как и X_1^2 , т.е. это χ^2 -распределение с одной степенью свободы. Тогда

$$f_{X_n, Y_{n+1}}(x, y) = k(x_n) \cdot k(x_1) = \frac{C_n}{\sqrt{2\pi}} x_{n+1}^{\frac{n}{2}-1} x_1^{-\frac{1}{2}} e^{-\frac{(x_n+x_1)}{2}}, \quad x_n, x_1 > 0. \quad \text{Введем}$$

преобразование $\begin{cases} x_{n+1} = x_n + x_1, \\ z_{n+1} = x_n. \end{cases}$ Тогда обратное преобразование будет

таково: $\begin{cases} x_n = z_{n+1}, \\ x_1 = x_{n+1} - z_{n+1}, \end{cases}$ а якобиан преобразования

$$\frac{\partial(x_n, x_1)}{\partial(x_{n+1}, z_{n+1})} = \frac{\begin{vmatrix} \frac{\partial x_n}{\partial x_{n+1}} & \frac{\partial x_n}{\partial z_{n+1}} \\ \frac{\partial x_1}{\partial x_{n+1}} & \frac{\partial x_1}{\partial z_{n+1}} \end{vmatrix}}{\begin{vmatrix} \frac{\partial x_n}{\partial x_{n+1}} & \frac{\partial x_n}{\partial z_{n+1}} \\ \frac{\partial x_1}{\partial x_{n+1}} & \frac{\partial x_1}{\partial z_{n+1}} \end{vmatrix}} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1 \neq 0. \quad \text{Следовательно, плот-}$$

ность распределения

$$f_{X_{n+1}, Z_{n+1}}(x_{n+1}, z_{n+1}) = \frac{C_n}{\sqrt{2\pi}} z_{n+1}^{\frac{n}{2}-1} (x_{n+1} - z_{n+1})^{-\frac{1}{2}} e^{-\frac{x_{n+1}}{2}}, \quad (2.1.5)$$

$$0 \leq z_{n+1} \leq x_{n+1}.$$

Проинтегрируем уравнение (2.1.5) по z_{n+1} , получим

$$\begin{aligned} f_{X_{n+1}}(x_{n+1}) &= \frac{C_n}{\sqrt{2\pi}} e^{-\frac{x_{n+1}}{2}} \cdot \int_0^{x_{n+1}} z_{n+1}^{\frac{n}{2}-1} (x_{n+1} - z_{n+1})^{-\frac{1}{2}} dz_{n+1} = \\ &= \left\langle \begin{matrix} z_{n+1} = tx_{n+1}, \\ z_{n+1} = 0, \quad t = 0, \\ z_{n+1} = x_{n+1}, \quad t = 1, \\ dz_{n+1} = x_{n+1} dt \end{matrix} \right\rangle = \frac{C_n}{\sqrt{2\pi}} e^{-\frac{x_{n+1}}{2}} \cdot \int_0^1 (1-t)^{-\frac{1}{2}} t^{\frac{n}{2}-1} x_{n+1}^{\frac{n}{2}-1+\frac{1}{2}} dt = \\ &= \frac{C_n}{\sqrt{2\pi}} \int_0^1 (1-t)^{-\frac{1}{2}} t^{\frac{n}{2}-1} e^{-\frac{x_{n+1}}{2}} x_{n+1}^{\frac{n+1}{2}-1} dt = C_{n+1} x_{n+1}^{\frac{n+1}{2}-1} e^{-\frac{x_{n+1}}{2}}, \quad x_{n+1} \geq 0 \quad (2.1.6) \end{aligned}$$

По форме записи выражение (2.1.6) аналогично выражению (2.1.4), только n увеличилось до $n+1$. Таким образом, доказательство по методу математической индукции завершено. Величина C_n определяется из ус-

ловия нормировки $\int_{-\infty}^{\infty} f_{X_n}(x_n) dx_n = 1$. Подставив значение $k_{X_n}(x_n)$ из

(2.1.4), получим $C_n \int_0^{\infty} x_n^{\frac{n}{2}-1} e^{-\frac{x_n}{2}} dx_n = 1$, но так как $\int_0^{\infty} e^{-x} x^{a-1} dx = \Gamma(a)$, то

$$\int_0^{\infty} x_n^{\frac{n}{2}-1} e^{-\frac{x_n}{2}} dx_n = \left\langle \frac{x_n}{2} = y, \right\rangle = \int_0^{\infty} 2(2y)^{\frac{n}{2}-1} e^{-y} dy = 2 \cdot 2^{\frac{n}{2}-1} \int_0^{\infty} y^{\frac{n}{2}-1} e^{-y} dy =$$

$$= 2^{\frac{n}{2}} \Gamma(n/2), \text{ тогда } C_n = \frac{1}{2^{\frac{n}{2}} \Gamma(n/2)}.$$

Следовательно, окончательно плот-

$$\text{ность } \chi^2 \text{-распределения имеет вид } k(x_n) = \frac{1}{2^{\frac{n}{2}} \Gamma(n/2)} x_n^{\frac{n}{2}-1} e^{-\frac{x_n}{2}}, \quad x_n \geq 0.$$

При $n = 2$ $k_2(x) = 1/2 e^{-\frac{x}{2}}$, $x \geq 0$, т.е. χ^2 -распределение с двумя степенями свободы является экспоненциальным распределением с $\lambda = 1/2$.

Можно привести другой более традиционный вывод этой же формулы. Величины X_i независимы, и каждая из них имеет по условию плот-

ность $1/\sqrt{2\pi} e^{-\frac{x_i^2}{2}}$, тогда совместная плотность величин X_1, X_2, \dots, X_n

выразится произведением $1/\sqrt{2\pi} e^{-\frac{x_1^2}{2}} \cdot 1/\sqrt{2\pi} e^{-\frac{x_2^2}{2}} \cdot \dots \cdot 1/\sqrt{2\pi} e^{-\frac{x_n^2}{2}} =$

$$= (1/\sqrt{2\pi})^n e^{-\frac{\sum x_i^2}{2}}, \text{ отсюда}$$

$$K_n(x) = P(\chi^2 < x) = \frac{1}{(2\pi)^{\frac{n}{2}} \sum x_i^2 < x} \int \dots \int e^{-\frac{1}{2}(x_1^2 + x_2^2 + \dots + x_n^2)} dx_1 dx_2 \dots dx_n. \quad (2.1.7)$$

Это n -мерный интеграл, распространенный на область, определяемую неравенством $\sum_i x_i^2 < x$. Область в общем случае представляет собой множество точек, лежащих внутри и на поверхности сферы n -мерного пространства радиуса \sqrt{x} с центром в начале координат.

Так как $k_n(x) = K'_n(x)$, найдем производную формулы (2.1.7) по определению. Дадим x приращение h , получим

$$K_n(x+h) - K_n(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int \int \dots \int_{x < \sum x_i^2 \leq x+h} e^{-\frac{1}{2}(x_1^2 + x_2^2 + \dots + x_n^2)} dx_1 dx_2 \dots dx_n.$$

Применим к этой формуле теорему о среднем, тогда

$$K_n(x+h) - K_n(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(x+\theta h)} \int \int \dots \int_{x < \sum x_i^2 \leq x+h} dx_1 dx_2 \dots dx_n. \quad (2.1.8)$$

Здесь $e^{-\frac{1}{2}(x+\theta h)}$ ($0 < \theta < 1$) есть некоторое среднее значение подынтегральной функции $e^{-\frac{1}{2}\sum x_i^2}$ в области интегрирования $x < \sum x_i^2 < x+h$.

Положим $S_n(x) = \int \int \dots \int_{\sum x_i^2 \leq x} dx_1 dx_2 \dots dx_n$. Это объем n -мерной сферы

радиуса \sqrt{x} . Интеграл в правой части уравнения (2.1.8) перепишем в виде

$$K_n(x+h) - K_n(x) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}(x+\theta h)} [S_n(x+h) - S_n(x)]. \quad (2.1.9)$$

Произведем в формуле для $S_n(x)$ замену переменных:

$$x_i = \sqrt{x} y_i, \quad dx_i = \sqrt{x} dy_i, \quad x_1^2 + x_2^2 + \dots + x_n^2 = x y_1^2 + x y_2^2 + \dots + x y_n^2 = x(y_1^2 + y_2^2 + \dots + y_n^2) \leq x, \quad \text{т.е.} \quad y_1^2 + y_2^2 + \dots + y_n^2 \leq 1. \quad \text{Отсюда}$$

$$S_n(x) = (\sqrt{x})^n \int \int \dots \int_{\sum y_i^2 \leq 1} dy_1 dy_2 \dots dy_n = C_1 x^{\frac{n}{2}}, \quad \text{где } C_1 - \text{объем единичной сферы } n\text{-мерного пространства.}$$

Наконец, формулу (2.1.9) запишем в виде

$$K_n(x+h) - K_n(x) = (2\pi)^{-\frac{n}{2}} C_1 e^{-\frac{1}{2}(x+\theta h)} \left[(x+h)^{\frac{n}{2}} - x^{\frac{n}{2}} \right], \quad \text{отсюда}$$

$$\frac{K_n(x+h) - K_n(x)}{h} = C_2 e^{-\frac{1}{2}(x+\theta h)} \frac{(x+h)^{\frac{n}{2}} - x^{\frac{n}{2}}}{h}. \quad \text{Перейдем к пределу при}$$

$h \rightarrow 0$. Так как при этом $\lim_{h \rightarrow 0} \frac{(x+h)^{\frac{n}{2}} - x^{\frac{n}{2}}}{n} = \frac{n}{2} x^{\frac{n}{2}-1}$ - производная степ-

пенной функции, то $\lim_{h \rightarrow 0} \frac{K_n(x+h) - K_n(x)}{h} = k_n(x) = C_3 e^{-\frac{1}{2}x} x^{\frac{n}{2}-1}$. Кон-

станту C_3 легче всего найти из условия нормировки $C_3 \int_0^{\infty} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx = 1$.

Как было показано ранее $C_3 = 1/2^{n/2} \Gamma(n/2)$, поэтому для плотности рас-
пределения получим выражение $k(x_n) = \frac{1}{2^{n/2} \Gamma(n/2)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, x \geq 0$.

Число степеней свободы n распределения можно связать с числом независимых величин, остающихся после оценки параметров или подбора распределения. Этот термин имеет разный смысл в различных задачах. На рис. 2.1 представлены χ^2 -кривые с числами степеней свободы n , равными 2, 4, 8 и 16. При $\chi^2 = 0$ тангенс угла наклона кривой обращается в бесконечность для $n = 3$, он остается конечным и ненулевым при $n = 4$ и обращается в нуль при $n > 4$. С ростом n кривая приближается к симметричной кривой. Справа изображен график функции распределения для $n = 4$.

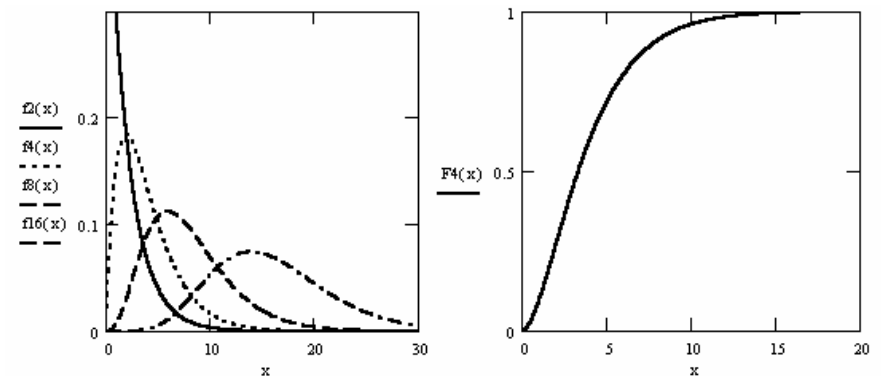


Рис. 2.1. Различные функции плотности и функция распределения χ^2 -распределения

Числовые характеристики распределения: 1) математическое ожидание $m_x = n$; 2) дисперсия $D_x = 2n$, $\sigma_x = \sqrt{2n}$; 3) мода $d_x = n - 2$, $n \geq 2$;

4) медиана $h_x \approx n - 0.67$; 5) коэффициент асимметрии $A = \frac{\mu_3}{\sigma_x^3} = \frac{\sqrt{2^3}}{\sqrt{n}}$;

6) коэффициент эксцесса $E = \frac{\mu_4}{\sigma_x^4} - 3 = 3 + \frac{12}{n}$.

Когда число степеней свободы стремится к бесконечности, A и E стремятся к нулю и трем соответственно, т.е. к значениям этих моментов для нормального распределения. Можно показать, что распределение данных случайных величин стремится при $n \rightarrow \infty$ к нормированному нормальному распределению. Приведем несколько наиболее употребительных формул приближения к нормальному распределению:

$$1) X_1 = \frac{\chi^2_{\frac{n-2}{n}}}{\sqrt{2n}}, \quad \chi^2_n \in N(n, 2n), \quad X_1 \in N(0, 1),$$

$$2) X_2 = \sqrt{2\chi^2_{\frac{n-2}{n}}} - \sqrt{2n-1} \quad (\text{аппроксимация Фишера}^*),$$

$$\sqrt{2\chi^2_{\frac{n-2}{n}}} \in N(\sqrt{2n-1}, 1), \quad X_2 \in N(0, 1),$$

$$3) X_3 = \left[\left(\frac{\chi^2_{\frac{n-2}{n}}}{n} \right)^{1/3} - 1 + \frac{2}{9n} \right] \sqrt{\frac{9n}{2}} \quad (\text{аппроксимация Вильсона}^{**} - \text{Хил-фери}),$$

$$\left(\frac{\chi^2_{\frac{n-2}{n}}}{n} \right)^{1/3} \in N\left(\frac{2}{9n} - 1, \frac{2}{9n}\right), \quad X_3 \in N(0, 1).$$

Распределение χ^2 обладает одним замечательным свойством: две независимые величины χ_1^2 и χ_2^2 , распределенные по закону χ^2 с n_1 и n_2 степенями свободы, при сложении дают в сумме величину $\chi_1^2 + \chi_2^2$, распределенную по закону χ^2 с $n_1 + n_2$ степенями свободы.

* Роналд Эйлмер Фишер (1890-1962) - английский математик.

** Эдвин Бидвел Вильсон (1879-1964) - английский математик.

2.2. t - распределение Стюдента***

Вторым из числа распределений, широко используемых в статистических проверках, является t -распределение Стюдента или просто t -распределение, впервые предложенное Госсетом и затем более строго обосновано Фишером. Оно лежит в основе множества процедур статистического анализа в науке и технике. На простом t -критерии основаны очень многие более сложные статистические критерии. Распределению Стюдента подчиняется статистика

$$t = z\sqrt{n}/\sqrt{v}, \quad (2.2.1)$$

где z и v независимы, z распределена нормально, $z \in N(0,1)$, а v подчиняется закону χ^2 с n степенями свободы. При этих условиях плотность вероятности величины t имеет вид

$$s_n(x) = B_n \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad B_n = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}. \quad (2.2.2)$$

Функция распределения обозначается через $S_n(x) = \int_{-\infty}^x s_n(t)dt$. Ее гра-

фик сильно напоминает график нормального распределения, с ростом n распределение t стремится к нормированному нормальному распределению $N(0,1)$. График функции распределения также очень похож на нормальный. Графики трех функций плотности вероятности: $f(x)$ - плотность стандартного нормального распределения, $f1(x)$ - плотность распределения Стюдента с одной степенью свободы, $f4(x)$ - плотность распределения Стюдента с четырьмя степенями свободы и график функции распределения Стюдента $F(x)$ представлены на рис. 2.2. Выведем фор-

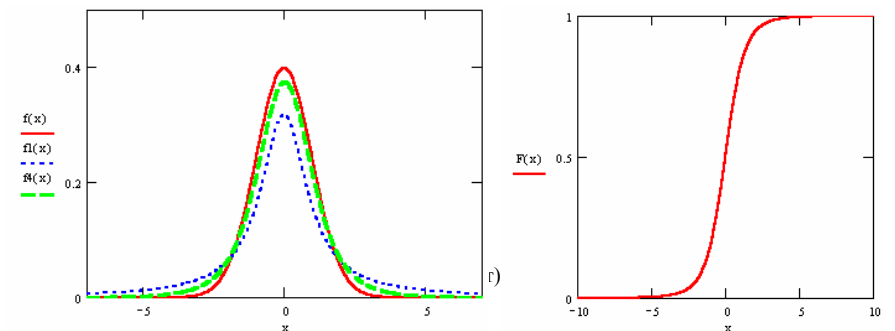


Рис. 2.2. Кривые плотности и распределения закона Стюдента

мулу (2.2.2) для функции плотности вероятности распределения Стьюдента. Так как $z \in N(0,1)$, а $v \in \chi_n^2$, то $f(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$, а

$$f_n(v) = \frac{1}{2^{\frac{n}{2}} \Gamma(n/2)} e^{-\frac{v}{2}} v^{\frac{n}{2}-1}. \text{ Плотность совместного распределения в этом}$$

случае будет равна

$$f(z, v) = C e^{-\frac{z^2}{2} - \frac{v}{2}} v^{\frac{n}{2}-1}, \quad C = 1 / \sqrt{2\pi} 2^{\frac{n}{2}} \Gamma(n/2). \quad (2.2.3)$$

Тогда

$$S_n(x) = P(t < x) = P\left(\frac{z\sqrt{n}}{\sqrt{v}} < x\right) = P\left(z < \frac{x\sqrt{v}}{\sqrt{n}}\right) = C \iint_{z < \frac{x\sqrt{v}}{\sqrt{n}}} e^{-\frac{z^2}{2} - \frac{v}{2}} v^{\frac{n}{2}-1} dv dz.$$

Область интегрирования определяется неравенством $-\infty < z < x\sqrt{v}/\sqrt{n}$ и представляет множество точек плоскости, ограниченное ветвью параболы $z = (x/\sqrt{n})\sqrt{v}$. Выполняя двойное интегрирование по z от $-\infty$ до $(x/\sqrt{n})\sqrt{v}$, а затем по v от 0 до ∞ , найдем $S_n(x) =$

$$= C \int_0^\infty v^{\frac{n}{2}-1} e^{-\frac{v}{2}} dv \int_{-\infty}^{\frac{x}{\sqrt{n}}\sqrt{v}} e^{-\frac{z^2}{2}} dz$$

(рис. 2.3). Вычислим сразу функцию плотности вероятности, дифференцируя полученное выражение по x в правой части под знаком интеграла. Тогда

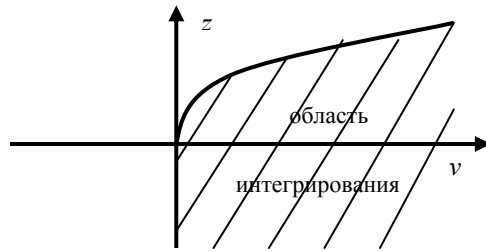


Рис. 2.3. Область существования статистики t

$$s_n(x) = S_n'(x) = C \int_0^\infty v^{\frac{n-1}{2}} e^{-\frac{v}{2}} \left\{ \frac{d}{dx} \int_{-\infty}^{\frac{x}{\sqrt{n}}\sqrt{v}} e^{-\frac{z^2}{2}} dz \right\} dv = C \int_0^\infty v^{\frac{n-1}{2}} e^{-\frac{v}{2}} e^{-\frac{x^2 v}{2n}} \frac{\sqrt{v}}{\sqrt{n}} dv =$$

$$= \frac{C}{\sqrt{n}} \int_0^\infty v^{\frac{n-1}{2}} e^{-\frac{v}{2} \left(1 + \frac{x^2}{n} \right)} dv.$$

Выполним подстановку $u = \frac{v}{2} \left(1 + (x^2/n) \right)$, $dv = \frac{2du}{1 + (x^2/n)}$,

$$e^{-\frac{v}{2} \left(1 + (x^2/n) \right)} = e^{-\frac{u}{\left(1 + (x^2/n) \right)} \left(1 + (x^2/n) \right)} = e^{-u}, \quad v^{(n-1)/2} = \frac{(2u)^{(n-1)/2}}{\left(1 + (x^2/n) \right)^{(n-1)/2}},$$

получим

$$s_n(x) = \frac{C}{\sqrt{n}} \frac{2^{\frac{n-1}{2}} \cdot 2}{\left(1 + \frac{x^2}{n} \right)^{\frac{n-1}{2}} \left(1 + \frac{x^2}{n} \right)^0} \int_0^\infty u^{\frac{n-1}{2}} e^{-u} du. \quad \text{Так как}$$

$$\int_0^\infty u^{\frac{n-1}{2}} e^{-u} du = \int_0^\infty u^{\frac{n+1}{2}-1} e^{-u} du = \Gamma\left(\frac{n+1}{2}\right), \quad \text{то } s_n(x) = \frac{2^{\frac{n+1}{2}}}{\left(1 + \frac{x^2}{n} \right)^{\frac{n+1}{2}}} \frac{C}{\sqrt{n}} \Gamma\left(\frac{n+1}{2}\right) =$$

$$= \frac{2^{\frac{n+1}{2}}}{\left(1 + \frac{x^2}{n} \right)^{\frac{n+1}{2}}} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n} \cdot \sqrt{2\pi} \cdot 2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \sqrt{n\pi}} \left(1 + \frac{x^2}{n} \right)^{-\frac{n+1}{2}}.$$

Числовые характеристики распределения: $m_X = 0$, $D_X = n/(n-2)$, $n > 2$; мода $d_X = 0$; медиана $h_X = 0$, коэффициент асимметрии $A = 0$, $n > 3$; коэффициент эксцесса $E = 6/(n-4)$, $n > 4$. Нормальная аппрок-

симация $N(0, \sqrt{n/(n-2)})$ очень хороша при $n \geq 30$, т.е.

$$X_1 = t / \sqrt{n/(n-2)} \in N(0, 1).$$

При больших n для квантилей распределения Стьюдента справедлива приближенная формула $t_p \approx \frac{u_p}{\sqrt{(1-1/4n)^2 - u_p^2/2n}}$, где u_p - квантиль порядка p стандартного нормального распределения.

2.3. F - распределение (распределение Фишера) или распределение дисперсионного отношения

Третье распределение, часто применяемое при анализе выборочных данных из нормальной совокупности, - это F -распределение. Прежде всего, оно используется в задачах, связанных с дисперсиями.

Если величины U и V независимы и каждая распределена как χ^2 с n_1 и n_2 степенями свободы, то $F = U/n_1 / (V/n_2)$ имеет плотность распределения вероятностей

$$f_F(x) = \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{x^{\frac{n_1}{2}-1}}{\left(1 + \frac{n_1 x}{n_2}\right)^{(n_1+n_2)/2}}, \quad x > 0. \quad (2.3.1)$$

Это двухпараметрическое семейство распределений с параметрами n_1 и n_2 , называемыми степенями свободы. Константа $\Gamma(n_1/2)\Gamma(n_2/2)/\Gamma((n_1 + n_2)/2)$ обозначается как $B(n_1/2, n_2/2)$. Это бета-функция, определяемая формулой

$$B(m, n) = \int_0^1 y^{m-1} (1-y)^{n-1} dy. \quad (2.3.2)$$

Графики трех функций плотностей вероятностей распределения Фишера: с двумя и пятью степенями свободы, с пятью и десятью степенями свободы и, наконец, с десятью и тридцатью степенями свободы, а также функция распределения приведены на рис. 2.4.

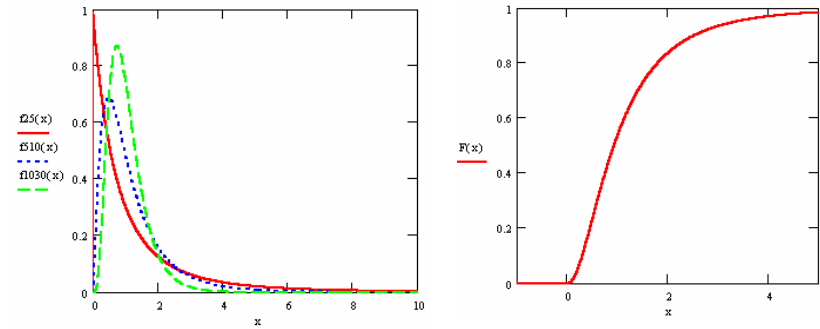


Рис. 2.4. Кривые плотности и распределения F -распределения

Приведем один из возможных выводов формулы (2.3.1). Плотности распределений случайных величин $Y = U/n_1$ и $Z = V/n_2$ выражаются

одинаково и имеют вид $k_Y(y) = \frac{1}{2^{(n_1/2)} \Gamma(n_1/2)} y^{\frac{n_1}{2}-1} e^{-\frac{y}{2}}$,

$k_Z(z) = \frac{1}{2^{(n_2/2)} \Gamma(n_2/2)} z^{\frac{n_2}{2}-1} e^{-\frac{z}{2}}$. Тогда их совместная функция распределения $P(Y/Z < u) = \iint_D k_Y(y) k_Z(z) dy dz$, где $D = \{(y, z) / y > 0, z > 0, y/z < u\}$.

Перейдем от двойного интеграла к повторному:

$$P(Y/Z < u) = \frac{1}{\Gamma(n_1/2) \Gamma(n_2/2) 2^{\frac{n_1+n_2}{2}}} \int_0^\infty dz \int_0^{zu} y^{\frac{n_1}{2}-1} z^{\frac{n_2}{2}-1} e^{-\frac{y}{2}} e^{-\frac{z}{2}} dy. \quad (2.3.3)$$

Произведя в (2.3.3) замену переменной по формуле $t = y/z$, получим

$$P(Y/Z < u) = \left\langle \begin{array}{l} zdt = dy, \\ y = 0, t = 0, \\ y = zu, t = u \end{array} \right\rangle =$$

$$= C_1 \int_0^\infty dz \int_0^u t^{\frac{n_1}{2}-1} z^{\frac{n_1}{2}-1} z^{\frac{n_2}{2}-1} z e^{-\frac{tz}{2}} e^{-\frac{z}{2}} dt =$$

$$= C_1 \int_0^u t^{\frac{n_1}{2}-1} dt \int_0^\infty z^{\frac{n_1+n_2}{2}-1} e^{-\frac{z(t+1)}{2}} dz,$$

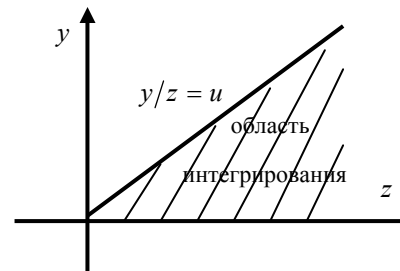


Рис. 2.5. Область существования дроби Фишера

$$C_1 = \frac{1}{\Gamma(n_1/2)\Gamma(n_2/2)2^{\frac{n_1+n_2}{2}}} \text{ (рис. 2.5). Но}$$

$$\int_0^\infty z^{\frac{n_1+n_2}{2}-1} e^{-\frac{z(t+1)}{2}} dz = \left\langle \begin{array}{l} \frac{z(t+1)}{2} = v, z = \frac{2v}{t+1}, \\ dz = \frac{2dv}{t+1}, e^{-\frac{z(t+1)}{2}} = e^{-v}, \\ z^{\frac{n_1+n_2}{2}-1} = \frac{2^{\frac{n_1+n_2}{2}-1}}{(t+1)^{\frac{n_1+n_2}{2}-1}} v^{\frac{n_1+n_2}{2}-1} \end{array} \right\rangle =$$

$$= \int_0^\infty \frac{2^{\frac{n_1+n_2}{2}-1}}{(t+1)^{\frac{n_1+n_2}{2}-1}} v^{\frac{n_1+n_2}{2}-1} e^{-v} \frac{2}{t+1} dv = \frac{2^{\frac{n_1+n_2}{2}}}{(t+1)^{\frac{n_1+n_2}{2}}} \Gamma\left(\frac{n_1+n_2}{2}\right),$$

так как $\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$. Тогда

$$P\left(\frac{Y}{Z} < u\right) = \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \int_0^u t^{\frac{n_1}{2}-1} (t+1)^{-\frac{n_1+n_2}{2}} dt. \quad (2.3.4)$$

Найдем, наконец, функцию плотности вероятности

$$f_F(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \frac{d}{dx} \int_0^{\frac{n_1}{n_2}x} t^{\frac{n_1}{2}-1} (t+1)^{-\frac{n_1+n_2}{2}} dt, \text{ так как}$$

$$u = \frac{y}{z}, \quad x = \frac{n_2 y}{n_1 z} = \frac{n_2}{n_1} u. \quad \text{Отсюда}$$

$$\begin{aligned}
f_F(x) &= \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}x\right)^{\frac{n_1}{2}-1} \left(\frac{n_1}{n_2}x + 1\right)^{-\frac{n_1+n_2}{2}} \frac{n_1}{n_2} = \\
&= \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{x^{\frac{n_1}{2}-1}}{\left(1 + \frac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}}. \quad (2.3.5)
\end{aligned}$$

При дифференцировании под знаком интеграла здесь, как и в предыдущем случае, использована формула

$$\frac{d}{dy} \int_{\alpha(y)}^{\beta(y)} f(x, y) dx = \int_{\alpha(y)}^{\beta(y)} \frac{\partial f(x, y)}{\partial y} dx + \beta'(y) f(\beta(y), y) - \alpha'(y) f(\alpha(y), y).$$

Формула (2.3.4) легко выражается в терминах бета-функции, поэтому вместо таблиц F -распределения можно использовать таблицы бета-функции. Функция плотности вероятности также как у χ^2 -распределения сильно асимметрична.

Числовые характеристики распределения:

1) математическое ожидание $m_x = \frac{n_2}{n_2 - 2}, n_2 > 2;$

2) дисперсия $D_x = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}, n_2 > 4;$

3) мода $d_X = \frac{n_2(n_1 - 2)}{n_1(n_2 + 2)}, n_1 > 1;$

4) коэффициент асимметрии $A = \frac{2(2n_1 + n_2 - 2)}{(n_2 - 6)} \sqrt{\frac{2(n_2 - 4)}{n_1(n_1 + n_2 - 2)}}, n_2 > 6;$

5) коэффициент эксцесса

$$E = \frac{12[(n_2 - 2)^2(n_2 - 4) + n_1(5n_2 - 22)(n_1 + n_2 - 2)]}{n_1(n_1 + n_2 - 2)(n_2 - 6)(n_2 - 8)}, n_2 > 8.$$

2.4. Распределение Колмогорова*

Важную роль в математической статистике играет распределение статистики, введенной А.Н. Колмогоровым:

$$D_n = \sup_{-\infty < x < +\infty} |F_n(x) - F(x)|, \quad (2.4.1)$$

где $F(x)$ - функция распределения случайной величины X , а $F_n(x)$ - эмпирическая функция распределения.

Теорема 2.1. (Колмогорова). Если функция распределения $F(x)$ непрерывна, то

$$\lim_{n \rightarrow \infty} P \left\{ \sqrt{n} \sup_{-\infty < x < +\infty} |F_n(x) - F(x)| < z \right\} = K(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}, \quad z > 0. \quad (2.4.2)$$

График функции распределения Колмогорова (рис. 2.6, справа) имеет ряд особенностей. Функция $K(z)$ очень медленно возрастает в промежутке $z \in [0, 0.5]$, затем очень быстро возрастает почти до единицы на отрезке

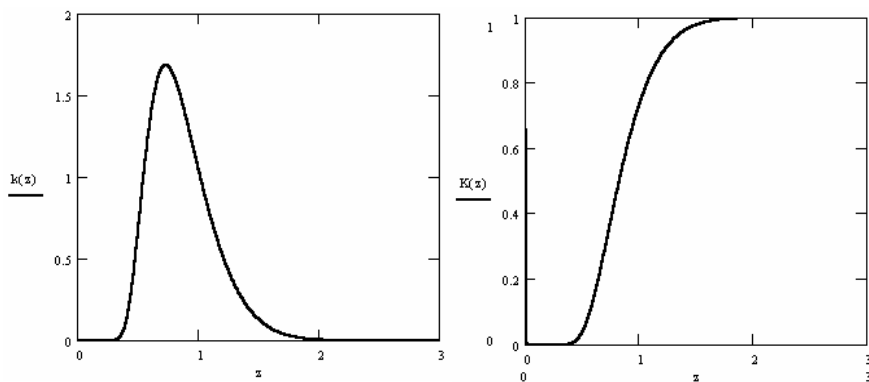


Рис. 2.6. Графики функций плотности вероятности и функции распределения статистики Колмогорова

$z \in [0.5, 1]$, потом следует опять медленный рост при $z \rightarrow \infty$. Найдем функцию плотности распределения Колмогорова

* Андрей Николаевич Колмогоров (1903-1987) - советский математик.

$$\begin{aligned}
f(z) = k(z) = K'_Z(z) &= \left[\sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2} \right]'_z = \left[1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 z^2} \right]'_z = \\
&= 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 z^2} (-2k^2 z) = -8z \sum_{k=1}^{\infty} (-1)^k k^2 e^{-2k^2 z^2} = k(z). \quad (2.4.3)
\end{aligned}$$

График функции плотности вероятности распределения изображен на рис. 2.6, слева. Найдем теперь основные числовые характеристики:

$$m_Z = \int_0^{\infty} zk(z)dz, \quad \text{так как } z > 0 \text{ по определению, тогда}$$

$$m_Z = \int_0^{\infty} z \left(-8z \sum_{k=1}^{\infty} (-1)^k k^2 e^{-2k^2 z^2} \right) dz = -8 \sum_{k=1}^{\infty} (-1)^k k^2 \int_0^{\infty} z^2 e^{-2k^2 z^2} dz. \quad \text{Вы-}$$

числим отдельно

$$\begin{aligned}
\int_0^{\infty} z^2 e^{-2k^2 z^2} dz &= \left\langle \begin{array}{l} z = u, \quad dz = du, \\ \frac{1}{2} e^{-2k^2 z^2} dz^2 = dv, \\ v = \frac{1}{2(-2k^2)} e^{-2k^2 z^2} \end{array} \right\rangle = \frac{z}{-4k^2} e^{-2k^2 z^2} \Big|_0^{\infty} + \frac{1}{4k^2} \int_0^{\infty} e^{-2k^2 z^2} dz = \\
&= \left\langle \begin{array}{l} \sqrt{2}kz = y, \\ dz = \frac{dy}{\sqrt{2}k} \end{array} \right\rangle = \frac{1}{4k^2} \int_0^{\infty} e^{-y^2} \frac{dy}{\sqrt{2}k} = \frac{1}{4\sqrt{2}k^3} \cdot \frac{\sqrt{\pi}}{2}. \quad \text{Здесь использован ин-}
\end{aligned}$$

$$\text{теграл Пуассона} \quad \int_{-\infty}^{\infty} e^{-t^2} dt = 2 \int_0^{\infty} e^{-t^2} dt = \sqrt{\pi}. \quad \text{Тогда}$$

$$\begin{aligned}
m_Z &= -8 \sum_{k=1}^{\infty} (-1)^k k^2 \frac{\sqrt{\pi}}{4\sqrt{2}k^3 \cdot 2} = - \sum_{k=1}^{\infty} (-1)^k \frac{1}{k} = \left\langle \sum_{k=1}^{\infty} (-1)^k \frac{1}{k} \right\rangle = \\
&= \sqrt{\frac{\pi}{2}} \ln 2 = 0.8687.
\end{aligned}$$

Аналогично

$$D_Z = \int_0^{\infty} (z - m_Z)^2 k(z) dz = \int_0^{\infty} \left(z - \ln 2 \sqrt{\frac{\pi}{2}} \right)^2 (-8z) \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 z^2} k^2 dz =$$

$$= -8 \sum_{k=1}^{\infty} (-1)^k k^2 \int_0^{\infty} \left(z - \ln 2 \sqrt{\frac{\pi}{2}} \right)^2 z e^{-2k^2 z^2} dz. \quad \text{Подсчитаем этот интеграл}$$

отдельно, раскрывая скобки и интегрируя по частям:

$$\begin{aligned} 1) \quad \int_0^{\infty} z^3 e^{-2k^2 z^2} dz &= \left\langle \begin{array}{l} z^2 = u, du = 2z dz, \\ \frac{1}{2} e^{-2k^2 z^2} dz^2 = dv, \\ v = -\frac{1}{4k^2} e^{-2k^2 z^2} \end{array} \right\rangle = \\ &= -\frac{z^2}{4k^2} e^{-2k^2 z^2} \Big|_0^{\infty} + \frac{2}{4k^2} \int_0^{\infty} z e^{-2k^2 z^2} dz = \frac{1}{-8k^4} e^{-2k^2 z^2} \Big|_0^{\infty} = \frac{1}{8k^4}; \\ 2) \quad -2 \ln 2 \sqrt{\frac{\pi}{2}} \int_0^{\infty} z^2 e^{-2k^2 z^2} dz &= -2 \ln 2 \sqrt{\frac{\pi}{2}} \frac{\sqrt{\pi}}{4\sqrt{2}k^3 \sqrt{2}} = -\frac{\pi \ln 2}{8k^3}; \\ 3) \quad \frac{\pi \ln^2 2}{2} \int_0^{\infty} z e^{-2k^2 z^2} dz &= \frac{\pi \ln^2 2}{4} \int_0^{\infty} e^{-2k^2 z^2} dz^2 = \\ &= \frac{\pi \ln^2 2}{4} \cdot \frac{1}{-2k^2} e^{-2k^2 z^2} \Big|_0^{\infty} = \frac{\pi \ln^2 2}{8k^2}, \quad \text{тогда} \\ D_Z &= -8 \sum_{k=1}^{\infty} (-1)^k k^2 \left[\frac{1}{8k^4} - \frac{\pi \ln 2}{8k^3} + \frac{\pi \ln^2 2}{8k^2} \right] = -8 \sum_{k=1}^{\infty} (-1)^k \frac{k^2}{8k^4} + 8 \sum_{k=1}^{\infty} (-1)^k \frac{\pi \ln 2}{8k} - \\ &- 8 \sum_{k=1}^{\infty} (-1)^k \frac{k^2 \pi \ln^2 2}{8k^2} = -\sum_{k=1}^{\infty} (-1)^k \frac{1}{k^2} + \pi \ln 2 \sum_{k=1}^{\infty} (-1)^k \frac{1}{k} - \pi \ln^2 2 \sum_{k=1}^{\infty} (-1)^k. \end{aligned}$$

Хотя $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k^2} = -\frac{\pi^2}{12}$, $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k} = -\ln 2$, но ряд $\sum_{k=1}^{\infty} (-1)^k$ расходится. Следовательно, D_Z не существует.

Оценим медиану. По определению $\int_{h_Z}^{\infty} k(z) dz = \frac{1}{2}$. В нашем случае

$$-8 \int_{h_Z}^{\infty} z \sum_{k=1}^{\infty} (-1)^k k^2 e^{-2k^2 z^2} dz = -8 \sum_{k=1}^{\infty} (-1)^k k^2 \int_{h_Z}^{\infty} z e^{-2k^2 z^2} dz =$$

$$= -\frac{8}{2} \sum_{k=1}^{\infty} (-1)^k k^2 \frac{1}{-2k^2} e^{-2k^2 z^2} \Big|_{h_Z}^{\infty} = -2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 h_Z^2} = \frac{1}{2}. \quad \text{Отсюда}$$

$$2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 h_Z^2} + 1 = -\frac{1}{2} + 1, \text{ но } 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 h_Z^2} = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 h_Z^2} =$$

$$= K(h_Z^2) = \frac{1}{2} \quad \text{Тогда } h_Z^2 = K^{-1}\left(\frac{1}{2}\right) = 0.8276. \text{ Итак, медиана распределения}$$

равна $h_Z = 0.9097$.

2.5. Гамма–распределение

Гамма–распределение полезно при представлении распределения величин (вес, длина), которые не могут быть отрицательными или значения которых ограничены снизу известным числом. Как и семейство распределений Вейбулла, семейство гамма–распределений включает экспоненциальное распределение как частный случай.

Гамма–распределение определяется формулой

$$f_X(x) = \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}} x^{\alpha} e^{-\frac{x}{\beta}}, \quad x > 0. \quad (2.5.1)$$

Распределение двухпараметрическое. Параметр масштаба $\beta > 0$, часто используется другой параметр λ , $\lambda = 1/\beta$. При $\alpha = 0$ уравнение (2.5.1) дает функцию плотности вероятности экспоненциального распределения.

$$f(x) = \frac{1}{\Gamma(1)\beta} e^{-\frac{x}{\beta}} = \lambda e^{-\lambda x}, \quad \lambda = \frac{1}{\beta}. \quad (2.5.2)$$

α в распределении должно быть больше -1 .

$$\text{Функция распределения } F(x) = \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}} \int_0^x t^{\alpha} e^{-\frac{t}{\beta}} dt. \quad (2.5.3)$$

Вид функции плотности вероятности и функции распределения сильно зависят от параметра α . На рис. 2.7 приведены графики функций плотностей (слева) и функций распределения (справа) для значений параметра α , равного нулю (для f_1 и F_1), двум (для f_2 и F_2) и семи (для f_3 и F_3).

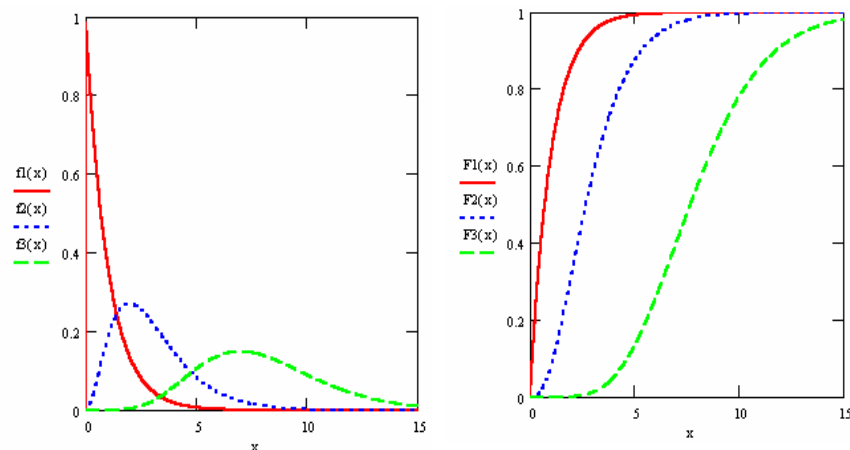


Рис. 2.7. Кривые плотности и распределения гамма-распределения

Числовые характеристики распределения:

- 1) математическое ожидание $m_X = \beta(\alpha + 1)$;
- 2) дисперсия $D_X = \beta^2(\alpha + 1)$;
- 3) мода $d_X = \beta\alpha$, $\alpha \geq -1$;
- 4) коэффициент асимметрии $A = 2/\sqrt{\alpha + 1}$;
- 5) коэффициент эксцесса $E = 6/(\alpha + 1)$.

При $\alpha + 1 = n/2$ и $\beta = 2$ гамма-распределение совпадает с χ^2 -распределением с n степенями свободы. Сумма двух независимых случайных величин, имеющих гамма-распределение с параметрами $\alpha_1 + 1$ и $\alpha_2 + 1$ соответственно, имеет гамма-распределение с параметром $\alpha_1 + \alpha_2 + 2$.

2.6. Распределение Вейбулла (Вейбулла – Гнеденко*)

Это распределение весьма широко применяется в последние два десятилетия. Особенно оно полезно в задачах долговечности и надежности. Его можно рассматривать как обобщение экспоненциального распределения, поскольку в нем три параметра, и оно сводится к экспоненциальному при подходящем выборе одного из них.

* Борис Владимирович Гнеденко (1912-1995) - советский математик.

Плотность распределения вероятностей смещенного (трехпараметрического) распределения Вейбулла определяется как

$$f_X(x) = \frac{c}{b} \left(\frac{x-a}{b} \right)^{c-1} e^{-\left(\frac{x-a}{b} \right)^c}, \quad x \geq a, \quad b > 0, \quad c > 0. \quad (2.6.1)$$

Параметр масштаба b иногда называют характерным временем жизни. Обычно распределение делают двухпараметрическим, полагая $a = 0$. Тогда получают так называемое классическое распределение Вейбулла с плотностью вероятности

$$f_X(x) = \begin{cases} \alpha \lambda x^{\alpha-1} e^{-\lambda x^\alpha}, & x > 0, \\ 0, & x \leq 0. \end{cases} \quad (2.6.2)$$

Здесь $\alpha, \lambda > 0$ и $\alpha = c$, $\lambda = 1/b^c$, а параметр α называется параметром формы. При $\alpha = 0$ и $\alpha = 1$ уравнение (2.6.2) превращается в функцию плотности экспоненциального распределения. Обычно проще работать с функцией распределения Вейбулла, которая имеет вид

$$F(x) = 1 - e^{-\lambda x^\alpha}. \quad (2.6.3)$$

Графики функции распределения и функции плотности вероятности зависят от значения параметра формы α (или c в формуле (2.6.1)). Далее на рис. 2.8 приведены графики функций плотности вероятностей и функ-

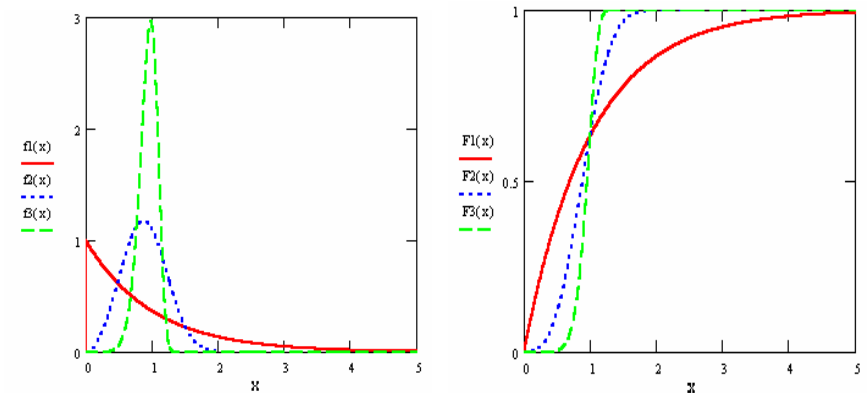


Рис. 2.8. Кривые плотности и распределения закона Вейбулла

ций распределения для ряда значений параметра формы: $f1$ и $F1$ для $\alpha = 1$, $f2$ и $F2$ для $\alpha = 3$, наконец, $f3$ и $F3$ для $\alpha = 8$.

Числовые характеристики двухпараметрического распределения выражаются через гамма-функцию и равны:

1) математическое ожидание $m_X = b\Gamma\left(\frac{c+1}{c}\right) = \lambda^{-\frac{1}{\alpha}}\Gamma\left(\frac{\alpha+1}{\alpha}\right)$;

2) дисперсия

$$D_X = b^2 \left[\Gamma\left(\frac{c+2}{c}\right) - \Gamma^2\left(\frac{c+1}{c}\right) \right] = \lambda^{-\frac{2}{\alpha}} \left[\Gamma\left(\frac{\alpha+2}{\alpha}\right) - \Gamma^2\left(\frac{\alpha+1}{\alpha}\right) \right];$$

3) мода $d_X = \begin{cases} b\left(1 - \frac{1}{c}\right)^{\frac{1}{c}}, & c \geq 1, \text{ или } d_X = \lambda^{-\frac{1}{\alpha}}\left(1 - \frac{1}{\alpha}\right)^{\frac{1}{\alpha}}, & \alpha \geq 1, \\ 0, & c < 1, \quad 0, & \alpha < 1; \end{cases}$

4) медиана $h_X = b(\ln 2)^{1/c}$;

5) коэффициент асимметрии

$$A = \frac{\Gamma\left(\frac{c+3}{c}\right) - 3\Gamma\left(\frac{c+2}{c}\right)\Gamma\left(\frac{c+1}{c}\right) + 2\Gamma^3\left(\frac{c+1}{c}\right)}{\left[\Gamma\left(\frac{c+2}{c}\right) - \Gamma^2\left(\frac{c+1}{c}\right)\right]^{3/2}};$$

6) коэффициент эксцесса

$$E = \frac{\Gamma\left(\frac{c+4}{c}\right) - 4\Gamma\left(\frac{c+3}{c}\right)\Gamma\left(\frac{c+1}{c}\right) + 6\Gamma\left(\frac{c+2}{c}\right)\Gamma^2\left(\frac{c+1}{c}\right) - 3\Gamma^4\left(\frac{c+1}{c}\right)}{\left[\Gamma\left(\frac{c+2}{c}\right) - \Gamma^2\left(\frac{c+1}{c}\right)\right]^2}.$$

Распределение Вейбулла часто используется в теории надежности для описания времени безотказной работы приборов.

2.7. Лабораторная работа № 2. Семейства вероятностных распределений в математических пакетах STATGRAPHICS и MANTCAD

Пакет STATGRAPHICS Plus for Windows предоставляет возможность работать с 22 наиболее распространенными распределениями вероятностей: Бернулли, биномиальным, дискретным равномерным, геометрическим, отрицательным биномиальным, Пуассона, бета-распределением, распределением χ^2 -квадрат, Эрланга*, экспоненциальным, распределением экстремального значения, F -распределением (распределением дисперсионного отношения), гамма-распределением, распределением Лапласа,

* Агнер Крауп Эрланг (1878-1929) – датский математик.

логистическим распределением, логнормальным, нормальным, распределением Парето*, Стьюдента, треугольным, равномерным и распределением Вейбулла.

Для доступа к процедурам, работающими с распределениями, в главном меню пакета необходимо выбрать пункт Plot→Probability Distribution (Графики→Распределение вероятностей). Появится дополнительное меню, содержащее все 22 перечисленных распределения. По умолчанию выделено распределение № 17 – нормальное. Рассмотрим, например, геометрическое распределение. Для этого отметим его в меню распределений и щелкнем по кнопке ОК; появится заставка распределения вероятностей со своим дополнительным меню, содержащим четыре пункта: Input Dialog (Диалог ввода), Tabular Options (Табличные процедуры), Graphics Options (Графические процедуры) и Save Results (Сохранение результата).

Для ввода параметров исследуемого распределения необходимо щелкнуть правой кнопкой мыши в любом месте заставки распределения вероятностей. Появится дополнительное (контекстное) меню следующего вида (рис. 2.9). Назначение процедур этого меню таково.

1. Pane Options (Панель процедур) позволяет задать объем генерируемой выборки выбранного распределения, уровни квантилей и тому подобное в зависимости от контекста вычислений.

2. Analysis Options (Процедуры анализа) позволяет задать параметры выбранного закона распределения. Это либо характеристики положения и рассеивания для непрерывных распределений, либо вероятности событий для дискретных. Пакет позволяет моделировать и выводить на экран до пяти однотипных распределений.

3. Print распечатывает содержимое текущего проекта или его части.

4. Copy to Gallery копирует содержимое текущего проекта в специальный инструмент – StatGallery. Он позволяет расположить в одном окне или на одном листе до девяти различных фрагментов текста и графических иллюстраций. Это часто бывает необходимым для составления отчетной документации.

Для рассматриваемого геометрического распределения выберем пункт Analysis Options. Появится новая закладка следующего вида (рис. 2.10).

Введем в поля ввода следующие вероятности: 0.05, 0.25, 0.5, 0.75 и 0.95. После щелчка левой

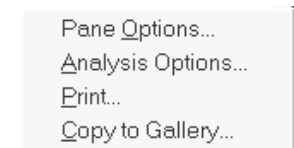


Рис. 2.9. Контекстное меню

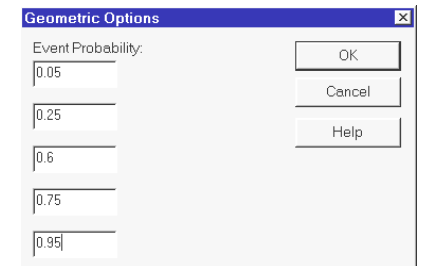


Рис. 2.10. Меню для задания параметров геометрического распределения

* Вильфридо Парето (1848-1923)-итальянский экономист и социолог.

кнопкой мыши по кнопке ОК эти данные заносятся в поле заставки геометрического распределения.

Формы представления выбранного распределения можно задать с помощью двух пунктов меню заставки геометрического распределения вероятностей Tabular Options и Graphics Options. Меню Tabular Options имеет следующий вид (рис. 2.11). Процедуры этого меню при задании пунктов

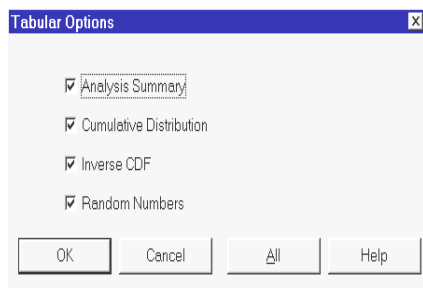


Рис. 2.11. Панель табличных параметров в анализе распределения вероятностей

следующим образом. Для одного значения переменной (по умолчанию $x_0 = 0$) вычисляются вероятности левого (Lower Tail Area) и правого (Upper Tail Area) хвостов распределения, т.е. вероятности $P(X < x_0)$ и $P(X > x_0)$, и вероятность (для дискретных распределений) или значение функции плотности (для непрерывных) при $x = x_0$. Переменную x_0 можно изменить или задать до пяти ее различных значений, щелкнув правой кнопкой мыши в поле заставки Cumulative Distribution и выбрав пункт меню Pane Options.

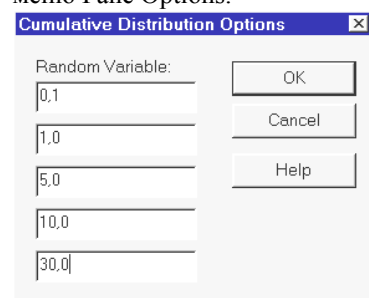


Рис. 2.12. Меню задания квантилей или процентных точек

Появится дополнительное подменю следующего вида (рис. 2.12). Введем в поля ввода этого меню такие значения: 0.1, 1.0, 5.0, 10.0 и 30.0.

Inverse CDF (Обратная функция распределения, значения процентилей) вычисляет для заданных на заставке геометрического распределения вероятностей 0.05, 0.25, 0.5, 0.75 и 0.95 процент наблюдений, лежащих левее указанного числа. Можно задать и иные значения. Для этого необходимо щелчком правой кнопки мыши вызвать пункт меню Pane Options и в появившейся заставке задать набор из пяти нужных числовых значений.

Random Numbers (Случайные числа) порождает последовательность

независимых одинаково распределенных случайных чисел, подчиняющихся выбранному распределению – одному из упомянутых двадцати двух.

В пункте Graphics Options (рис. 2.13) можно построить графики следующих функций.

Density/Mass Function – функция плотности вероятности для непрерывных распределений или графическое изображение ряда распределения для дискретных распределений. Графики функций выдаются с соответствующими заголовками и автоматически оцифровываются. Если на график выводится несколько кривых, то они обозначаются различными типами линий – непрерывной, пунктирной, точечной и другими. Справа от графика указывается легенда – связь между типами линий и параметрами кривых, выводимых на график.

CDF (Cumulative Distribution Function – функция распределения). Первые две формы представления распределений из этого меню являются наиболее употребительными. Последние три еще не использовались нами на практике.

Survivor Function (Функция выживаемости) равна единице минус функция распределения. Это хорошо видно при сравнении графиков обеих функций.

Log Survivor Function (Логарифм функции выживаемости). Имеется в виду натуральный логарифм этой функции.

Hazard Function (Функция риска). Функцией риска называется частное от деления плотности распределения на функцию выживаемости.

После нажатия кнопки ОК меню Graphical Options на правой половине экрана монитора выводятся графики всех заданных функций. Двойной щелчок левой кнопки мыши на любом графике или заставке разворачивает и сворачивает их на весь экран. Заполним теперь лист StatGallery и оформим отчет о проделанной лабораторной работе. В любом месте открытого поля StatGallery щелкнем правой кнопкой мыши. Появится дополнительное меню (рис. 2.14). Это меню задает порядок и форму расположения текстовой и графической информации на листе отчета. Назначение процедур этого меню следующее:

1. Modify Arrangement (Задание классификации) вызывает подменю StatGallery Options, распределяющее рисунки и

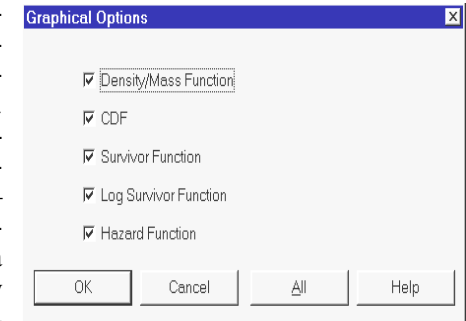


Рис. 2.13. Панель графических параметров в анализе распределения вероятностей



Рис. 2.14. Дополнительное меню StatGallery

текст различным образом в поле StatGallery (рис. 2.15). С его помощью информацию на листе отчета можно расположить девятью способами.

2. Print – распечатывает содержимое StatGallery.

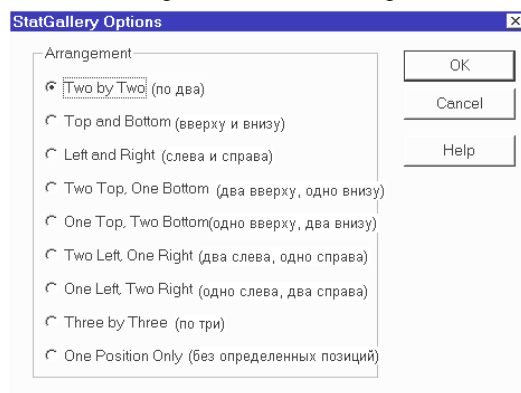


Рис. 2.15. Меню StatGallery, распределяющее рисунки и текст на листе отчета

3. Erase Graph – стирает часть информации из поля отчета. Вызываемое подменю имеет строение, аналогичное предыдущему подменю.

4. Move Graph – перемещает части отчета по полю StatGallery. При этом можно заменить предыдущую информацию новой (Replace), или затереть ее наложением (Overlay).

5. Clear Gallery – очищает поле StatGallery.

Вызовем пункт меню StatGallery Options и зададим расположение шести элементов отчета способом «по три». В отчет включим заставки Probability Distributions (Top Left), Cumulative Distributions (Center Left), Inverse CDF (Bottom Left) и графики трех функций Probability (Top Center), Cumulative Probability (Center) и Log Survivor Probability (Bottom Center). В результате получим лист отчета (рис. 2.16). К сожалению, более трех элементов в строку и столбец поля StatGallery поместить нельзя.

Воспользуемся, наконец, пунктом дополнительного меню заставки распределения вероятностей Save Results и сохраним полученные результаты. Появится дополнительное подменю следующего вида (рис. 2.17). В полях Save везде поставим галочки, а в полях Target Variables (Плановые переменные) наберем имена случайных выборок Geom1,..., Geom5.

После нажатия клавиш File→Save Data File As и набора имени Geom сгенерированные выборки будут помещены в базу данных пакета STATGRAPHICS.

В пакете реализовано уникальное средство для сохранения результатов работы и создания собственных статистических проектов. Все, что пользователь считает ценным в своем варианте анализа (методы, параметры статистических процедур, графика, табличные схемы и так далее) можно сохранить в виде нового файла StatFolio. Затем этот файл по мере надобности можно изменять и дополнять, используя многократно. Сохраним и мы результаты нашей работы, выбрав File→Save StatFolio As→Geometric.

Probability Distributions					

Distribution: Geometric					
Parameters:		Event prob.			
Dist. 1	0,05				
Dist. 2	0,25				
Dist. 3	0,6				
Dist. 4	0,75				
Dist. 5	0,95				
Cumulative Distribution					

Distribution: Geometric					
Lower Tail Area (<)					
Variable	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
0,1	0,0975	0,4375	0,84	0,9375	0,9975
1	0,05	0,25	0,6	0,75	0,95
5	0,226219	0,762695	0,98976	0,999023	1,0
10	0,401263	0,943686	0,999895	0,999999	1,0
30	0,785361	0,999821	1,0	1,0	1,0
Probability Mass (=)					
Variable	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
0,1	0,0	0,0	0,0	0,0	0,0
1	0,0475	0,1875	0,24	0,1875	0,0475
5	0,038689	0,0593262	0,006144	0,000732422	2,96875E-7
10	0,0299368	0,0140784	0,0000629146	7,15256E-7	9,27734E-14
30	0,0107319	0,0000446455	6,91753E-13	6,50521E-19	8,84756E-40
Upper Tail Area (>)					
Variable	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
0,1	0,9025	0,5625	0,16	0,0625	0,0025
1	0,9025	0,5625	0,16	0,0625	0,0025
5	0,735092	0,177979	0,004096	0,000244141	1,5625E-8
10	0,5688	0,0422351	0,000041943	2,38419E-7	4,88498E-15
30	0,203907	0,000133937	4,61187E-13	0,0	0,0
Inverse CDF					

Distribution: Geometric					
CDF	Dist. 1	Dist. 2	Dist. 3	Dist. 4	Dist. 5
0,01	0	0	0	0	0
0,1	2	0	0	0	0
0,5	13	2	0	0	0
0,9	44	8	2	1	0
0,99	89	16	5	3	1
Random Numbers					

To generate random numbers from the selected distribution, use the save button on the analysis toolbar.					
Random numbers to be generated: 200					

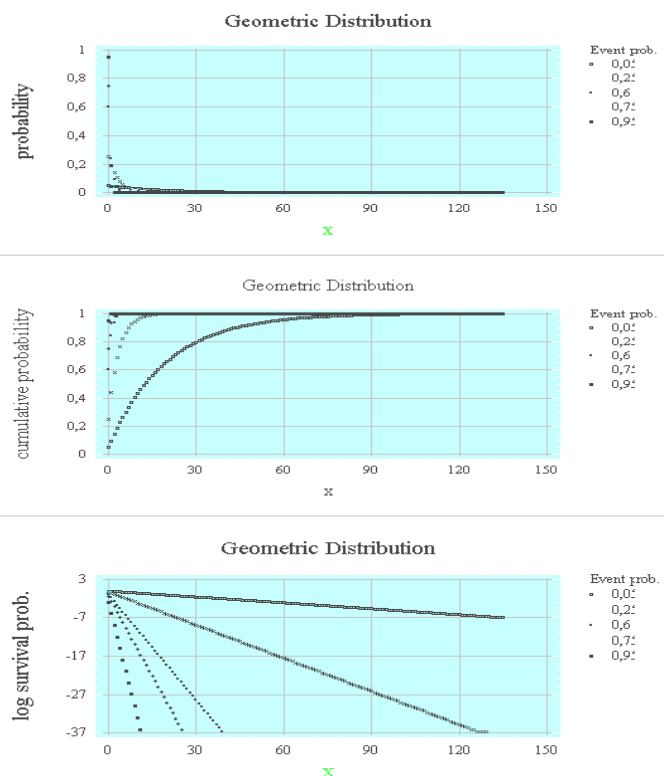


Рис. 2.16. Некоторые табличные и графические характеристики геометрического распределения

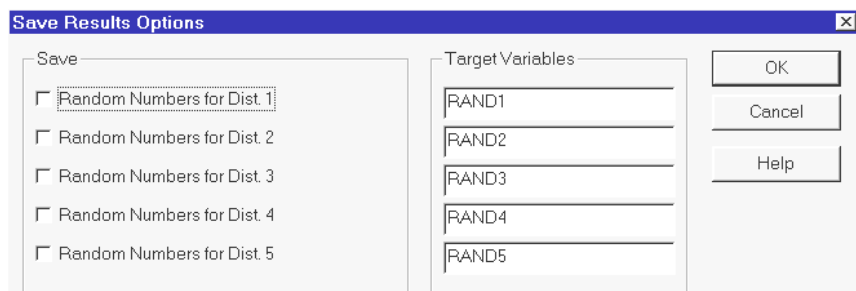


Рис. 2.17. Меню пункта Save Results

Задание № 1. По номеру вашей фамилии в журнале преподавателя выбрать распределение в пункте Probability Distributions (если номер больше 22, выбирать номер минус 15), задать пять однотипных распределений, варьируя параметры выбранного распределения, вычислить и вывести на экран дисплея все пункты меню Tabular Options и Graphical Options. Объем выборок задать равным 50. Полученные результаты записать на лист отчета в StatGallery и сохранить в личном статистическом проекте под оригинальным именем (имя: LAB2№группыФИО).

В пакете MATHCAD статистические функции условно разделяются на четыре раздела: статистики совокупностей, распределения вероятностей, гистограммы и случайные числа. Пакет MATHCAD имеет более широкую область применения, статистические функции составляют лишь одну из его более чем двадцати глав. В связи с этим статистические процедуры не сопровождаются непосредственно связанной с ними графикой. Результаты каждой процедуры требуют поэтому индивидуального графического отображения.

Приведем перечень и дадим описание основным встроенным статистическим функциям пакета MATHCAD.

1. Статистики совокупностей. MATHCAD содержит всего шесть функций для вычисления статистических оценок случайных совокупностей.

mean(A) - возвращает среднее значение элементов массива A размерности $m \times n$ по формуле

$$\text{mean}(A) = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} A_{i,j}; \quad (2.7.1)$$

median(A) - возвращает медиану. Медианой называется величина, больше и меньше которой в вариационном ряду расположено одинаковое число членов ряда. Если число элементов массива A четно, то медиана определяется как среднее арифметическое двух центральных элементов вариационного ряда;

var(A) - вычисляет дисперсию элементов массива A ;

cvar(A,B) - возвращает ковариацию по формуле

$$\text{cvar}(A, B) = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [A_{i,j} - \text{mean}(A)] [\overline{B_{i,j} - \text{mean}(B)}], \quad (2.7.2)$$

где черта указывает комплексно-сопряженную величину;

stdev(A) - вычисляет среднее квадратическое отклонение

$$\text{stdev}(A) = \sqrt{\text{var}(A)} ;$$

наконец, функция **corr(A, B)** находит коэффициент корреляции для двух массивов A и B .

2. Распределение вероятностей и случайные числа. В пакете MATH-CAD можно работать лишь с шестнадцатью статистическими распределениями: биномиальным, геометрическим, отрицательным биномиальным, Пуассона, бета-распределением, распределением χ^2 -квадрат, экспоненциальным, F -распределением, гамма-распределением, логистическим, логнормальным, нормальным, распределением Стюдента, равномерным, распределениями Вейбулла и Коши.

Функции плотности вероятности, функции распределения, обращения функций распределения и функции, моделирующие случайные числа, имеют похожие названия, отличающиеся лишь первой буквой в имени. Эти функции и их параметры приведены в табл. 1.

Т а б л и ц а 1

Номер по порядку и название распределения	Функции плотности вероятности	Функции распределения	Обратные функции распределения (значения процентиля)	Функции, генерирующие случайные числа
1. Биномиальное	dbinom(k,n,p)	pbinom(k,n,p)	qbinom(α ,n,p), где $P(X \leq x) = \alpha$	rbinom (m,n,p), где m- число случайных чисел
2. Геометрическое	dgeom(k,p)	pgeom(k,p)	qgeom(α ,p)	rgeom(m,p)
3. Отрицательное биномиальное	dnbinom (k,n,p)	pnbinom (k,n,p)	onbinom (α ,n,p)	rnbinom (m,n,p)
4. Пуассона	dpois(k, λ)	ppois(k, λ)	qpois(α , λ)	rpois(m, λ)
5. Бета – распределение	dbeta(x,n1,n2)	pbeta(x,n1,n2)	qbeta(α ,n1,n2)	rbeta(m,n1,n2)
6. χ^2 -квадрат распределение	dchisq(x,n)	pchisq(x,n)	qchisq(α ,n)	rchisq(m,n)

Окончание табл. 1

Номер по порядку и название распределения	Функции плотности вероятности	Функции распределения	Обратные функции распределения (значения процентилей)	Функции, генерирующие случайные числа
7. Экспоненциальное	dexp(x,λ)	rexpr(x,λ)	qexp(α,λ)	rexpr(m,λ)
8. F - распределение	dF(x,n1,n2)	pF(x,n1,n2)	qF(α,n1,n2)	rF(m,n1,n2)
9. Гамма – распределение	dgamma(x,n)	pgamma(x,n)	qgamma(α,n)	rgamma(m,n)
10. Логистическое	dlogis(x,l,n)	plogis(x,l,n)	qlogis(α,l,n)	rlogis(m,l,n)
11. Логнормальное	dlnorm(x,μ,σ)	plnorm(x,μ,σ)	qlnorm(α,μ,σ)	rlnorm(m,μ,σ)
12. Нормальное	dnorm(x,μ,σ)	pnorm(x,μ,σ), cnorm(x)= pnorm(x,0,1)	qnorm(α,μ,σ)	rnorm(m,μ,σ)
13. Стьюдента	dt(x,n)	pt(x,n)	qt(α,n)	rt(m,n)
14. Равномерное	dunif(x,a,b)	punif(x,a,b)	qunif(α,a,b)	runif(m,a,b), rnd(x)= runif(1,0,x)
15. Вейбулла	dweibull(x,n)	pweibull(x,n)	qweibull(α,n)	rweibull(m,n)
16. Коши	dcauchy(x,l,n)	pcauchy(x,l,n)	qcauchy(α,l,n)	rcauchy(m,l,n)

Функции, вырабатывающие случайные числа, генерируют псевдослучайные последовательности (подробнее см. разд. 3). Эти последовательности зависят от некоторого целого числа, называемого стартовым значением. Для изменения стартового значения в пункте меню пакета MATHCAD Математика надо выбрать Генератор случайных чисел и ввести необходимое целое число.

3. Для вычисления частотного распределения и построения гистограмм в пакете MATHCAD имеется одна функция **hist(int,A)**. Она возвращает вектор, представляющий частоты, с которыми величины, содержащиеся в векторе A , попадают в интервалы, представляемые вектором **int**. Элементы массива **int** должны быть упорядочены по возрастанию. Возвращаемый результат – вектор, содержащий на один элемент меньше, чем **int**. Его элементы – частоты f_i есть числа $n(A)$ значений в массиве A , удовлетворяющих условию $\text{int}_i < n(A) < \text{int}_{i+1}$.

В качестве примера в пакете MATHCAD рассмотрим моделирование распределения Коши. Это распределение имеет плотность вероятности

$$f(x) = \frac{\lambda}{\pi[\lambda^2 + (x - \mu)^2]}, \quad -\infty < x < \infty, \quad (2.7.3)$$

где μ - параметр положения (медиана), $\lambda > 0$ - параметр рассеивания (срединное отклонение). Математического ожидания и моментов распределе- ние не имеет.

Наберем в пакете MATHCAD следующую программу:

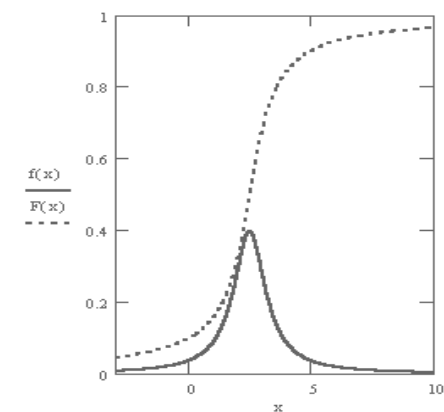
ORIGIN := 1 $\mu := 2.5$ $\lambda := 0.8$

$x1 := \text{rcauchy}(50, \mu, \lambda)$ $x1 := \text{sort}(x1)$

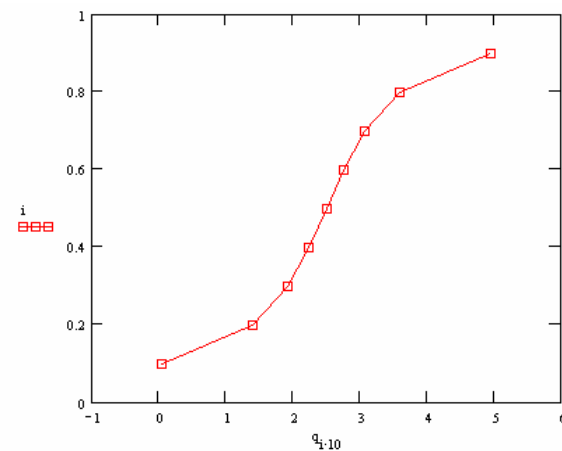
$f(x) := \text{dcauchy}(x, \mu, \lambda)$ $F(x) := \text{pcauchy}(x, \mu, \lambda)$

	1
1	-0.643
2	-0.118
3	-0.029
4	0.246
5	0.606
6	0.962
7	0.988
8	1.225
9	1.276
10	1.322
11	1.577
12	1.663
13	1.919
14	1.941
15	1.975

$x1 =$



$i := 0.1, 0.2, \dots, 1$ $q_{i*10} := \text{qcauchy}(i, \mu, \lambda)$

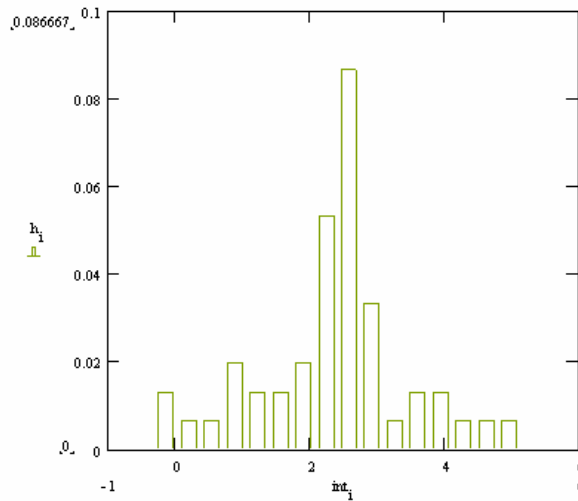


$m := \text{mean}(x1)$
 $m = 2.506$
 $med := \text{median}(x1)$
 $med = 2.654$
 $D := \text{var}(x1)$
 $D = 1.723$
 $\sigma := \text{stdev}(x1)$
 $\sigma = 1.313$
 $xmin := \text{min}(x1)$
 $xmax := \text{max}(x1)$
 $xmin = -0.643$
 $xmax = 6.159$
 $E := 0.477 \cdot \sqrt{2} \cdot \sigma$
 $E = 0.885$
 $R := xmax - xmin$

```

R = 6.802  m := 20  L :=  $\frac{R}{m}$ 
i := 1...20  inti := xmin +  $\frac{L}{2} \cdot (2 \cdot i - 1)$ 
h := hist(int, x1)
i := 1...20  hi :=  $\frac{h_i}{50}$ 

```



Значение E , равное срединному отклонению, - это параметр рассеивания распределения Коши. В нашем случае он должен быть близок к заданному значению λ . Следует отметить, что функции **mean** и **var** по формуле (2.7.1) и формуле, аналогичной (2.7.2), формально находят числовые значения

смоделированной выборки, которые, однако, вовсе не являются математическим ожиданием и дисперсией распределения Коши.

Итак, на примере данного распределения выполнен примерно такой же объем работ и построены те же графики, которые в пакете STAT-GRAPHICS получаются автоматически заданием лишь соответствующего пункта меню.

Задание № 2. По номеру вашей фамилии в журнале преподавателя выбрать распределение из табл. 1 (если номер больше 16, выбирать номер минус 15), самостоятельно задать все необходимые параметры распределения; построить графики функции плотности вероятности, функции распределения, функции значений процентиля, смоделировать случайную выборку данного распределения длиной в 100 единиц, вычислить числовые характеристики выборки и построить ее гистограмму.

3. МЕТОД СТАТИСТИЧЕСКИХ ИСПЫТАНИЙ (МЕТОД МОНТЕ-КАРЛО)

3.1. Общие принципы метода статистических испытаний

Различают физическое и математическое моделирование. При физическом моделировании модель воспроизводит изучаемую систему с сохранением ее физической природы. Классическим примером физического моделирования является продувка масштабных моделей летательных аппаратов в аэродинамических трубах. Более широкие возможности предоставляет математическое моделирование. При изучении любой системы этим методом необходимо построить ее математическую модель. Как правило, реальная система находится под воздействием случайных факторов или сам механизм функционирования содержит элементы случайности. Математическая модель, содержащая случайные элементы, называется вероятностной моделью.

Существуют различные пути исследования вероятностной модели: 1) аналитическое исследование; 2) аналитическое исследование с применением численных методов; 3) аппаратное моделирование; 4) статистическое моделирование.

Непосредственное экспериментальное изучение сложных случайных явлений часто требует чрезмерно больших затрат средств и времени. Тогда прибегают к статистическому моделированию изучаемых явлений. Современная вычислительная техника дает возможность имитировать практически без ограничений сложнейшие явления и процессы. Это привело к созданию метода статистического моделирования как научного метода исследования, позволяющего сочетать теоретические расчеты с имитацией различных экспериментов.

Свою историю статистическое моделирование начинает от метода Монте-Карло, предложенного фон Нейманом^{*} и Уламом^{**} в 1940 году для решения детерминированных задач с помощью случайных величин, имитируемых на ЭВМ. Метод Монте-Карло, или метод статистических испытаний, относится к классу статистических и служит для получения некоторых сведений о распределении случайной величины X после получения ряда ее реализаций, т.е. решает типичную задачу математической статистики, которая изучает методы оценки параметров распределений случайной величины на основе ее реализаций.

Метод статистических испытаний применяют для решения не только тех задач, в которых в явном виде имеются случайные явления, но также и

^{*} Джон фон Нейман (1903-1957) - американский математик.

^{**} Станислав Марцин Улам (1909-1984) – американский математик

для решения многих математических задач, не содержащих таких явлений. В этом случае искусственно подбирают такое случайное явление, характеристики которого связаны с результатом решения исходной задачи. Для определения числовых значений этих характеристик используется метод статистических испытаний.

В задаче оценки математического ожидания, например, по методу Монте-Карло традиционной оценкой является среднее арифметическое.

Идея метода статистических испытаний основана на законе больших чисел. Наиболее простая вычислительная схема при этом заключается в следующем. Если решается задача оценки среднего значения некоторой случайной величины, то вычисляются N независимых реализаций случайной величины и ее математическое ожидание (среднее значение) оценивается с помощью среднего арифметического этих реализаций. Оценка погрешности может быть получена, например, по неравенству Чебышева*

$$P\left(\left|\frac{S_n}{N} - a\right| \leq \varepsilon\right) \geq 1 - \frac{\sigma^2}{N\varepsilon^2} \quad (3.1.1)$$

и имеет вероятностный характер. Если положить $\gamma = \frac{\sigma^2}{N\varepsilon^2}$, то получим

$$P\left(\left|\frac{1}{N} \sum_{i=1}^N X_i - MX\right| \leq \frac{\sigma}{(N\gamma)^{\frac{1}{2}}}\right) > 1 - \gamma, \quad (3.1.2)$$

где γ - малая величина. С вероятностью $1 - \gamma$ среднее арифметическое отличается от MX не более чем на $\sigma/\sqrt{N\gamma}$, т.е. погрешность убывает как $1/\sqrt{N}$. Также может быть оценено и N , гарантирующее необходимую точность с заданной вероятностью.

Моделирование на ЭВМ случайных элементов подчиняется двум основным принципам:

1) сходство между случайным элементом - оригиналом и его моделью на ЭВМ состоит в совпадении или близости вероятностных законов распределения или числовых характеристик;

2) всякий случайный элемент конструируется как некоторая борелевская функция от простейших, так называемых базовых случайных величин.

* Пафнутий Львович Чебышев (1821-1894) – русский математик и механик.

3.2. Датчики базовой случайной величины (БСВ)

Датчик БСВ - устройство, позволяющее по запросу получить реализацию x или несколько независимых реализаций x_1, x_2, \dots, x_n базовой случайной величины X . Существуют три типа датчиков: табличные, физические и программные.

Табличный датчик БСВ - это таблица случайных чисел, представляющая собой экспериментально полученную выборку реализаций равномерно распределенной на промежутке $[0, 1]$ случайной величины. Можно заранее составить таблицы этих значений, используя, например, физические генераторы. Существуют и готовые таблицы случайных чисел. Однако при расчетах на ЭВМ такие таблицы, как правило, не используются. Их хранение во внутренней памяти ЭВМ обычно невозможно вследствие ее загруженности информацией, относящейся непосредственно к задаче.

Физический датчик БСВ - специальное радиоэлектронное устройство, служащее приставкой к ЭВМ. Оно состоит из источника флуктуационного шума (например, флуктуационно шумящей радиолампы), значение которого в произвольный момент времени является случайной величиной $X \geq 0$ с плотностью вероятности $f_X(x)$. В качестве физического датчика БСВ может быть использован и источник радиоактивного распада. Счетчик подсчитывает количество радиоактивных частиц за некоторое время Δt . Если число частиц четное, то в разряд посылается единица, если нечетное, то ноль. При параллельной работе k генераторов будет получено значение k - разрядной двоичной дроби. Время Δt выбирается таким, чтобы вероятность получения в разряде единицы, так же как и вероятность получения нуля, была равна 0.5. Недостатки физического датчика БСВ: невозможность повторения некоторой ранее полученной реализации x ; схемная нестабильность, приводящая к необходимости контроля работы датчика при очередном его использовании.

Программный датчик БСВ (псевдослучайные последовательности чисел). Возможен и следующий подход к моделированию случайной величины X , имеющей равномерное распределение на $[0, 1]$. Эту величину получают с помощью некоторой рекуррентной формулы, причем X обладает статистическими свойствами, близкими к свойствам равномерного распределения на отрезке $[0, 1]$. Полученную последовательность чисел называют псевдослучайной.

Один из первых методов получения псевдослучайной последовательности чисел был предложен Джоном фон Нейманом. Он называется методом середины квадратов.

Возьмем некоторое число r_0 . Пусть $r_0 = 0.9876$. Возведем его в квадрат $r_0^2 = 0.97535376$. Выберем четыре средние цифры этого числа и положим $r_1 = 0.5353$. Затем возводим r_1 в квадрат: $r_1^2 = 0.28654609$ и снова выбираем четыре средние цифры. Получаем $r_2 = 0.6546$ и так далее.

Метод псевдослучайных последовательностей прост и экономичен; на получение каждого числа затрачивается всего несколько простых операций. Однако он имеет ряд существенных недостатков, главный из которых - трудность теоретической оценки статистических свойств псевдослучайной последовательности. Помимо этого, числа, входящие в вырабатываемую программным способом псевдослучайную последовательность, зависимы между собой, а сама последовательность является периодической, так как в ЭВМ может быть представлено только конечное число различных чисел и повторное появление какого-либо числа и всех последующих за ним чисел неизбежно.

3.3. Моделирование на ЭВМ стандартной равномерно распределенной случайной величины (базовой случайной величины)

Рассмотрим случайную величину

$$\sum_{i=1}^{n-1} \frac{\beta_i}{2^i} + \frac{\beta_n}{2^n}, \quad (3.3.1)$$

где $\beta_i, i = 1, 2, \dots, n$ независимы в совокупности, $\beta_i, i = 1, 2, \dots, n-1$ - дискретные случайные величины, принимающие значения 0, 1 с равными вероятностями, т.е. $P(\beta_i = 0) = P(\beta_i = 1) = 1/2$, а β_n равномерно распределена на $[0, 1]$. Путем довольно несложных преобразований методом математической индукции доказывается, что

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{n-1} \frac{\beta_i}{2^i} \xrightarrow{P} R[0, 1] \quad (3.3.2)$$

Таким образом, идея подобного моделирования сводится к тому, что реализации дискретной случайной величины - индикатора с $P(X = 0) = P(X = 1) = 1/2$ являются двоичными цифрами случайной величины $\alpha \in R[0, 1]$, т.е. α представима в виде $\alpha = 0.\beta_1\beta_2\beta_3\dots$. Если речь идет о равномерном распределении на $[0, 1]$, то имеется в виду реализация α с бесконечным числом значащих цифр. Это предполагает либо сколь угодно большую точность отсчета, либо бесконечное число реализаций n . Эта задача решается приближенно, решение в любом случае не тривиально.

Подавляющее большинство алгоритмов работает на основе рекуррентных соображений следующим путем. Пусть $x_{i+k} = f(x_{i+k-1}, x_{i+k-2}, \dots, x_i)$, где k фиксировано, $i = 1, 2, \dots$, f - целочисленная функция целочисленных аргументов; x_0, x_1, \dots, x_{k-1} - целые (пусковые) константы - задаются. Например, f может быть задана в виде

$$x_{i+k} = \left(\sum_{j=0}^{k-1} b_j x_{i+j} + \theta \right) \bmod P, \quad (3.3.3)$$

где b_j и θ - целые числа, x_{i+k} - остаток от деления целочисленной линейной функции $\sum_{j=0}^{k-1} b_j x_{i+j} + \theta$ на число P . Это так называемый метод

Леметра* или метод сравнений. Если $k = 1$ и $\theta = 0$, то $x_{i+1} = b_0 x_i$. Это мультипликативный метод. Имеется обширная литература, в которой обсуждаются свойства последовательностей, полученных таким методом. Опишем более подробно суть мультипликативного конгруэнтного метода (метода вычетов).

Псевдослучайная последовательность вычисляется по рекуррентным формулам

$$x_i = \frac{x_i^*}{P}, \quad x_i^* = (\beta x_{i-1}^*) \bmod P, \quad i = 1, 2, \dots, \quad (3.3.4)$$

где β, P, x_0^* - параметры программного датчика; β - множитель; P - модуль; x_0^* - стартовое значение. Операция $y = (z) \bmod P$ означает

$y = z - P \cdot \left\lfloor \frac{z}{P} \right\rfloor$, $\left\lfloor \dots \right\rfloor$ - операция деления нацело. Тогда неотрицательные

числа $x_0^*, x_1^*, \dots \in \{0, 1, 2, \dots, P-1\}$. Отсюда следует, что 1) последовательность $\{x_i^*\}$, а значит и $\{x_i\}$ всегда «зацикливается», т.е. начиная с некоторого номера $i \geq i_0$ образуется цикл, повторяющийся бесконечное число раз; 2) период последовательности $T \leq P-1$.

Параметры β, P и x_0^* определяются таким образом, чтобы величина T была максимальной. Наиболее распространены три варианта выбора P :

* Жорж Леметр (1894-1966) - бельгийский математик, физик и астроном.

1) $P = 2^q$, где q - число двоичных разрядов, используемых для задания целой константы в ЭВМ. Например, $P = 2^{31} = 2147483648$. Так как $T \leq P-1$, то целесообразно модуль P выбирать максимально возможным;

2) $P = 10^q$;

3) P - простое число.

3.4. Моделирование дискретной случайной величины при помощи случайных событий

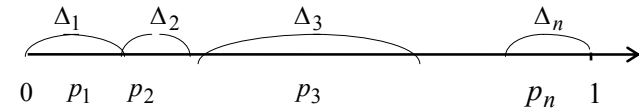
Если случайная величина дискретна, то ее моделирование можно свести к моделированию независимых испытаний. Пусть имеется следующий ряд распределения

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

Обозначим событие $A_i = (X = x_i)$. Тогда нахождение значения, принятого случайной величиной X , сводится к определению того, какое из событий A_1, A_2, \dots, A_n появится. События A_i несовместны и образуют полную группу событий, и для их моделирования можно использовать следующую процедуру.

Пусть в результате k независимых испытаний может произойти одно из двух противоположных событий A и $B = \bar{A}$. Известно, что $P(A) = p$, $P(B) = 1 - p$. Построим последовательность значений r_1, r_2, \dots, r_k случайной величины $R \in R[0, 1]$. Если $r_i < p$, $i = 1, 2, \dots, k$, то считаем, что в i -м испытании наступило событие A . Если же $r_i > p$, то считаем, что произошло событие B . Это действительно так, ибо $P(R < p) = P(0 < R < p) = P(A)$ и $P(p < R < 1) = P(R > p) = P(B)$.

Разделим теперь отрезок $[0, 1]$ на n участков $\Delta_1, \Delta_2, \dots, \Delta_n$, длины которых соответственно равны p_1, p_2, \dots, p_n . Получаем как и прежде последовательность значений r_1, r_2, \dots, r_k случайной величины R . Если $r_i \in \Delta_m$, то считаем, что в i -м испытании наступило событие A_m , так как $P(R \in \Delta_m) = \text{длине отрезка } \Delta_m = p_m = P(A_m)$.



Помимо описанного общего алгоритма моделирования дискретной случайной величины для многих законов существуют специальные алгоритмы. Рассмотрим два примера: моделирование биномиального и пуассоновского распределений.

Моделирование случайной величины с биномиальным распределением.

$$\text{Если } Y \in B(n, p), \text{ то } P_n(m) = C_n^m p^m (1-p)^{n-m}, \quad (3.4.1)$$

где p - вероятность появления события в каждом отдельно взятом испытании, n - число испытаний. Введем индикатор - случайную величину X_i , показывающую, появилось или нет интересующее нас событие в i -м испытании. Величина X_i , очевидно, может принимать только два значения: либо 1 с вероятностью p , либо 0 с вероятностью $1-p$. Итак,

X_i	1	0
P	p	$1-p$

ряд распределения индикатора. Тогда в n испытаниях интересующее нас событие появится m раз $m = X_1 + X_2 + \dots + X_n$, где m будет очередным значением случайной величины X , распределенной биномиально с параметрами n, p .

Итак, определение значения случайной величины $Y = m$ сводится к следующей процедуре:

1) получают последовательность значений r_1, r_2, \dots, r_n случайной величины $R \in R[0, 1]$;

2) для каждого числа $r_i, i = 1, 2, \dots, n$ проверяют, выполняется ли неравенство $r_i < p$. Если неравенство выполняется, то полагают $X_i = 1$, в противном случае считают $X_i = 0$;

3) находят сумму значений n случайных величин X_i , т.е.

$$Y = m = \sum_{i=1}^n X_i ;$$

4) процедуру повторяют необходимое число раз (п. 1-3), получают последовательность значений m_1, m_2, m_3, \dots случайной величины $X \in B(n, p)$. Описанный метод называется методом браковки. В вычислительной практике часто используется метод, базирующийся на геометрическом распределении.

Моделирование случайной величины, распределенной по закону Пуассона. Распределение Пуассона $P(\lambda)$ - предельная форма биномиального распределения $B(n, p)$ при $n \rightarrow \infty$, $p \rightarrow 0$, $np \rightarrow \lambda$, т.е.

$$\lim_{\substack{n \rightarrow \infty, \\ np \rightarrow \lambda}} C_n^m p^m (1-p)^{n-m} = \frac{\lambda^m e^{-\lambda}}{m!}. \quad (3.4.2)$$

Алгоритм моделирования случайной величины, распределенной по закону Пуассона, учитывая (3.4.2), может быть, например, таков:

1) задать λ и выбрать n такое, чтобы вероятность $p = \lambda/n$ была достаточно малой ($p < 0.01$);

2) получить последовательность значений r_1, r_2, \dots, r_n случайной величины $R \in R[0, 1]$;

3) для каждого числа r_i , $i = 1, 2, \dots, n$ проверить выполнение неравенства $r_i < p$. Если оно выполняется, то $X_i = 1$, в противном случае считают $X_i = 0$;

4) вычислить $Y = \sum_{i=1}^n X_i$. Это и есть значение случайной величины $Y \in P(\lambda)$;

5) п. 1-4 повторить требуемое число раз.

3.5. Моделирование непрерывных случайных величин

Для моделирования непрерывных случайных величин разработано несколько общих методов. Рассмотрим некоторые из них.

Метод обратной функции основан на теореме Смирнова*.

Теорема 3.1. (Смирнова). Если X удовлетворяет уравнению

$$\int_{-\infty}^X dF_X(t) = \alpha, \text{ т.е.}$$

* Николай Васильевич Смирнов (1900-1966) - советский математик.

$$X = F_X^{-1}(\alpha), \quad (3.5.1)$$

где α - величина, распределенная равномерно на $[0, 1]$, то X распределено по закону $F_X(t)$.

Действительно, введем случайную величину $\alpha = F_X(t)$, обратим внимание на то, что так как $0 \leq F_X \leq 1$, то и $0 \leq \alpha \leq 1$. Найдем функцию распределения случайной величины α :

$$F_\alpha(z) = P(\alpha < z) = \begin{cases} 0, & z \leq 0, \\ P\{F_X(t) < z\} = P\{t < F_X^{-1}(z)\} = F_X\{F_X^{-1}(z)\} = \alpha, & 0 \leq \alpha \leq 1, \\ 1, & z \geq 1. \end{cases}$$

Таким образом, α - случайная величина, имеющая равномерное распределение на отрезке $[0, 1]$.

Порядок действия при моделировании конкретного распределения методом обратной функции следующий:

1) разыгрывается реализация α равномерной случайной величины $\alpha \in R[0, 1]$;

2) решается уравнение $F(x) = \alpha$. Его решение $X = F^{-1}(\alpha)$ дает случайную величину X с заданным законом распределения $F(x)$. Здесь F^{-1} функция обратная к $F(x)$. Во многих случаях удается найти явное выражение для $F^{-1}(\alpha)$. Тогда моделирование происходит наиболее просто.

Недостатком описанного метода являются аналитические трудности при вычислении F^{-1} . В «чистом виде» метод обратной функции используется редко на практике, так как для многих распределений, например, нормального, даже $F(x)$ не выражается через элементарные функции, а табулирование $F^{-1}(\alpha)$ существенно усложняет моделирование. На практике метод обратной функции дополняют аппроксимацией $F(x)$ или сочетают с другими методами.

Метод суммирования основан на центральной предельной теореме (ЦПТ). Например, в простейшем случае ЦПТ может быть сформулирована так.

Теорема 3.2. Если случайные величины X_1, X_2, \dots, X_n независимы, одинаково распределены и их математические ожидания и дисперсии конечны, то при увеличении n закон распределения суммы $X_1 + X_2 + \dots + X_n$ неограниченно приближается к нормальному.

Практически оказывается, что для получения хорошего приближения к нормальному распределению достаточно сравнительно небольшого чис-

ла слагаемых. Пусть r_1, r_2, \dots, r_n - независимые случайные величины $r_i \in R[0, 1]$. Обозначим через Y сумму этих величин: $Y = r_1 + r_2 + \dots + r_n$, тогда $Mr_i = 0.5$ и $Dr_i = 1/12$, $i = 1, 2, \dots, n$. Отсюда $MY = 0.5n$, $DY = n/12$. При достаточно большом n по ЦПТ можно считать, что Y имеет нормальный закон распределения с математическим ожиданием $MY = 0.5n$ и дисперсией $DY = n/12$, т.е.

$$Y \in N(0.5n, \sqrt{n/12}). \quad (3.5.2)$$

Перейдем к стандартной нормально распределенной случайной величине $U = \frac{Y - MY}{\sqrt{DY}} = (Y - 0.5n) \sqrt{\frac{12}{n}} = (Y - 0.5n) \frac{6}{\sqrt{3n}}$, $U \in N(0, 1)$. Например, при $n = 12$

$$X = MX + \sigma_X U = MX + \sigma_X \left(\sum_{i=1}^n r_i - 0.5n \right) \frac{6}{\sqrt{3n}} = MX + \sigma_X \left(\sum_{i=1}^{12} r_i - 6 \right). \quad (3.5.3)$$

Тогда $X \in N(MX, \sigma_X)$, а $r_i \in R[0, 1]$. Таким образом, имея двенадцать значений случайной величины R , получаем значение нормальной случайной величины X , имея следующие двенадцать значений R , получаем следующее значение X и так далее.

Для других распределений основной принцип метода суммирования и его опора на ЦПТ остаются без изменения. Моделирующие формулы типа (3.5.3) для каждого конкретного распределения разные.

Рассмотрим еще один пример моделирования гамма-распределения.

Его плотность $f(x) = \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}} x^\alpha e^{-\frac{x}{\beta}}$, $x > 0$. Положим $\lambda = \frac{1}{\beta}$, тогда

$$f(x) = \frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)} x^\alpha e^{-\lambda x}, x > 0. \quad (3.5.4)$$

Предлагаемый способ моделирования гамма-распределения основывается на следующей теореме.

Теорема 3.3. Если y_1, y_2, \dots, y_n - независимые стандартные экспоненциально распределенные случайные величины, то $x = \sum_{i=1}^n y_i$ имеет гамма-распределение с параметром $\alpha = n - 1$.

Если y - экспоненциально распределенная случайная величина, то $f(y) = \lambda e^{-\lambda y}$, $y \geq 0$, λ - параметр экспоненциального распределения. Для моделирования экспоненциального распределения применим метод обратной функции.

$$F(y) = \lambda \int_0^y e^{-\lambda t} dt = 1 - e^{-\lambda y}, \quad F(y) = \gamma \Rightarrow 1 - e^{-\lambda y} = \gamma, \quad \text{где } \gamma \in R[0, 1]$$

$1 - \gamma = e^{-\lambda y}$, $-\lambda y = \ln(1 - \gamma) = \ln \gamma$, так как если $\gamma \in R[0, 1]$, то и $(1 - \gamma) \in R[0, 1]$.

$$\text{Итак,} \quad y = -\frac{1}{\lambda} \ln \gamma \quad (3.5.5)$$

формула, моделирующая показательное распределение. По теореме 3.3

$$x = \sum_{i=1}^{\alpha+1} y_i = \sum_{i=1}^{\alpha+1} \left(-\frac{1}{\lambda} \right) \ln \gamma_i = -\frac{1}{\lambda} \sum_{i=1}^{\alpha+1} \ln \gamma_i = -\frac{1}{\lambda} \ln \left(\prod_{i=1}^{\alpha+1} \gamma_i \right). \quad (3.5.6)$$

Формула (3.5.6) справедлива, если α - целое. В случае дробного α формулу (3.5.6) модернизируют следующим образом:

$$\begin{cases} x = -\frac{1}{\lambda} \ln \left(\prod_{i=1}^N \gamma_i \right) + x^*, & \alpha + 1 = N + 0.5, \\ x^* = -\frac{1}{\lambda} \ln \gamma_{N+1} \cdot \cos^2(2\pi \gamma_{N+1}). \end{cases} \quad (3.5.7)$$

При $0 < \alpha + 1 < 1$ существуют модификации рассмотренного метода. Для значений параметра $(\alpha + 1) > 1$, можно в комбинации с указанной техникой использовать свойство аддитивности гамма-распределения.

3.6. Лабораторная работа № 3. Моделирование некоторых распределений с помощью базовых случайных величин в пакете MATHCAD

В качестве базового распределения для моделирования многих случайных величин используется стандартное равномерное распределение $R[0, 1]$. Само равномерное распределение может быть получено несколькими способами с помощью линейных, нелинейных и смешанных формул метода сравнений (см. подразд. 3.3). Рассмотрим формулу

$$x_{n+1} = (bx_n + \theta) \bmod P \quad (3.6.1)$$

частный случай формулы (3.3.3), вырабатывающую последовательность целых чисел, равномерно распределенных в области $\{0,1,2,\dots,P-1\}$. Эта формула используется в программе URAND (Universal RANDom number generator) [22]. Как уже упоминалось ранее, пусковые константы b, θ и P должны выбираться практически таким образом, чтобы вырабатываемая случайная последовательность наиболее полно отвечала требуемым вероятностным законам распределения и числовым характеристикам этих законов.

Для формулы (3.6.1) в [11] следующим образом подытожены теоретико-числовые ограничения на выбор b, θ и P :

$$\left\{ \begin{array}{l} 1) \ (b) \bmod 8 = 5, \\ 2) \ \frac{P}{100} < b < P - \sqrt{P}, \\ 3) \ \frac{\theta}{P} \approx \frac{1}{2} - \frac{1}{6}\sqrt{3} \approx 0.21132, \ \theta - \text{нечетное число.} \end{array} \right. \quad (3.6.2)$$

Следует заметить, что формула (3.6.1) использует целые числа, которые в разных ЭВМ имеют разную длину и хранятся по-разному. При выполнении арифметических операций с целыми числами большое значение имеют особенности машинной арифметики [26]. В пакете MATHCAD реально приходится оперировать с вещественными числами, поэтому приведенная ниже программа Urand, взятая из [22] и адаптированная в пакете MATHCAD, нуждается в подробном тестировании, которое автором не проводилось.

Программа реализует формулу (3.6.1) для $P = 2^{31}$ и предназначена для вычисления псевдослучайного числа, имитирующего реализацию случайной величины со стандартным равномерным распределением $R[0,1]$.

ORIGIN := 1

iy := -1023 t := runif(100,0,1)

```
MassUrand(iy,k) :=
| for i ∈ 1..k
|   | u1 ← Urand(iy)
|   | iy ← u1 · 2147483648
|   u
```

```

Urand(iy) := m2 ← 1073741824
             halfm ← m2
             mic ← halfm · atan(1.0)
             ib ← mod(mic, 8)
             ib ← mic - ib + 5
             mic ← halfm ·  $\left(1 - \frac{1}{\sqrt{3}}\right)$ 
             iθ ← mod(mic, 2)
             iθ ← mic - iθ + 1
             s ←  $\frac{0.5}{\text{halfm}}$ 
             iy ← iy · ib + iθ
             iy ← mod(iy, 2147483648)
             iy ← (iy + m2) + m2 if iy < 0
             iy ← (iy - m2) - m2 if  $\frac{\text{iy}}{2} > m2$ 
             while iy > 2147483648
                 iy ← iy - 2147483648
             u ← iy · s
             u

```

	1
1	$1.268419437 \cdot 10^{-3}$
2	0.19332302
3	0.585006099
4	0.350308103
5	0.822837725
6	0.174128995
t = 7	0.710495408
8	0.303986053
9	0.091411268
10	0.147313412
11	0.988508474
12	0.119079732
13	$8.922664449 \cdot 10^{-3}$
14	0.531664159

$k := \text{Urand}(iy) \quad k = 0.48016216$
 $k1 := \text{Urand}(k) \quad k1 = 0.399884106$
 $d := \text{MassUrand}(iy, 100)$

Приведем в заключение еще одну программу, вычисляющую псевдо-случайные числа со стандартным равномерным распределением, и входящую в типовое математическое обеспечение ЭВМ фирмы IBM и называемую Randu.

```

Randu(ix,k) := for i ∈ 1..k
  iy ← ix - 65539
  iy ← iy + 2147483647 + 1 if iy < 0
  while iy > 2147483648
    iy ← iy - 2147483648
  randi ←  $\frac{iy}{2147483648}$ 
  ix ← iy
rand

```

	1
1	0.48016216
2	0.041056693
3	0.224624626
4	0.749138683
5	0.603225827
6	0.842525303
7	0.306550622
8	0.381240189
9	0.269118547
10	0.637852132
11	0.517660022
12	0.525602877
13	0.487882137
14	0.011663139
15	0.543287391

$ix := 19510 \quad u := \text{Randu}(ix,1) \quad u = 0.595$

$u1 := \text{Randu}(ix,100)$

	1
1	0.595425204
2	0.572469461
3	0.075989925
4	0.303714405
5	0.138377101
6	0.096832964
7	0.335603871
8	0.142126556
9	0.832324491
10	0.714807945
11	0.797927254
12	0.354292016
13	0.944406808
14	0.47781271
15	0.367214986

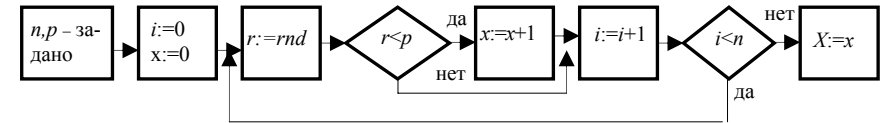
Программа Randu реализует ту же формулу (3.6.1), где $b = 65539$, $\theta = 0$ и $P = 2^{31}$. Анализ свойств получаемых ею псевдослучайных чисел приведен в работе [22]. Там показано, что в последовательности, вырабатываемой программой Randu, наблюдается крайне высокая корреляция между тремя подряд идущими случайными числами.

Итак, при моделировании стандартного равномерного распределения в пакете MATHCAD при проведении студенческих лабораторных работ можно пользоваться двумя встроенными функциями пакета **runif(m,a,b)** и **rnd(x)**, а также двумя приведенными подпрограммами.

Другие распределения из списка лабораторной работы № 2 моделируются чаще всего двумя самыми распространенными способами: методом обратной функции и методом суммирования. Далее приведен список распределений и моделирующий каждое конкретное распределение алгоритм.

1. Биномиальное распределение. Метод браковки, описанный в подразд. 3.4 – стандартный способ имитационного моделирования дис-

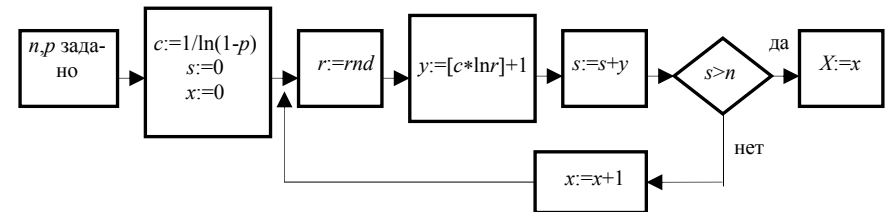
кретной случайной величины. По блок-схеме этого метода, имеющей очень простой вид, последовательно выполняют следующие действия:



- задают n и p по схеме Бернулли;
- получают очередное стандартное равномерно распределенное число r ;
- количество чисел r , которые меньше p , есть случайное число $B(n, p)$, распределенное биномиально.

Использование геометрического распределения. Если p мало, то метод, который работает быстрее чем метод браковки, состоит в суммировании геометрически распределенных случайных чисел до тех пор, пока их сумма не превзойдет n . Количество слагаемых минус единица и есть биномиальное случайное число $B(n, p) = k - 1$, где k - минимальное число, такое, что $\sum_{i=1}^k y_i > n$. y_i - геометрически распределенное случайное число.

Блок-схема этого метода такова:



2. Геометрическое распределение. Случайные числа $G(p)$ получают из случайных чисел, распределенных равномерно на $[0,1]$, с помощью соотношения

$$x_i = \left\lceil \frac{\ln r_i}{\ln(1-p)} \right\rceil, \quad (3.6.3)$$

где $\lceil \dots \rceil$ - целая часть числа. Алгоритм основан на следующей теореме.

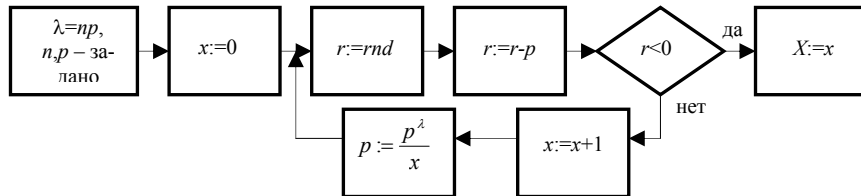
Теорема 3.4. Если $r \in R[0,1]$ - базовая случайная величина, то случайная величина $x = \lceil \ln r / \ln(1-p) \rceil$, где $\lceil x \rceil$ - целая часть x , имеет распределение $P(X = x) = p(1-p)^x$, $x = 0, 1, 2, \dots$

Доказательство

Так как для $r \in R[0,1]$ по определению $F_r(x) = P(r < x) = \begin{cases} 0, & x \leq 0, \\ x, & 0 < x < 1, \\ 1, & x \geq 1, \end{cases}$

$$\begin{aligned} P(X = x) &= P\left\{x \leq \frac{\ln r}{\ln(1-p)} < x+1\right\} = P\{x \ln(1-p) \geq \ln r > (x+1) \ln(1-p)\} = \\ &= P\{\ln(1-p)^{x+1} < \ln r \leq \ln(1-p)^x\} = P\{(1-p)^{x+1} < r \leq (1-p)^x\} = \\ &= (1-p)^x - (1-p)^{x+1} = (1-p)^x(1-1+p) = p(1-p)^x = pq^x. \end{aligned}$$

3. Распределение Пуассона. Первый способ моделирования описан в подразд. 3.4 и употребляется когда p мало. Его блок-схема такова.



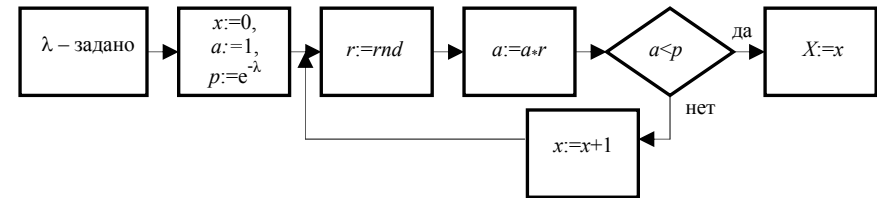
- 1) Задают n , p и $\lambda = np$ так, чтобы $n \rightarrow \infty$, $p \rightarrow 0$ ($p < 0.01$).
- 2) Получают очередное стандартное равномерно распределенное число r .
- 3) Количество чисел r , меньших p , есть случайное число $P(\lambda)$, распределенное по закону Пуассона.

Второй способ основан на связи пуассоновского с показательным и эрланговским распределениями и базируется на теореме 3.5.

Теорема 3.5. Случайная величина x , определенная соотношением

$$x = \min \left\{ N : \prod_{k=1}^{N+1} r_k < e^{-\lambda}, N = 0, 1, 2, \dots \right\} \text{ распределена по закону Пуассона.}$$

Блок-схема алгоритма выглядит следующим образом.



4. Экспоненциальное распределение. Самый распространенный метод моделирования экспоненциального распределения – метод обратной функции, описанный в подразд. 3.5. Моделирующая формула имеет вид

$$x_i = -\frac{1}{\lambda} \ln r_i, \quad r_i \in R[0,1]. \quad (3.6.4)$$

5. Классическое распределение Вейбулла. Распределение Вейбулла с параметром $\alpha = 1$ совпадает с экспоненциальным распределением со средним $b = (1/\lambda)^{1/c}$, поэтому моделируется также методом обратной функции. Действительно, если $F(x) = 1 - e^{-\lambda x^\alpha} = r$, $r \in R[0,1]$, то $1 - r = e^{-\lambda x^\alpha}$, $r = e^{-\lambda x^\alpha}$, $-\lambda x^\alpha = \ln r$, $x^\alpha = -1/\lambda \ln r$, $x = (-1/\lambda \ln r)^{1/\alpha}$. Таким образом, моделирующая формула имеет следующий вид:

$$x_i = (-1/\lambda \ln r_i)^{1/\alpha}, \quad r_i \in R[0,1]. \quad (3.6.5)$$

6. Распределение Парето. Функция плотности вероятности этого распределения равна $f(x) = \alpha/x_0 (x_0/x)^{\alpha+1}$, $x > x_0$, где x_0 - параметр положения, левая граница области возможных значений ($x_0 > 0$), α - параметр формы ($\alpha > 0$). Функция распределения $F(x) = 1 - (x_0/x)^\alpha$, $x > x_0$ легко обращается, в результате получается моделирующая формула

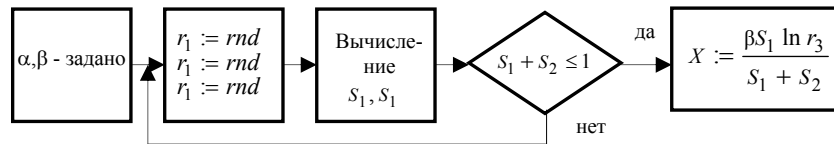
$$x_i = x_0 (1/r_i)^{1/\alpha}, \quad r_i \in R[0,1]. \quad (3.6.6)$$

7. Распределение Эрланга. Распределение Эрланга $(\alpha + 1)$ -го порядка – это гамма-распределение с целым параметром α (см. формулу (2.5.1)). Графики и формулы оценивания параметров для гамма-распределения применимы также и для распределения Эрланга. Распределение моделируется методом суммирования по теореме 3.3. Основная моделирующая формула имеет вид

$$x_i = -\beta \ln \left(\prod_{i=1}^{\alpha+1} r_i \right), \quad r_i \in R[0,1]. \quad (3.6.7)$$

8. Гамма-распределение. I. В том случае, когда α - целое число, моделирование происходит по формуле (3.6.7).

II. Если α - не целое, может быть использован следующий алгоритм. Сначала для $-1 < \alpha < 0$ выберем r_1, r_2 и r_3 - три независимые случайные величины, равномерно распределенные на отрезке $[0,1]$. Положим $S_1 = r_1^{1/(\alpha+1)}$, $S_2 = r_2^{1/\alpha}$. Если $S_1 + S_2 > 1$, возьмем вместо r_1, r_2 другую пару таких же случайных величин. Так будем поступать до тех пор, пока не получим $S_1 + S_2 \leq 1$. В этом случае случайная величина $x = \beta S_1 \ln r_3 / (S_1 + S_2)$ будет иметь гамма-распределение с параметрами α и β . Блок-схема этого алгоритма имеет вид



Наконец, для параметра формы $\alpha > 0$ псевдослучайные числа, подчиненные гамма-распределению, дает формула $y = x - \beta \ln \left(\prod_{i=1}^m r_i \right)$, где случайные числа x имеют гамма-распределение с параметрами β и $\alpha_1 = \alpha - [\alpha]$, $m = [\alpha + 1]$. Здесь $[...]$ - целая часть числа.

Итак, в случае нецелого α моделирующие формулы для гамма-распределения выглядят следующим образом.

$$\left\{ \begin{array}{l} \text{для } -1 < \alpha < 0 \quad S_1 = r_1^{1/\alpha+1}, \quad S_2 = r_2^{1/\alpha}, \\ \quad \text{если } S_1 + S_2 \leq 1, \quad x_i = \frac{\beta S_1 \ln r_3}{S_1 + S_2}, \\ \\ \text{для } \alpha_1 > 0 \quad y_i = x_i - \beta \ln \left(\prod_{i=1}^m r_i \right), \\ \alpha_1 = \alpha - [\alpha], \quad m = [\alpha + 1], \quad r_i \in R[0,1]. \end{array} \right. \quad (3.6.8)$$

9. Распределение Коши. Функция плотности вероятности этого распределения приведена в лабораторной работе № 2 (см. формулу (2.7.3)).

Функция распределения имеет вид $F(x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arctg} \frac{x - \mu}{\lambda}$. Методом обращения из нее легко получается моделирующая формула

$$x_i = \lambda \operatorname{tg} \left[\pi \left(r_i - \frac{1}{2} \right) \right] + \mu, \quad r_i \in R[0,1]. \quad (3.6.9)$$

10. Нормальное распределение. Наиболее употребительный метод моделирования нормального распределения – метод суммирования. Моделирующая формула (3.5.3) $x_i = m_x + \sigma_x \left(\sum_{i=1}^{12} r_i - 6 \right)$, $r_i \in R[0,1]$, дает случайную величину, распределенную нормально с математическим ожиданием $M[X] = m_x$ и дисперсией $D[X] = \sigma^2$. Для моделирования стандартной нормальной случайной величины применяется формула

$$x_i = \sum_{i=1}^{12} r_i - 6. \quad (3.6.10)$$

11. Логарифмически нормальное (логнормальное) распределение. Логнормальное распределение имеет плотность вероятности равную $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\ln^2(x/m)/2\sigma^2}$, где m - параметр масштаба (медиана), σ - параметр формы. Если случайная величина X распределена по логнормальному закону с параметрами m, σ , то случайная величина $Y = \ln X$ подчиняется нормальному закону распределения с математическим ожиданием $\mu = \ln m$ и средним квадратическим отклонением σ .

Так как $X = e^Y$, то моделирующая формула, очевидно, имеет вид

$$x_i = m \exp \left[\sigma \left(\sum_{i=1}^{12} r_i - 6 \right) \right], \quad r_i \in R[0,1]. \quad (3.6.11)$$

12. χ^2 - распределение подробно описано в подразд. 2.1. Для генерирования случайных чисел используются несколько алгоритмов. Самый распространенный алгоритм опирается на определение χ^2 - распределения и реализуется формулой

$$x_i = \sum_{j=1}^n u_j^2, \quad u_j \in N(0,1). \quad (3.6.12)$$

Величины x_i , полученные по формуле (3.6.12), имеют χ^2 - распределение с n степенями свободы.

13. t - распределение Стьюдента. Распределение Стьюдента связано с многими распределениями и может быть аппроксимировано ими при соответствующих значениях числа степеней свободы. Например, при $n=1$ оно совпадает с распределением Коши с параметром положения $\mu=0$ и параметром масштаба $\lambda=1$. При $n \rightarrow \infty$ распределение Стьюдента сходится к стандартному нормальному распределению. При произвольном n для генерирования случайных чисел используется статистика (2.2.1). Тогда $t_n = u / \sqrt{\chi_n^2 / n}$, $u \in N(0,1)$, $u = \sum_{i=1}^{12} r_i - 6$, $\chi_n^2 = \sum_{i=1}^n u_i^2$, $r_i \in R[0,1]$. Моделирующая формула выглядит следующим образом:

$$\begin{cases} t_i = \frac{u_i}{\sqrt{\frac{1}{n} \sum_{j=1}^n u_j^2}}, \\ u_i = \sum_{k=1}^{12} r_k - 6, \quad r_k \in R[0,1]. \end{cases} \quad (3.6.13)$$

14. F -распределение. Как и в двух предыдущих случаях моделирование псевдослучайных чисел основано на определении основной статистики F -распределения. Случайная величина, имеющая распределение Фишера, связана с независимыми случайными величинами $\chi_{n_1}^2$ и $\chi_{n_2}^2$ следующим соотношением $x_{n_1, n_2} = \chi_{n_1}^2 \cdot n_2 / (\chi_{n_2}^2 \cdot n_1)$. Тогда моделирующая формула имеет вид

$$\begin{cases} x_{n_1, n_2} = \frac{n_2 \cdot \sum_{i=1}^{n_1} u_i^2}{n_1 \cdot \sum_{i=1}^{n_2} u_i^2}, \\ u_i = \sum_{j=1}^{12} r_j - 6, \quad r_j \in R[0,1] \end{cases} \quad (3.6.14)$$

15. Логистическое распределение. Функция плотности вероятности этого распределения равна $f(x) = \frac{e^{(x-\mu)/\lambda}}{\lambda(1 + e^{(x-\mu)/\lambda})^2}$, где μ - параметр положения ($-\infty < \mu < \infty$), а λ - параметр масштаба ($\lambda > 0$). Функция распреде-

ления $F(x) = 1 / (1 + e^{-(x-\mu)/\lambda})$. После несложных преобразований методом обращения получается следующая моделирующая формула

$$x_i = \mu - \lambda \ln \frac{1-r_i}{r_i}, \quad r_i \in R[0,1]. \quad (3.6.15)$$

Задание № 1. По номеру фамилии студента в журнале преподавателя (если номер больше 15, считать номер минус 15) выбрать одно из рассмотренных пятнадцати распределений и смоделировать по соответствующим формулам выборку псевдослучайных чисел объемом 100 единиц. Построить график этой выборки. Для выборок, получаемых методом суммирования, определить эффект влияния количества слагаемых в теореме 3.2.

Стандартные равномерно распределенные случайные числа получать с помощью подпрограмм URAND или RANDU.

Смоделировать выборку такого же объема с помощью программ пакета MATHCAD (см. табл. 1), построить график этой выборки и сравнить оба полученных графика.

4. ТОЧЕЧНЫЕ И ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ И ИХ СВОЙСТВА

4.1. Статистические характеристики вариационных рядов и показатели их качества

Построив вариационный ряд и изобразив его графически, можно получить первоначальное представление о закономерностях в ряду наблюдений. Однако часто этого не достаточно. Поэтому для дальнейшего изучения изменения значений случайной величины используют числовые характеристики вариационных рядов. Поскольку эти характеристики вычисляются по статистическим данным, их обычно называют статистическими характеристиками или оценками.

Пусть по значениям измерений некоторой случайной величины требуется найти число, близкое к неизвестному значению измеряемого параметра. Например, пусть по значениям выборки объема n необходимо оценить неизвестный параметр θ закона распределения случайной величины X $P(X \leq x) = F(\theta, x)$.

Точечной оценкой $\hat{\theta}$ неизвестного параметра θ называется произвольная функция элементов выборки $\hat{\theta} = f_{\theta}(x_1, x_2, \dots, x_n)$. Значения этой функции при полученных в результате измерений $X_i = x_i, i = 1, 2, \dots, n$, будут считаться приближенным значением параметра θ .

Любая функция результатов опытов, которая не зависит от неизвестных статистических характеристик, называется статистикой.

Точечной оценкой статистической характеристики θ (параметра) называется статистика, реализация которой, полученная в результате опытов, принимается за неизвестное истинное значение параметра θ .

Ясно, что статистика – случайная величина. Но не всякая статистика может быть оценкой θ . Чтобы статистика могла служить оценкой неизвестного параметра, необходимо, чтобы ее распределение было сосредоточено в достаточной близости от неизвестного значения θ . Оценка должна быть «хорошей», т.е. обладать рядом положительных качеств.

Показатель качества – это некоторая характеристика, определяющая соответствие оценки ее назначению, т.е. ее пригодность для получения решения поставленной задачи. Показатели качества могут измеряться в разных шкалах: количественных (шкала интервалов, отношений, абсолютных разностей), порядковых (шкала порядка, рангов, баллов), номинальных (шкала наименований). Показатели качества могут быть проблемно-ориентированными, наиболее удобными, физически понятными при решении задач определенного класса, или универсальными, пригодными для различных классов задач. Количественные показатели могут быть представлены в абсолютных или относительных единицах. Для функциональных характеристик, представляющих собой функции $f(\theta)$ некоторого аргумента θ , показатели могут быть локальными или глобальными. Локальные показатели характеризуют качество оценки $f(\theta)$ при фиксированных значениях аргумента θ , а глобальные – вдоль всего диапазона изменения θ . Локальным показателем является, например, дисперсия $D(f(\theta))$, глобальным $\max_{\theta} D(f(\theta))$.

Вся совокупность показателей качества может быть сгруппирована в четыре класса: функциональные, метрологические, технические и экономические (эффективности и эксплуатации).

4.2. Типовые принципы, используемые при построении оценок [5]

1. Принцип несмещенности π_1 . Согласно этому принципу оператор f , применяемый к выборке \bar{x} , должен выбираться так, чтобы оценка $\hat{\theta}$ была несмещенной или асимптотически несмещенной, т.е.

$$M(\hat{\theta}) = \theta \text{ или } \lim_{n \rightarrow \infty} M(\hat{\theta}) = \theta, \forall \theta \in F, \quad (4.2.1)$$

где F – класс возможных оценок. Показателем качества в этом случае является значение смещения $\varepsilon_{\hat{\theta}} = M(\hat{\theta}) - \theta$. Чем меньше $\varepsilon_{\hat{\theta}}$, тем качествен-

нее оценка и ее алгоритм. $\varepsilon_{\hat{\theta}} = 0$ соответствует оптимальному по этому показателю алгоритму.

2. Принцип состоятельности π_2 . В качестве оценки следует выбирать состоятельную оценку, которая при неограниченном объеме выборки n сходится по вероятности к оцениваемой характеристике θ , т.е.

$$P_{\delta} = P\left(\left|\hat{\theta} - \theta\right| \leq \delta\right) \xrightarrow{P} 1, \delta > 0, n \rightarrow \infty. \quad (4.2.2)$$

Не всякая состоятельная оценка несмещенная, но всякая состоятельная оценка, имеющая асимптотически конечное среднее, будет асимптотически несмещенной. Показателем качества здесь может быть, например, значение объема выборки n_1 , начиная с которого P_{δ} меньше заданного δ . Меньший объем выборки соответствует более качественной оценке.

3. Принцип минимума среднего квадрата отклонения (эффективности) π_3 . Лучшей (более эффективной) считается та оценка, которая имеет меньшее значение среднего квадрата уклонения (ошибки, квадратичного риска)

$$\bar{\sigma}_{\hat{\theta}}^{-2} = M\left[\left(\hat{\theta} - \theta\right)^2\right] = \varepsilon_{\hat{\theta}}^2 + \sigma_{\hat{\theta}}^2, \quad \bar{\sigma}_{\hat{\theta}}^{-2} \rightarrow \min. \quad (4.2.3)$$

Для каждой оцениваемой характеристики или параметра θ можно попытаться найти нижнюю грань $\inf_f \bar{\sigma}_{\hat{\theta}}^{-2}$ вдоль всех возможных операторов f . Оценка $\hat{\theta}$, для которой достигается $\inf_f \bar{\sigma}_{\hat{\theta}}^{-2}$, называется эффективной.

4. Принцип минимума дисперсии или объема эллипсоида рассеивания π_4 . В тех случаях, когда смещение $\varepsilon_{\hat{\theta}}$ известно, его можно учесть в результатах измерений. В других случаях оно может не влиять на результат. Тогда необходимо выбирать такой оператор f , который обеспечивает минимум дисперсии $\sigma_{\hat{\theta}}^2$. Для несмещенных оценок это эквивалентно минимуму $\bar{\sigma}_{\hat{\theta}}^{-2} = \sigma_{\hat{\theta}}^2$. Для векторных параметров рассматривается минимум объема эллипсоида рассеивания. Можно попытаться найти нижнюю грань дисперсии $\inf_f \sigma_{\hat{\theta}}^2$ и соответствующий ей оператор f . Показателем качества оценки $\hat{\theta}$ являются дисперсия оценки $D(\hat{\theta})$ или дисперсия объема эллипсоида рассеивания или значение n_1 , при котором обеспечивается требование к дисперсии оценки или к объему эллипсоида.

5. Принцип минимума ширины доверительных интервалов π_5 .

Показатель качества таких оценок - значение ширины доверительного интервала. По этому принципу обычно конструируются интервальные оценки.

6. Принцип минимума меры близости π_6 . Показатель качества - значение меры близости, например, $d(\hat{\theta}, \theta)$, $0 \leq d \leq 1$, причем $d(\hat{\theta}, \theta) \rightarrow \min$. Условие предпочтительности оценки имеет вид $d(\hat{\theta}_1) \leq d(\hat{\theta}_2)$.

7. Принцип извлечения максимума информации, содержащейся в выборке π_7 . Выбираются такие алгоритмы и оценки, которые содержат в себе максимум информации, имеющейся в выборке об измеряемой характеристике θ . Показателем качества оценки является значение разности информации, содержащихся в оценке и в выборке, например, $i(\hat{\theta}) = \frac{I(\hat{\theta}, \theta)}{I(x, \theta)}$, $i(\hat{\theta}) \rightarrow \max$. Условие предпочтительности $i(\hat{\theta}_1) \geq i(\hat{\theta}_2)$.

8. Принцип минимума потерь от использования оценки и проблемной ориентации π_8 . Выбирается оценка, дающая меньшие потери, если вместо истинного значения θ принимается $\hat{\theta}$. Показателем качества могут быть абсолютные или отнесенные к наилучшей оценке значения средних потерь байесовского типа, различные показатели качества решения $r(\hat{\theta}) = \frac{R(\hat{\theta})}{\inf_f R(\hat{\theta})}$, $r(\hat{\theta}) \rightarrow \min$.

9. Принцип асимптотической определенности π_9 . Согласно этому принципу асимптотические (при $n \rightarrow \infty$) свойства оценок должны быть четко определенными, что обеспечивает метрологическую определенность измерения характеристик. Асимптотическая определенность может сводиться, например, к асимптотической нормальности оценок, их несмещенности, эффективности и тому подобное.

10. Принцип инвариантности по наблюдениям π_{10} или по измеряемой характеристике π_{11} . Оценка $\hat{\theta}$ называется инвариантной по наблюдениям, если для любого взаимно однозначного отображения $f(X)$ имеет место равенство $\hat{\theta}(\bar{x}) = \hat{\theta}(f(\bar{x}))$. Оценка $\hat{\theta}$ называется инвариантной по измеряемой характеристике θ , если для произвольной однозначной функции f выражение $f(\hat{\theta})$ есть оценка для $f(\theta)$ той же структуры, того же типа, что и оценка $\hat{\theta}$.

11. Принцип устойчивости и корректности π_{12} . Оценка $\hat{\theta}$ должна быть мало критичной к отклонениям условий ее нахождения от номинальных (вида вероятностной модели, наличия помех и тому подобное). Небольшие отклонения условий не должны приводить к большим отклонениям значений оценок, ее точностных показателей. Показателем качества может быть абсолютное или относительное значение меры разброса смещений и дисперсий оценок при переходе от одной модели к другой в заданном классе.

12. Принцип минимума необходимой априорной информации π_{13} . Лучшей считается та оценка $\hat{\theta}$, которая при прочих равных условиях требует меньше априорных данных.

Из других принципов можно отметить принцип простоты реализации π_{14} , принцип адаптируемости к априорным и исходным данным π_{15} , принцип транзитивности π_{16} , заключающийся в независимости результатов оценивания от способа разбиения алгоритма на части, принцип самообучения и самоорганизации π_{17} , принцип универсальности π_{18} , состоящий в том, что алгоритм оценки $\hat{\theta}$ оказывается пригодным для оценки различных характеристик случайных элементов одного типа или одинаковых характеристик разнотипных случайных элементов.

Все приведенные принципы взаимосвязаны, а иногда и противоречивы, стремление выполнить один принцип противоречит возможности выполнить другой. Кроме того, для выбранного алгоритма f выполнение некоторых свойств может оказаться принципиально невозможным.

4.3. Точечные оценки вероятности по частоте, математического ожидания и дисперсии

1. Оценка вероятности по частоте. Пусть неизвестный параметр θ есть p - неизвестная вероятность события A , а ее оценка $\hat{\theta} - p^* = n_A/n$ - частота этого события по классической схеме случаев. Пусть также n_{A_i} - индикатор события A в этой схеме случаев. Распределение n_{A_i} , очевидно, таково

n_{A_i}	0	1
p_i	$1 - p_{A_i}$	p_{A_i}

Здесь $p_{A_i} = p_A$. Тогда $p^* = \frac{n_A}{n} = \frac{\sum_{i=1}^n n_{A_i}}{n} = \frac{n_{A_1} + n_{A_2} + \dots + n_{A_n}}{n} = \frac{s_n}{n}$.

По определению $M(n_{A_i}) = \sum_{i=1}^2 x_i p_i = 0 \cdot q + 1 \cdot p = p$,

$D(n_{A_i}) = \sum_{i=1}^2 x_i^2 p_i - m_{n_{A_i}}^2 = 0^2 \cdot q + 1^2 \cdot p - p^2 = p - p^2 = pq$. Отсюда

$$M(p^*) = M\left(\frac{n_A}{n}\right) = M\left(\sum_{i=1}^n \frac{n_{A_i}}{n}\right) = \frac{1}{n} \sum_{i=1}^n M(n_{A_i}) = \frac{1}{n} \cdot p \cdot \sum_{i=1}^n 1 = \frac{1}{n} \cdot p \cdot n = p.$$

Аналогично, $D(p^*) = D\left(\frac{n_A}{n}\right) = D\left(\sum_{i=1}^n \frac{n_{A_i}}{n}\right) = \frac{1}{n^2} \sum_{i=1}^n D(n_{A_i}) = \frac{1}{n^2} \cdot n \cdot pq = \frac{pq}{n}$.

Таким образом, $M(p^*) = p$, т.е. оценка вероятности по частоте не смещена.

По неравенству Чебышева $P(|p^* - p| \geq \varepsilon) \leq \frac{D(p^*)}{\varepsilon^2} = \frac{pq}{n\varepsilon^2}$. Перейдя к

противоположному событию, получим $P(|p^* - p| < \varepsilon) > 1 - \frac{pq}{n\varepsilon^2}$, т.е.

$P(|p^* - p| < \varepsilon) \xrightarrow{P} 1$. Следовательно, оценка вероятности по частоте – со-

стоятельная оценка. К тому же $\lim_{n \rightarrow \infty} D(p^*) = \lim_{n \rightarrow \infty} \frac{pq}{n} = 0$, таким образом,

это асимптотически эффективная оценка.

Для доказательства эффективности оценки необходимо выяснить, имеет ли она по сравнению с другими оценками, которых может быть достаточно много, наименьшую дисперсию или нет. В некоторых случаях этот минимум хорошо известен; тогда, сравнив с ним дисперсию рассматриваемой оценки, можно ответить на поставленный вопрос.

Так для случайной величины X , распределенной по нормальному закону с дисперсией $D(X)$, нижняя граница для дисперсий различных несмещенных оценок равна pq/n . Так как $D(p^*)$ совпадает с минимальной оценкой, то частота p^* , будучи несмещенной оценкой, является также и эффективной оценкой вероятности p .

2. Оценка математического ожидания. Пусть результаты наблюдений x_1, x_2, \dots, x_n случайной величины X независимы и

$M(x_1) = M(x_2) = \dots = M(x_n) = M(X) = m_X$. Дисперсии всех наблюдений должны быть конечны и $D(x_i) = D_X$, $i = \overline{1, n}$. В этих условиях в качестве точечной оценки $\theta = M(X)$ используется среднее арифметическое результатов наблюдений $\hat{\theta} = m_X^* = \frac{1}{n} \sum_{i=1}^n x_i$. Найдем математическое ожидание и

дисперсию этой оценки: $M(m_X^*) = M\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n M(x_i) = \frac{1}{n} \cdot m_X \cdot n = m_X$. Таким образом, легко доказывается, что m_X^* - несмещенная оценка m_X . Дисперсия оценки также практически очевидна:

$$D(m_X^*) = \frac{1}{n^2} \sum_{i=1}^n D(x_i) = \frac{1}{n^2} \cdot D_X \cdot n = \frac{D_X}{n}.$$

Воспользуемся опять неравенством Чебышева, получим $P(|m_X - m_X^*| \geq \varepsilon) \leq \frac{D(m_X^*)}{\varepsilon^2} = \frac{D_X}{n\varepsilon^2}$ или $P(|m_X - m_X^*| < \varepsilon) > 1 - \frac{D_X}{n\varepsilon^2}$. Очевидно, что $\lim_{n \rightarrow \infty} P(|m_X - m_X^*| < \varepsilon) = 1$, т.е. оценка m_X средним арифметическим - состоятельная оценка.

Эффективность или неэффективность оценки зависит от вида распределения случайной величины X . Если X - нормальная случайная величина, то эта оценка будет эффективной. Для других распределений этого может и не быть. Асимптотическую эффективность, однако, можно легко установить: $\lim_{n \rightarrow \infty} D(m_X^*) = \lim_{n \rightarrow \infty} \frac{D_X}{n} = 0$.

3. Оценка дисперсии. Естественной оценкой дисперсии случайной величины X служит ее выборочная дисперсия, т.е. если $\theta = D(X)$, то

$\hat{\theta} = D_X^* = \frac{1}{n} \sum_{i=1}^n (x_i - m_X^*)^2$, так как $m_X^* = m_X$. Представим формулу для

D_X^* в несколько ином виде через центрированные величины:

$$\begin{aligned} D_X^* &= \frac{1}{n} \sum_{i=1}^n (x_i - m_X - m_X^* + m_X)^2 = \frac{1}{n} \sum_{i=1}^n [(x_i - m_X) - (m_X^* - m_X)]^2 = \\ &= \frac{1}{n} \sum_{i=1}^n \left(\overset{\circ}{x_i - m_X} - \overset{\circ}{m_X^* - m_X} \right)^2 = \frac{1}{n} \left(\sum_{i=1}^n \left(\overset{\circ}{x_i} \right)^2 - 2 \overset{\circ}{m_X^*} \sum_{i=1}^n \overset{\circ}{x_i} + n \left(\overset{\circ}{m_X^*} \right)^2 \right), \text{ но} \end{aligned}$$

$$\begin{aligned}
& -\frac{2m_{\circ}^*}{n} \sum_{i=1}^n x_i = -\frac{2m_{\circ}^*}{n} \sum_{i=1}^n (x_i - m_X) = -2m_{\circ}^* \left(\frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n m_X \right) = \\
& = -2m_{\circ}^* \left(\underbrace{m_X^* - \frac{nm_X}{n}}_{m_{\circ}^*} \right) = -2 \left(m_{\circ}^* \right)^2.
\end{aligned}$$

Тогда $D_X^* = \frac{1}{n} \sum_{i=1}^n \left(x_i \right)^2 - \left(m_{\circ}^* \right)^2$. Математическое ожидание этого

$$\text{выражения легко находится: } M(D_X^*) = \frac{1}{n} \sum_{i=1}^n M \left[\left(x_i \right)^2 \right] - M \left[\left(m_{\circ}^* \right)^2 \right] =$$

$$= \frac{1}{n} \sum_{i=1}^n D(x_i) - D(m_X^*) = \frac{1}{n} \cdot D_X \cdot n - \frac{D_X}{n} = \frac{n-1}{n} D_X.$$

Таким образом, оценка D_X^* - смещенная оценка. Смещение здесь равно $-D_X/n$ и при $n \rightarrow \infty$ стремится к нулю. Чтобы получить несмещенную оценку достаточно D_X^* умножить на $\frac{n}{n-1}$. В результате полу-

чим $\tilde{D}_X = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_X^*)^2 = \frac{n}{n-1} D_X^*$ - несмещенную оценку дисперсии.

Для оценки состоятельности надо найти $D(D_X^*)$. Это сделать довольно трудно. Можно показать, что $D(D_X^*) = O(1/n)$ и выражается через центральные моменты вплоть до четвертого порядка. Приведем без доказательства формулы дисперсий смещенной и несмещенной оценок:

$$\begin{aligned}
D(D_X^*) &= \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}, \\
D(\tilde{D}_X) &= \frac{n(\mu_4 - \mu_2^2)}{(n-1)^2} - \frac{2(\mu_4 - 2\mu_2^2)}{(n-1)^2} + \frac{\mu_4 - 3\mu_2^2}{n(n-1)^2}.
\end{aligned}$$

Тогда, опять используя неравенство Чебышева, будем иметь

$$P \left(\left| D_X^* - D_X \right| \geq \varepsilon \right) \leq \frac{D(D_X^*)}{\varepsilon^2} = O \left(\frac{1}{n} \right) \quad \text{и} \quad \lim_{n \rightarrow \infty} P \left(\left| D_X^* - D_X \right| < \varepsilon \right) = 1, \quad \text{т.е.}$$

оценка D_X^* - состоятельная оценка, так же как и \tilde{D}_X .

Если распределение нормально, то $\mu_4 = 3\mu_2^2$ и тогда $D(D_X^*) = \frac{2\mu_2^2(n-1)}{n^2} = \frac{2D_X^2(n-1)}{n^2}$, а $D(\tilde{D}_X) = \frac{2D_X^2}{n-1}$. Следовательно, обе оценки смещенная и несмещенная асимптотически эффективны.

Имея оценку дисперсии, можно получить еще один интересный результат для нормального распределения. Видно, что $D(D_X^*) < D(\tilde{D}_X)$, так как $\frac{2(n-1)}{n^2} < \frac{2n}{n^2} = \frac{2}{n} < \frac{2}{n-1}$. Таким образом, смещенная оценка дисперсии точнее несмещенной.

4.4. Неравенство Крамера* - Рао

В вычислительных процедурах математической статистики желательно употреблять только те оценки, которые по возможности принимали бы значения, наиболее близкие к неизвестному параметру. Наличие у оценки свойств несмещенности, состоятельности и эффективности дает возможность выбирать такие оценки. Однако практика показывает, что состоятельная оценка может быть смещенной, наоборот, несмещенная оценка не обязана быть состоятельной. Несмещенная оценка может быть неэффективной и тому подобное.

Имеются несколько подходов к нахождению несмещенных оценок с минимальной дисперсией. Такие оценки существуют не всегда, но их нахождение всегда чрезвычайно сложно. Одним из путей построения таких оценок является использование неравенства Крамера – Рао, которое дает нижнюю границу для дисперсии несмещенной оценки.

Пусть $\hat{\theta}_n$ - несмещенная оценка неизвестного параметра θ , построенная по выборке объема n . Тогда

$$D(\hat{\theta}_n) \geq 1/nI, \quad (4.4.1)$$

где $I = I(\theta)$ - информация Фишера, определяемая в дискретном случае формулой

$$I = M\left[\left(\ln p(x, \theta)\right)'_{\theta}\right]^2 = \sum_{i=1}^n \left[\frac{p'_{\theta}(x_i, \theta)}{p(x_i, \theta)} \right]^2 p(x_i, \theta), \quad (4.4.2)$$

а в непрерывном – формулой

* Карл Харальд Крамер (1893-1985) – шведский математик.

$$I = M\left[\left(\ln f(x, \theta)\right)'_{\theta}\right]^2 = \int_{-\infty}^{\infty} \left[\frac{f'_{\theta}(x, \theta)}{f(x, \theta)}\right] f(x, \theta) dx. \quad (4.4.3)$$

Таким образом, дисперсия любой несмещенной оценки не может быть меньше $1/nI$. Эффективностью несмещенной оценки $\hat{\theta}_n$ называют по Крамеру – Рао величину

$$e = 1/nID(\hat{\theta}_n). \quad (4.4.4)$$

Ясно, что при таком определении эффективность любой оценки $\hat{\theta}_n$ при каждом θ заключена между нулем и единицей, причем чем она ближе к единице при каком-либо θ , тем лучше оценка $\hat{\theta}_n$ при этом значении неизвестного параметра. Если $e(\theta) = 1$ при любом θ , то оценка называется эффективной по Крамеру – Рао.

Пример. Рассмотрим оценку $\hat{\theta} = p^*$ неизвестной вероятности успеха $\theta = p$ в схеме Бернулли. Ранее в подразд. 4.3 доказана несмещенность этой оценки и получена формула $D(p^*) = pq/n$. Найдем информацию Фишера. Так как распределение случайной величины X в каждом опыте в схеме Бернулли совпадает с распределением индикатора, т.е.

X	$x_0 = 0$	$x_1 = 1$
p	$1 - p$	p

то $P(x_0, p) = P(0, p) = 1 - p$, $P(x_1, p) = P(1, p) = p$. Следовательно,

$$I = \left[\frac{P'_p(0, p)}{P(0, p)}\right]^2 P(0, p) + \left[\frac{P'_p(1, p)}{P(1, p)}\right]^2 P(1, p) = \left(\frac{-1}{1-p}\right)^2 (1-p) + \left(\frac{1}{p}\right)^2 p =$$

$$= \frac{1}{1-p} + \frac{1}{p} = \frac{1}{pq}.$$

Эффективность будет равна $e = \frac{1}{nID(p^*)} = \frac{1}{n \cdot (1/pq) \cdot (pq/n)} = 1$, т.е.

оценка $\hat{\theta} = p^*$ эффективна по Крамеру – Рао.

Надо заметить, что эффективные по Крамеру – Рао оценки существуют крайне редко.

Другой путь к построению эффективных оценок состоит во введении понятия достаточной статистики.

k -мерная статистика $s = (s_1, s_2, \dots, s_k)^T = (s_1(x_1, x_2, \dots, x_n), \dots, s_k(x_1, x_2, \dots, x_n))^T$ называется достаточной для параметра θ , если условное распределение $F_{x_1, x_2, \dots, x_n}(x_1, x_2, \dots, x_n / S = s)$ выборки x_1, x_2, \dots, x_n при условии $S = s$ не зависит от параметра θ .

Это определение на практике для проверки достаточности конкретных статистик использовать весьма сложно, поэтому часто пользуются факторизационной теоремой Неймана – Фишера.

Теорема 4.1 (Неймана – Фишера). Для того чтобы статистика $s = s(x_1, x_2, \dots, x_n)$ была достаточной для параметра θ , необходимо и достаточно, чтобы ряд распределения $P(x_1, x_2, \dots, x_n, \theta) = P(x_1, \theta)P(x_2, \theta) \dots P(x_n, \theta)$ в дискретном случае или плотность распределения $f(x_1, x_2, \dots, x_n, \theta) = f(x_1, \theta)f(x_2, \theta) \dots f(x_n, \theta)$ в непрерывном случае выборки x_1, x_2, \dots, x_n были представимы в виде

$$P(x_1, x_2, \dots, x_n, \theta) = A(x_1, x_2, \dots, x_n)B(s, \theta),$$

$$f(x_1, x_2, \dots, x_n, \theta) = A(x_1, x_2, \dots, x_n)B(s, \theta),$$

где функция $A(x_1, x_2, \dots, x_n)$ зависит только от x_1, x_2, \dots, x_n , а функция $B(s, \theta)$ только от s и θ .

Пример. Пусть x_1, x_2, \dots, x_n - выборка из генеральной совокупности с теоретической функцией распределения, являющейся нормальной со средним θ_1 и дисперсией θ_2 . Покажем, что двумерная статистика $s = (s_1, s_2)^T$, где $s_1 = 1/n(x_1 + x_2 + \dots + x_n)$, $s_2 = (x_1 - s_1)^2 + (x_2 - s_1)^2 + \dots + (x_n - s_1)^2$ является достаточной для двумерного параметра $\theta = (\theta_1, \theta_2)^T$. Действительно, формула для n -мерного нормального вектора $f(x_1, x_2, \dots, x_n, \theta) = \frac{1}{(2\pi\theta_2)^{\frac{n}{2}}} \exp\left(-\frac{s_2 + n(s_1 - \theta_1)^2}{2\theta_2}\right)$ имеет вид, указанный в теореме Неймана – Фишера, в котором

$$A(x_1, x_2, \dots, x_n) = 1, \quad B(s, \theta) = \frac{1}{(2\pi\theta_2)^{\frac{n}{2}}} \exp\left(-\frac{s_2 + n(s_1 - \theta_1)^2}{2\theta_2}\right).$$

Видно, что смысл достаточной статистики s заключается в том, что она включает в себя всю ту информацию о неизвестном параметре θ , которая содержится в исходной выборке x_1, x_2, \dots, x_n . На практике достаточные статистики играют важную роль. Они обладают рядом важных

свойств, например, при некоторых условиях имеют минимальную дисперсию и тому подобное.

4.5. Методы получения точечных оценок

1. Метод моментов. Пусть имеется выборка x_1, x_2, \dots, x_n из генеральной совокупности с теоретической функцией распределения $F(x)$, принадлежащей k -параметрическому семейству $F(\bar{x}, \theta_1, \theta_2, \dots, \theta_k)$ с неизвестными параметрами $\theta_1, \theta_2, \dots, \theta_k$, которые нужно оценить. Так как вид $F(x)$ известен, можно вычислить первые k теоретических моментов распределения, ибо формулы для этих моментов тоже известны. Эти моменты будут зависеть и от k неизвестных параметров $\theta_1, \theta_2, \dots, \theta_k$:

$$\begin{cases} v_1 = M(X) = v_1(\theta_1, \theta_2, \dots, \theta_k), \\ v_2 = M(X^2) = v_2(\theta_1, \theta_2, \dots, \theta_k), \\ \vdots \\ v_k = M(X^k) = v_k(\theta_1, \theta_2, \dots, \theta_k). \end{cases} \quad (4.5.1)$$

Суть метода моментов заключается в том, что так как выборочные моменты являются состоятельными оценками теоретических моментов, можно в системе (4.5.1) теоретические моменты v_1, v_2, \dots, v_k заменить выборочными $v_1^*, v_2^*, \dots, v_k^*$, а затем решить систему (4.5.1) относительно неизвестных параметров $\theta_1, \theta_2, \dots, \theta_k$, т.е. найти оценки $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$. Вместо системы (4.5.1) реально приходится решать систему

$$\left\{ \begin{array}{l} v_1^* = v_1(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k), \\ v_2^* = v_2(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k), \\ \dots \\ v_k^* = v_k(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k). \end{array} \right. \quad (4.5.2)$$

Часто получается, что найденные оценки $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ будут состоятельными оценками $\theta_1, \theta_2, \dots, \theta_k$. Справедлива следующая теорема об асимптотической нормальности оценок, полученных методом моментов.

Теорема 4.2. При некоторых условиях, наложенных на семейство $F(\bar{x}, \theta_1, \theta_2, \dots, \theta_k)$, совместное распределение случайных величин $\sqrt{n}(\hat{\theta}_1 - \theta_1), \sqrt{n}(\hat{\theta}_2 - \theta_2), \dots, \sqrt{n}(\hat{\theta}_k - \theta_k)$ при $n \rightarrow \infty$ сходится к k -мерному нормальному закону с нулевыми средними и ковариационной

матрицей, зависящей от теоретических моментов v_1, v_2, \dots, v_k и матрицы $\|\partial v_i / \partial \theta_j\|$.

Практически моментами выше четвертого пользоваться нежелательно, так как точность их вычисления резко падает с увеличением порядка моментов. В методе моментов не обязательно использовать первые k моментов. Иногда в этом методе привлекают более или менее произвольные функции от элементов выборки.

Оценки, полученные методом моментов, имеют эффективность по Крамеру – Рао, существенно меньшую единицы, и могут быть смещенными. Но они часто используются из-за простоты получения, иногда в качестве начального приближения.

2. Метод максимального правдоподобия. Один из важнейших методов для отыскания оценок параметров по данным выборки был предложен Р. Фишером и носит название метода наибольшего (или максимального) правдоподобия. Пусть имеется выборка объема n : x_1, x_2, \dots, x_n из генеральной совокупности с теоретической функцией распределения $F(x)$. Если случайная величина X , представленная этой выборкой, дискретна, то ее ряд распределения $P(X = x_i)$, $i = \overline{1, n}$. Пусть распределение имеет k неизвестных параметров $\theta_1, \theta_2, \dots, \theta_k$, которые нужно оценить. Тогда функция $L = L(x_1, x_2, \dots, x_n, \theta_1, \theta_2, \dots, \theta_k) = P(x_1, \theta_1, \theta_2, \dots, \theta_k) \times P(x_2, \theta_1, \theta_2, \dots, \theta_k) \cdot \dots \cdot P(x_n, \theta_1, \theta_2, \dots, \theta_k)$ называется функцией правдоподобия. Ее значение – это вероятность произведения событий, $X = x_2, \dots, X = x_n$, или, иначе, совместная вероятность появления чисел x_1, x_2, \dots, x_n . Чем больше значение L , тем правдоподобнее или более вероятно появление в результате наблюдений чисел x_1, x_2, \dots, x_n . Отсюда и название функции – функция правдоподобия результатов наблюдений. Если наблюдаемая случайная величина X непрерывна, то функция правдоподобия имеет аналогичный вид, с той лишь разницей, что вместо вероятностей $P(x_i, \theta_1, \theta_2, \dots, \theta_k)$ фигурируют значения функции плотности $f(x_i, \theta_1, \theta_2, \dots, \theta_k)$.

Метод нахождения оценок неизвестных параметров, основанный на требовании максимизации функции правдоподобия, называется методом максимального правдоподобия, а найденные этим методом оценки – оценками максимального правдоподобия.

Функции L или $\ln L$, рассматриваемые как функции параметров $\bar{\theta} = (\theta_1, \theta_2, \dots, \theta_k)^T$, достигают максимума при одном и том же значении параметра $\bar{\theta}$, так как $\ln L$ – монотонно возрастающая функция. Поэтому

вместо отыскания максимума функции L находят (что удобнее) максимум функции $\ln L$. Функция $\ln L$ называется логарифмической функцией правдоподобия.

По этому методу за оценку параметров $\hat{\theta}_1 = \hat{\theta}_1(x_1, x_2, \dots, x_n)$, $\hat{\theta}_2 = \hat{\theta}_2(x_1, x_2, \dots, x_n), \dots, \hat{\theta}_k = \hat{\theta}_k(x_1, x_2, \dots, x_n)$ принимаются значения аргументов функции L или $\ln L$, при которых вероятность получения данных значений выборки максимальна. Очевидно, что для этого необходимо $\partial L / \partial \bar{\theta} = 0$ или $\partial \ln L / \partial \bar{\theta} = 0$.

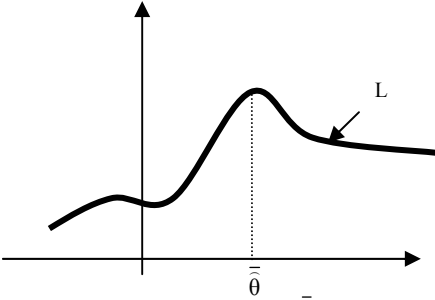


Рис. 4.1. Оценка параметра $\bar{\theta}$ на графике функции правдоподобия

Решая эту в общем случае систему нелинейных уравнений, находят значения параметров $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ (рис. 4.1).

Пример. Найдём оценку максимального правдоподобия для вероятности успеха в схеме Бернулли. Можно вероятность p рассматривать как параметр, входящий в распределение дискретной двузначной случайной величины X , принимающей только два значения 1 и 0 в зависимости от того, появится ли рассматриваемое событие в текущем испытании или не появится. Тогда $P(X = m) = p^m (1 - p)^{n-m}$, где n - количество испытаний, а m - число успехов в схеме Бернулли. Если m не фиксировать заранее, то $L = \prod_{k=1}^n p^{z_k} (1 - p)^{1-z_k}$, где $z_k = 0, 1$ - индикатор появления рассматриваемого события, $z_k = 1$, если это событие появится в k -м испытании и $z_k = 0$, если не появится. Очевидно, что $z_1 + z_2 + \dots + z_n = m$.

Тогда $\ln L = \sum_{k=1}^n [z_k \ln p + (1 - z_k) \ln(1 - p)]$ и $\frac{\partial \ln L}{\partial p} = \sum_{k=1}^n \left(\frac{z_k}{p} - \frac{1 - z_k}{1 - p} \right) = \frac{1}{p} \sum_{k=1}^n z_k - \frac{1}{1 - p} \sum_{k=1}^n (1 - z_k) = \frac{\bar{z} \cdot n}{p} - \frac{n}{1 - p} + \frac{n \cdot \bar{z}}{1 - p} = 0$. Здесь $\bar{z} = \frac{1}{n} \sum_{k=1}^n z_k$. Отсюда $\hat{p} = \bar{z} = \frac{1}{n} \sum_{k=1}^n z_k = \frac{m}{n}$.

Пример. Рассмотрим случайную величину X , подчиненную закону Пуассона с неизвестным параметром λ . Произведя выборку, получим наблюдаемые значения x_1, x_2, \dots, x_n . Величина X может принять любое из

значений $0, 1, 2, \dots$. Так как $P(X = x) = (\lambda^x / x!) e^{-\lambda}$, $x = 0, 1, 2, \dots$, то функция

$$\text{правдоподобия имеет вид } L(x_1, x_2, \dots, x_n, \lambda) = \prod_{k=1}^n \frac{\lambda^{x_k}}{x_k!} e^{-\lambda} =$$

$$= \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \cdot \frac{\lambda^{x_2}}{x_2!} e^{-\lambda} \cdot \dots \cdot \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} = e^{-n\lambda} \frac{\lambda^{\sum_{k=1}^n x_k}}{x_1! x_2! \dots x_n!}.$$

Найдем производную функции $\ln L$ по λ : $\ln L(x_1, x_2, \dots, x_n, \lambda) =$

$$= \sum_{k=1}^n x_k \ln \lambda - n\lambda - \ln \prod_{k=1}^n x_k! , \quad \frac{\partial \ln L}{\partial \lambda} = \frac{1}{\lambda} \sum_{k=1}^n x_k - n = 0 , \quad \hat{\lambda} = \frac{1}{n} \sum_{k=1}^n x_k .$$

В заключение необходимо убедиться, что найденный стандартным методом матанализа экстремум – максимум. Представляем читателям сделать это самостоятельно.

Пример. Пусть величины x_i , $i = \overline{1, n}$ имеют нормальное распределение. А неизвестных параметров два – матожидание и дисперсия. В этом случае

$$L(x_1, x_2, \dots, x_n, m_X, D_X) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi D_X}} \exp \left[-\frac{(x_k - m_X)^2}{2D_X} \right] = \left(\frac{1}{\sqrt{2\pi D_X}} \right)^n \times \\ \times \exp \left[-\sum_{k=1}^n \frac{(x_k - m_X)^2}{2D_X} \right], \text{ а } \ln L = -\frac{n}{2} (\ln 2\pi + \ln D_X) - \frac{1}{2D_X} \sum_{k=1}^n (x_k - m_X)^2 .$$

Для оценок \hat{m}_X и \hat{D}_X получим систему двух уравнений:

$$\begin{cases} \frac{\partial \ln L}{\partial m_X} = \frac{1}{\hat{D}_X} \sum_{k=1}^n (x_k - \hat{m}_X) = 0, \\ \frac{\partial \ln L}{\partial D_X} = -\frac{n}{2\hat{D}_X} + \frac{1}{2\hat{D}_X^2} \sum_{k=1}^n (x_k - \hat{m}_X)^2 = 0. \end{cases}$$

$$\text{Отсюда } \hat{m}_X = \frac{1}{n} \sum_{k=1}^n x_k , \text{ а } \hat{D}_X = \frac{1}{n} \sum_{k=1}^n (x_k - \hat{m}_X)^2 .$$

Метод максимального правдоподобия обладает важными достоинствами: он всегда приводит к состоятельным (хотя иногда и смещенным) оценкам, распределенным асимптотически нормально, имеющим наименьшую возможную дисперсию по сравнению с другими, также асимптотически нормальными оценками.

Однако далеко не для всех практических задач метод максимального правдоподобия дает удовлетворительные результаты. Дело в том, что предположение о принадлежности неизвестной плотности распределения

определенному параметрическому семейству (нормальному, показательному или какому-то другому) на практике выполняется лишь приближенно. Метод, который принимает это предположение безоговорочно, может привести к результатам, не имеющим даже приблизительно правильного характера. Так может происходить и при определенных, хоть и небольших, отклонениях от начальных предположений.

4.6. Сущность интервального оценивания

Поскольку все точечные оценки основаны на данных выборки, следовательно, они являются случайными величинами. В предыдущих подразделах были оценены их математические ожидания и дисперсии. Интервальные оценки учитывают факт случайности точечных оценок и дают представление об их точности и надежности. Рассмотрим интервальную оценку на примере математического ожидания.

Найдем ε из равенства $P(|m_X^* - m_X| < \varepsilon) = \beta$, где $\beta = 0.9, 0.95, 0.99$, т.е. событие $|m_X^* - m_X| < \varepsilon$ практически достоверное. Снимем модуль под знаком вероятности, получим $P(m_X^* - \varepsilon < m_X < m_X^* + \varepsilon) = \beta$. Это означает, что m_X с вероятностью β попадает в интервал $I_\beta = (m_X^* - \varepsilon, m_X^* + \varepsilon)$. В данном случае, поскольку m_X не случайно, а m_X^* случайно, то I_β тоже случайная величина. Поэтому правильнее говорить, что с вероятностью β случайный интервал I_β

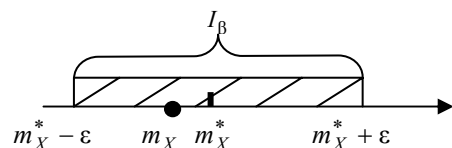


Рис. 4.2. Доверительный интервал для параметра m_X

длиной 2ε накрывает точку m_X (рис. 4.2).

Вероятность β называется доверительной вероятностью, а I_β - доверительным интервалом. Границы доверительного

интервала могут быть вычислены точно и приближенно.

4.7. Приближенные и точные доверительные интервалы для параметров распределений

1. Приближенное оценивание - это оценивание длин доверительных интервалов - базируется на центральной предельной теореме. Пусть произведено n независимых опытов над случайной величиной X , характе-

ристики которой – математическое ожидание и дисперсия – неизвестны.

Для этих параметров получены оценки $m_X^* = \frac{1}{n} \sum_{i=1}^n x_i$,

$$D_X^* = \frac{1}{n} \sum_{i=1}^n (x_i - m_X^*)^2.$$

Вид распределения случайной величины X может быть произвольным. Требуется построить доверительный интервал I_β , соответствующий доверительной вероятности β , для математического ожидания m_X .

Оценка математического ожидания – величина m_X^* представляет собой сумму n независимых одинаково распределенных случайных величин x_i , и, согласно центральной предельной теореме, при $n \rightarrow \infty$ ее закон распределения превратится в нормальный.

Итак, если $Y = \sum_{i=1}^n x_i$, то $P(Y < y) = F(y) \rightarrow \Phi\left(\frac{y - m_Y}{\sigma_Y}\right)$, где Φ – функция Лапласа. Если использовать стандартизированное среднее арифметическое, то

$$P\left(\frac{Y - m_Y}{\sigma_Y} < y\right) = P\left(\frac{(1/n) \sum_{i=1}^n x_i - m_X}{\sqrt{D_X/n}} < x\right) \approx \Phi(x),$$

поскольку, как было показано в предыдущих подразделах, $m_Y = m_X$ и $D_Y = D_X/n$.

Пусть D_X нам известно, тогда известно и $D_Y = D_X/n$. Найдем ε_β

из равенства $P(|m_X^* - m_X| < \varepsilon_\beta) = \beta$. Так как $m_X^* = \frac{1}{n} \sum_{i=1}^n x_i$, то

$M(m_X^*) = m_X$ и $D(m_X^*) = D_X/n$. Распишем исходное равенство для определения длины доверительного интервала подробнее:

$$\begin{aligned} P(m_X^* - \varepsilon_\beta < m_X < m_X^* + \varepsilon_\beta) &= \beta \approx \Phi\left(\frac{m_X^* + \varepsilon_\beta - m_X}{\sqrt{D_X/n}}\right) - \Phi\left(\frac{m_X^* - \varepsilon_\beta - m_X}{\sqrt{D_X/n}}\right) \approx \\ &\approx \Phi\left(\frac{\varepsilon_\beta}{\sqrt{D_X/n}}\right) - \Phi\left(\frac{-\varepsilon_\beta}{\sqrt{D_X/n}}\right) = \Phi\left(\frac{\varepsilon_\beta}{\sqrt{D_X/n}}\right) - \left[1 - \Phi\left(\frac{\varepsilon_\beta}{\sqrt{D_X/n}}\right)\right] = 2\Phi\left(\frac{\varepsilon_\beta}{\sqrt{D_X/n}}\right) - 1. \end{aligned}$$

Здесь, чтобы привести выражение в правой части к одной функции Лапласа, были сокращены m_X и m_X^* . Так как $m_X \neq m_X^*$, этим допущена еще одна неточность, помимо использования центральной предельной теоремы.

Итак, окончательно $P(|m_X^* - m_X| < \varepsilon_\beta) \approx 2\Phi(\varepsilon_\beta / \sqrt{D_X/n}) - 1$. Приравнивая правую часть найденного равенства β , найдем приближенные границы доверительного интервала $2\Phi(\varepsilon_\beta / \sqrt{D_X/n}) - 1 = \beta$. Тогда

$$\varepsilon_\beta = \sqrt{D_X/n} \Phi^{-1}((1 + \beta)/2), \quad \Phi(x) = 1/\sqrt{2\pi} \int_{-\infty}^x e^{-t^2/2} dt. \quad \text{Отсюда}$$

$$I_\beta = (m_X^* - \varepsilon_\beta, m_X^* + \varepsilon_\beta).$$

На практике, конечно, очень часто D_X не известна, поэтому ее приходится заменять смещенной или несмещенной оценкой дисперсии. Это еще более «размывает» границы приближенного доверительного интервала для математического ожидания.

Построим теперь приближенный доверительный интервал для дисперсии. Все ранее приведенные предположения о распределении случайной величины X остаются в силе. Построим интервал для несмещенной

оценки дисперсии, т.е. для $\hat{D}_X = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_X^*)^2$. Величины, стоящие

под знаком суммы, уже не могут считаться независимыми, так как в каждое слагаемое входит m_X^* , зависящее от всех x_i . Поэтому непосредственно центральную предельную теорему применить нельзя. Однако можно показать, что при $n \rightarrow \infty$ распределение $\sum_{i=1}^n (x_i - m_X^*)^2$ тоже стремится к

нормальному. Тогда имеем $M(\hat{D}_X) = D_X$, $D(\hat{D}_X) = \frac{2}{n-1} D_X^2$ (см. подразд. 4.3). Далее поступим как в случае с математическим ожиданием:

$$P(\hat{D}_X - \varepsilon_\beta < D_X < \hat{D}_X + \varepsilon_\beta) = \beta \approx \Phi\left(\frac{\hat{D}_X + \varepsilon_\beta - D_X}{\sqrt{2/(n-1)} D_X}\right) - \Phi\left(\frac{\hat{D}_X - \varepsilon_\beta - D_X}{\sqrt{2/(n-1)} D_X}\right).$$

Тогда $P(|\hat{D}_X - D_X| < \varepsilon_\beta) \approx 2\Phi\left(\frac{\varepsilon_\beta}{\sqrt{2/(n-1)} D_X}\right) - 1 = \beta$. Отсюда

$$\varepsilon_\beta = \sqrt{2/(n-1)} D_X \Phi^{-1}((1 + \beta)/2).$$

Естественно, в тех случаях когда D_X не известна вместо нее употребляется ее оценка. Это еще более снижает точность доверительного интервала. Наконец, если ε_β найдено, то $I_\beta = (\bar{D}_X - \varepsilon_\beta, \bar{D}_X + \varepsilon_\beta)$.

2. Точное оценивание. Точный доверительный интервал для математического ожидания строится на основе распределения Стьюдента, а для дисперсии - на основе χ^2 -распределения. Для точного нахождения длин доверительных интервалов совершенно необходимо заранее знать вид закона распределения случайной величины X , тогда как для применения приближенных методов это не обязательно. Длина любого доверительного интервала находится из распределения каких-то статистик, а распределения этих статистик выводятся на основе известных вероятностных законов.

Пусть выборка x_1, x_2, \dots, x_n взята из нормальной генеральной совокупности с определенными математическим ожиданием и дисперсией, т.е. $x_i \in N(m_X, D_X)$, $i = \overline{1, n}$. Рассмотрим две вспомогательные статистики.

По определению (см. подразд. 2.1) χ_n^2 -распределение с n степенями свободы есть сумма квадратов независимых случайных величин, каждая из которых имеет стандартное нормальное распределение, т.е. $\chi_n^2 = x_1^2 + x_2^2 + \dots + x_n^2$, $x_i \in N(0,1)$, $i = \overline{1, n}$. Рассмотрим формулу для смещенной оценки дисперсии $D_X^* = \frac{1}{n} \sum_{i=1}^n (x_i - m_X^*)^2$. Здесь

$x_i \in N(m_X, D_X)$, $i = \overline{1, n}$, тогда $(x_i - m_X)/\sqrt{D_X} \in N(0,1)$, но так как $M(m_X^*) = m_X$, то и $(x_i - m_X^*)/\sqrt{D_X} \in N(0,1)$. Следовательно,

$$\left(\frac{x_1 - m_X^*}{\sqrt{D_X}}\right)^2 + \left(\frac{x_2 - m_X^*}{\sqrt{D_X}}\right)^2 + \dots + \left(\frac{x_n - m_X^*}{\sqrt{D_X}}\right)^2 = \chi_n^2, \text{ но } \sum_{i=1}^n \left(\frac{x_i - m_X^*}{\sqrt{D_X}}\right)^2 = \frac{D_X^* n}{D_X}.$$

Тогда статистика $D_X^* n / D_X$ имеет χ^2 -распределение с $n-1$ степенью свободы, так как на x_i наложено одно ограничение (связь) при вычислении m_X^* . Аналогично доказывается, что статистика $\bar{D}_X(n-1)/D_X$ имеет χ^2 -распределение с $n-1$ степенью свободы.

Таким же образом рассмотрим дробь Стьюдента $t = z\sqrt{n}/\sqrt{v}$ (см. подразд. 2.2). Здесь $z \in N(0,1)$, а $v \in \chi_n^2$. Пусть $z = \frac{m_X^* - m_X}{\sqrt{D_X/n}} \in N(0,1)$, а

роль статистики v будет играть дробь $v = D_X^* n / D_X$. Тогда

$$t = \frac{z\sqrt{n}}{\sqrt{v}} = \frac{\sqrt{n}(m_X^* - m_X) / \sqrt{D_X/n}}{\sqrt{nD_X^*/D_X}} = \frac{\sqrt{n}(m_X^* - m_X)}{\sqrt{D_X^*}}, \text{ причем эта статистика}$$

имеет распределение Стьюдента с $n-1$ степенью свободы. Аналогичным образом полученная статистика $t = \sqrt{n}(m_X^* - m_X) / \sqrt{\hat{D}_X}$ будет распределена по закону Стьюдента с $n-1$ степенью свободы. Напишем вновь исходное равенство для длины доверительного интервала $P(|m_X^* - m_X| < \varepsilon_\beta) = \beta$ и преобразуем его следующим образом:

$$P\left(\frac{|m_X^* - m_X|\sqrt{n}}{\sqrt{\hat{D}_X}} < \frac{\varepsilon_\beta\sqrt{n}}{\sqrt{\hat{D}_X}}\right) = \beta \quad \text{или} \quad P\left(\left|\frac{(m_X^* - m_X)\sqrt{n}}{\sqrt{\hat{D}_X}}\right| < \frac{\varepsilon_\beta\sqrt{n}}{\sqrt{\hat{D}_X}}\right) =$$

$$P\left(|t| < \frac{\varepsilon_\beta\sqrt{n}}{\sqrt{\hat{D}_X}} = t_\beta\right) = P(|t| < t_\beta) = \beta, \text{ где случайная величина } t \text{ имеет рас-}$$

пределение Стьюдента с $n-1$ степенью свободы. Но

$$P(|t| < t_\beta) = \int_{-t_\beta}^{t_\beta} s_{n-1}(t) dt = 2 \int_0^{t_\beta} s_{n-1}(t) dt = \beta. \text{ Итак, } 2 \int_0^{t_\beta} s_{n-1}(t) dt = \beta.$$

Величину t_β можно найти обратным интерполированием по таблице распределения Стьюдента. Тогда $\varepsilon_\beta = t_\beta \sqrt{\hat{D}_X/n}$, а сам интервал будет иметь вид $I_\beta = (m_X^* - t_\beta \sqrt{\hat{D}_X/n}, m_X^* + t_\beta \sqrt{\hat{D}_X/n})$.

Построим, наконец, точный доверительный интервал для дисперсии при тех же предположениях относительно выборки, что и в предыдущем случае.

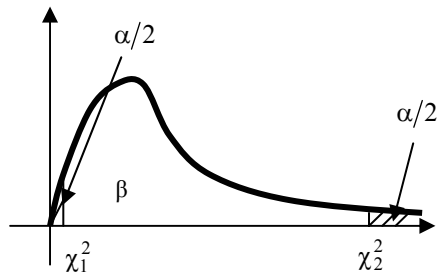


Рис. 4.3. Доверительный интервал для дисперсии, построенный на основе χ^2 -распределения

Так как χ^2 -распределение несимметрично, то условимся интервал, в которой попадает случайная величина с χ^2 -распределением, с заданной вероятностью β выбирать так, чтобы с левого и правого конца кривой плотности вероятности выхода случайной

величины за точки χ_1^2 и χ_2^2 были одинаковы и равны $\alpha/2 = (1 - \beta)/2$ (рис. 4.3). Тогда $P\left(\chi_1^2 < \frac{(n-1)\bar{D}_X}{D_X} < \chi_2^2\right) = 1 - P(\chi^2 < \chi_1^2) - P(\chi^2 > \chi_2^2) = \beta$. Переворачивая неравенство внутри вероятности, окончательно будем иметь $P\left(\frac{(n-1)\bar{D}_X}{\chi_2^2} < D_X < \frac{(n-1)\bar{D}_X}{\chi_1^2}\right) = \beta$. Величины χ_1^2 и χ_2^2 находят по таблицам χ^2 -распределения из равенств $P(\chi^2 > \chi_1^2) = 1 - \frac{\alpha}{2} = \frac{1 + \beta}{2}$, $P(\chi^2 > \chi_2^2) = \frac{\alpha}{2} = \frac{1 - \beta}{2}$.

4.8. Лабораторная работа № 4. Оценивание параметров вероятностных распределений в пакетах STATGRAPHICS и MATHCAD

При построении оценок параметров распределений к ним предъявляются различные требования, такие как: несмещенность, эффективность, устойчивость к отклонениям от модели и тому подобное. Постоянно предлагаются новые концепции и подходы к оцениванию, а также конкретные алгоритмы их реализации. Свой вклад в разнообразие оценок вносят и различные способы параметризации распределений. Все это порождает множество различных оценок одних и тех же параметров. Поэтому трудно ожидать, что в том или ином статистическом пакете обязательно найдется процедура, в точности реализующая требуемый алгоритм. Однако почти все пакеты выводят значения наиболее распространенных оценок параметров стандартных вероятностных распределений.

В пакете STATGRAPHICS Plus for Windows большинство точечных оценок получается по методу максимального правдоподобия, а интервальные оценки для математического ожидания и дисперсии строятся точные. При этом в комментариях в StatAdvisor подчеркивается, что выборка должна быть взята из нормальной генеральной совокупности, иначе доверительные интервалы не точны и должны быть скорректированы.

Получим точечные и интервальные оценки для распределения Парето, описанного в подразд. 3.6. Для этого в начале смоделируем выборку этого распределения объемом в 100 единиц. Моделирование выборок псевдослучайных чисел в пакете STATGRAPHICS описано в лабораторной работе № 2 (подразд. 2.7). Для этого необходимо в головном меню пакета выбрать пункт Plot→Probability Distribution и в появившемся дополнительном меню отметить распределение № 18 – Парето. После

щелчка по кнопке ОК появится заставка распределения Парето. Функция плотности вероятности этого распределения равна $f(x) = \alpha/x_0 (x_0/x)^{\alpha+1}$, $x > x_0$. В пакете автоматически задается $x_0 = 1$, таким образом, для полного определения распределения необходимо выбрать параметр формы α .

Щелкнем правой кнопкой мыши в любом месте заставки распределения Парето и в появившемся дополнительном меню выберем пункт Analysis Options. Зададим не пять, как позволяет пакет, а одно распределение с параметром формы (Shape), равным четырем. Далее в меню заставки распределения Парето выберем пункт Tabular Options и зададим в нем пункт Random Numbers. После щелчка по кнопке ОК будет автоматически смоделирована выборка псевдослучайных чисел, подчиненных распределению Парето, объемом 100 единиц. Сохраним эту выборку с помощью пункта меню Save Results под именем Pareto.

Для получения точечных и интервальных оценок параметров распределений в пакете STATGRAPHICS выберем в головном меню пункт Describe (Описание данных)→Numeric Data (Числовые данные)→One-Variable Analysis (Анализ одной переменной). Появится заставка дополнительного меню анализа одной переменной, в котором в окне Data необходимо указать имя выборки Pareto и нажать кнопку ОК.

Появится поле Analysis Summary (Сводка анализа). Выберем пункт дополнительного меню Tabular Options и зададим в нем вывод информации по разделам Analysis Summary, Summary Statistic (Описание данных) и Confidence Intervals (Доверительные интервалы). После щелчка по кнопке ОК на экран будет выведена информация, представленная на левой половине рис. 4.5.

Следует заметить, что по умолчанию пакет выводит значения лишь восьми общих статистик из девятнадцати. Если необходимы значения других точечных характеристик распределения, то их вывод на экран можно задать, щелкнув правой кнопкой мыши в поле Summary Statistics и выбрав пункт Pane Options дополнительного меню.

Аналогичный щелчок в поле Confidence Intervals и выбор пункта Pane

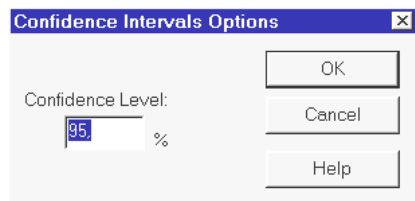


Рис. 4.4. Меню задания величины доверительной вероятности

Options вызывает дополнительное меню (рис. 4.4), которое задает величину доверительной вероятности β (Confidence Level). По умолчанию задается значение 95%. Выберем $\beta = 99$ и щелкнем по кнопке ОК. В поле Confidence Intervals немедленно изменятся границы точных доверительных интервалов для матема-

тического ожидания и стандартного отклонения. В пункте дополнительного меню Graphics Options отметим разделы Scatterplot (Диаграмма рассеивания), Frequency Histogram (Частотная гистограмма) и Density Trace (График функции плотности). Тогда вид информации, выводимой на экран дисплея, будет полностью соответствовать изображенной на рис. 4.5.

Оценка параметра формы распределения Парето по методу максимального правдоподобия в пакете не находится, по элементам выборки этот параметр рассчитывается следующим образом:

$$\hat{\alpha} = 1 / \left(\frac{1}{n} \sum_{i=1}^n \ln x_i - \ln x_0 \right) . \quad (4.8.1)$$

Косвенно его можно оценить по оценке математического ожидания

$$m_X^* = \frac{\alpha}{\alpha - 1} x_0, \quad \alpha > 1, \quad \text{тогда, так как } x_0 = 1, \quad \text{то } \hat{\alpha} = \frac{m_X^*}{m_X^* - 1} = \frac{1.28435}{1.28435 - 1} \approx 4.52.$$

В пакете MATHCAD нет встроенных процедур оценок максимального правдоподобия и построения доверительных интервалов, поэтому их придется программировать самостоятельно. Кроме того, в табл. 1 (см. подразд. 2.7) нет распределения Парето, следовательно, будем моделировать его по формуле (3.6.6) с использованием стандартных равномерных случайных чисел, получаемых по программам URAND или RUNIF (см. лабораторную работу №3, подразд. 3.6).

Сначала, так же как в пакете STATGRAPHICS, смоделируем выборку из генеральной совокупности с функцией распределения Парето объемом 100 единиц. Это можно сделать следующим образом.

$$\text{ORIGIN} := 1 \quad n := 100 \quad x_0 := 1 \quad \text{alfa} := 4 \quad c := \frac{1}{\text{alfa}} \quad t := \text{runif}(n, 0, 1) \quad i := 1 \dots 100 \quad d_i := x_0 * \left(\frac{1}{t_i} \right)^c$$

$$Mx := \text{mean}(d) \quad Mx = 1.375 \quad Dx := \text{var}(d) \quad \sigma x := \sqrt{Dx}$$

$$Dx = 0.433 \quad \sigma x = 0.658$$

Получены точечные оценки математического ожидания и дисперсии. Оценка математического ожидания практически совпадает с аналогичной оценкой в пакете STATGRAPHICS, там Average=1.284. Дисперсия же значительно больше. Это связано с моделирующей формулой (3.6.6); если элементы выборки, полученной в пакете STATGRAPHICS, изменялись примерно от единицы до трех, то в пакете MATHCAD разброс элементов анало-

t =	1	1.268·10 ⁻³	d =	1	5.299
	2	0.193		2	1.508
	3	0.585		3	1.143
	4	0.35		4	1.3
	5	0.823		5	1.05
	6	0.174		6	1.548
	7	0.71		7	1.089
	8	0.304		8	1.347
	9	0.091		9	1.819
	10	0.147		10	1.614
	11	0.989		11	1.003
	12	0.119		12	1.702

гичной выборки составляет от единицы до пяти, т.е. масштаб рассеивания значительно больше.

Analysis Summary

Data variable: Pareto

100 values ranging from 1,00147 to 3,10832

The StatAdvisor

This procedure is designed to summarize a single sample of data. It will calculate various statistics and graphs. Also included in the procedure are confidence intervals and hypothesis tests. Use the Tabular Options and Graphical Options buttons on the analysis toolbar to access these different procedures.

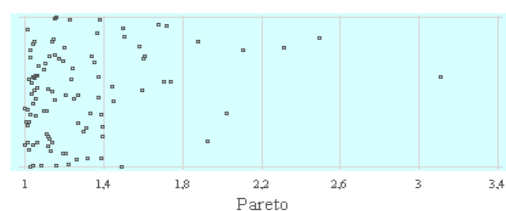
Summary Statistics for Pareto

Count = 100
Average = 1,28435
Median = 1,15642
Mode =
Geometric mean = 1,25011
Variance = 0,118775
Standard deviation = 0,344637
Standard error = 0,0344637
Minimum = 1,00147
Maximum = 3,10832
Range = 2,10685
Lower quartile = 1,05436
Upper quartile = 1,38063
Interquartile range = 0,32627
Skewness = 2,58741
Std. skewness = 10,5631
Kurtosis = 8,86239
Std. kurtosis = 18,0903
Coeff. of variation = 26,8337%
Sum = 128,435

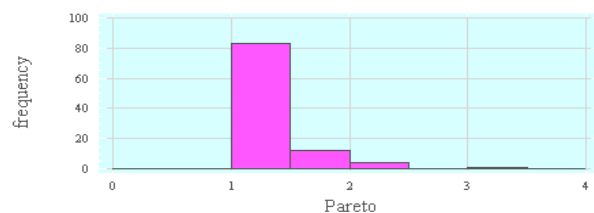
Confidence Intervals for Pareto

99,0% confidence interval for mean: 1,28435 +/- 0,0905159 [1,19383;1,3
99,0% confidence interval for standard deviation: [0,290866;0,420471]

Scatterplot for Pareto



Histogram for Pareto



Density Trace for Pareto

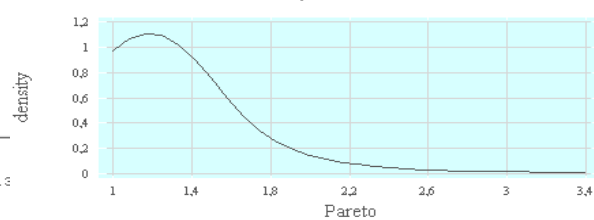


Рис. 4.5. Характеристики распределения Парето

Оценим теперь параметр формы α распределения Парето по методу максимального правдоподобия. Для этого составим функцию правдоподобия с учетом того, что $x_0 = 1$:

$$L(x, \alpha) = \prod_{i=1}^n \alpha \left(\frac{1}{x_i} \right)^{\alpha+1} = \alpha^n \prod_{i=1}^n \left(\frac{1}{x_i} \right)^{\alpha+1}. \quad (4.8.2)$$

Логарифм функции правдоподобия

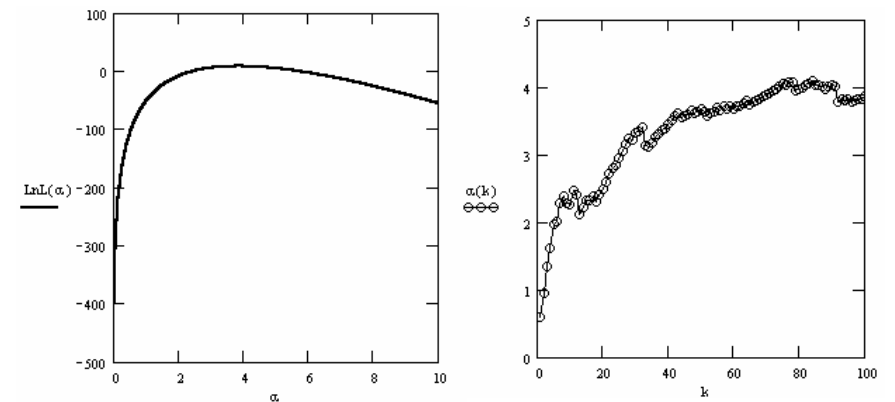
$$\ln L(x, \alpha) = n \ln \alpha + \sum_{i=1}^n \ln \left(\frac{1}{x_i} \right)^{\alpha+1} = n \ln \alpha - (\alpha + 1) \sum_{i=1}^n \ln x_i.$$

Тогда $\frac{\partial \ln L(x, \alpha)}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^n \ln x_i = 0$ и $\hat{\alpha} = n / \sum_{i=1}^n \ln x_i$, т.е. получена

формула (4.8.1) с учетом $x_0 = 1$. Вычислим в пакете MATHCAD логарифм функции правдоподобия и построим ее график:

$$\text{Ln}L(\alpha) := n * \ln(\alpha) - (\alpha + 1) * \sum_{i=1}^n \ln(d_i)$$

$$k := 1 \dots 100 \quad \alpha(k) := \frac{k}{\sum_{i=1}^k \ln(d_i)} \quad \alpha(10) = 2.262 \quad \alpha(50) = 3.697 \quad \alpha(100) = 3.865$$



Построим, наконец, 99%-ные доверительные интервалы для математического ожидания и дисперсии, точечные оценки которых дают программы **mean** и **var**. Поскольку в пакете MATHCAD имеются встроенные функции для вычисления процентиля нормального распределения, распределения Стьюдента и χ^2 -распределения, то легко строятся по формулам подразд. 4.7 любые доверительные интервалы. Построим сначала приближенные интервалы.

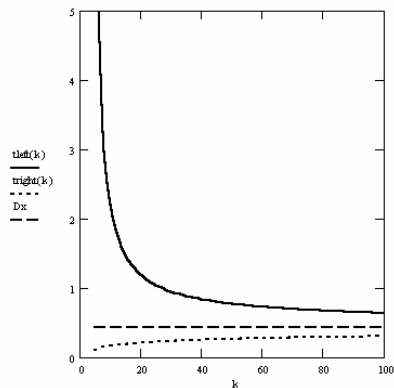
$$\begin{aligned}\beta &:= 0.99 \quad t1 := \text{qnorm}\left(\frac{1+\beta}{2}, 0, 1\right) \quad t1 = 2.576 \quad \varepsilon := \sqrt{\frac{Dx}{n}} * t1 \\ Mxl &:= Mx - \varepsilon \quad Mxr := Mx + \varepsilon \quad Mxl = 1.205 \quad Mxr = 1.544 \\ \varepsilon1 &:= \sqrt{\frac{2}{n-1}} * Dx * t1 \quad \varepsilon1 = 0.158 \quad Dx1 := Dx - \varepsilon1 \quad Dxr := Dx + \varepsilon1 \\ Dx1 &= 0.274 \quad Dxr = 0.591\end{aligned}$$

Итак, доверительные интервалы, базирующиеся на предположениях ЦПТ, вычислены. Допуская, что выборка взята из нормальной генеральной совокупности (а наша выборка имеет распределение Парето!), построим «точные» интервалы.

$$\begin{aligned}t1 &:= \text{qt}\left(\frac{1+\beta}{2}, n\right) \quad t1 = 2.626 \quad \varepsilon := \sqrt{\frac{Dx}{n}} * t1 \quad \varepsilon = 0.173 \\ Mxl1 &:= Mx - \varepsilon \quad Mxr1 := Mx + \varepsilon \quad Mxl1 = 1.202 \quad Mxr1 = 1.548 \\ t1 &:= \text{qchisq}\left(\frac{1-\beta}{2}, n-1\right) \quad t1 = 66.510 \quad t2 := \text{qchisq}\left(\frac{1+\beta}{2}, n-1\right) \quad t2 = 138.987 \\ Dx11 &:= Dx * \frac{(n-1)}{t2} \quad Dxr1 := Dx * \frac{(n-1)}{t1} \quad Dx11 = 0.308 \quad Dxr1 = 0.644\end{aligned}$$

В заключение исследуем изменение длины точного доверительного интервала, например, для дисперсии в зависимости от объема выборки.

$$\begin{aligned}\beta1 &:= \frac{1-\beta}{2} \quad \beta2 := \frac{1+\beta}{2} \\ k &:= 5 \dots 100 \quad \text{tright}(k) := Dx * \frac{k-1}{\text{qchisq}(\beta2, k-1)} \\ tleft(k) &:= Dx * \frac{k-1}{\text{qchisq}(\beta1, k-1)}\end{aligned}$$



Задание №1. По номеру фамилии студента в журнале преподавателя выбрать распределение из табл. 1 (если номер больше 16, выбирать номер минус 15) и получить точечные и интервальные оценки ($\beta = 0.95$) математического ожидания и дисперсии в пакетах STATGRAPHICS и MATHCAD. В пакете MATHCAD, кроме того, по методу максимального правдоподобия оценить параметры выбранного распределения.

5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ. КРИТЕРИИ СОГЛАСИЯ

5.1. Понятие статистической гипотезы. Основные этапы проверки гипотез

Во многих случаях результаты наблюдений используются для проверки предположений относительно тех или иных свойств распределений. Такие задачи очень часты, например, они возникают при сравнении различных технологических процессов и методов. Рассмотрение подобных задач в строгой математической постановке приводит к понятию статистической гипотезы.

Под статистической гипотезой понимают всякое предположение о генеральной совокупности (о распределении вероятностей), проверяемое по результатам наблюдений. Для проверки естественнонаучных гипотез часто применяется такой принцип: гипотезу отвергают, если происходит то, что при ее справедливости происходить не должно. Проверка статистических гипотез происходит также, только место невозможных событий занимают события практически невозможные. Причина этого проста: пригодных для проверки невозможных событий, как правило, просто нет.

Статистическая гипотеза называется простой, если она полностью задает распределение вероятностей. Сложная гипотеза указывает не одно распределение, а некоторое множество распределений. Например, простая гипотеза о том, что случайная величина X распределена по стандартному нормальному закону, т.е. $X \in N(0,1)$, немедленно становится сложной, если $m_X \neq 0$ или $D_X \neq 1$. В задачах практики часто бывает известен вид закона распределения X и надо проверить лишь предположения о значениях параметров данного распределения.

Если в гипотезе речь идет о соответствии числовых параметров данного распределения какому-то конкретному значению, то такая гипотеза называется параметрической.

Проверяемая гипотеза называется нулевой гипотезой и обозначается чаще всего H_0 . Вместе с ней рассматривается одна из альтернативных или конкурирующих гипотез, обозначаемых H_1 . Правило, на котором основывается решение о нулевой гипотезе, называется критерием. Все решения принимаются на основе выборки, следовательно, на основе какой-нибудь статистики. Эта статистика называется статистикой z критерия.

Выберем уровень вероятности α , $\alpha > 0$. Условимся считать событие практически невозможным, если его вероятность меньше α . Когда речь идет о проверке гипотезы, число α называют уровнем значимости. Принцип проверки гипотез очень прост. В соответствии с этим принципом маловероятные события считаются невозможными, а имеющие большую вероятность – достоверными.

Пусть W – множество значений статистики z , а ω – такое множество $\omega \in W$, что при гипотезе H_0

$$P(z \in \omega / H_0) = \alpha. \quad (5.1.1)$$

Пусть, наконец, z_α – выборочное значение статистики z . Тогда критерий формулируется следующим образом: гипотеза H_0 отклоняется при $z_\alpha \in \omega$ и принимается, если $z_\alpha \in W \setminus \omega$. Множество ω всех значений статистики z , при которых гипотеза H_0 отклоняется, называется критической областью; область $W \setminus \omega$ называется областью принятия решения.

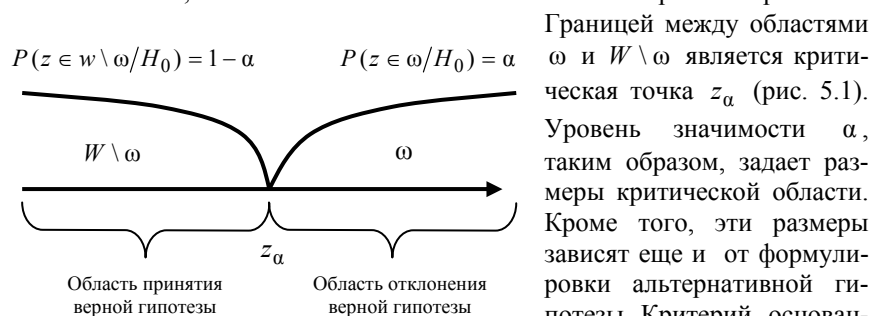


Рис. 5.1. Область принятия решения и критическая область статистики z

Границей между областями ω и $W \setminus \omega$ является критическая точка z_α (рис. 5.1). Уровень значимости α , таким образом, задает размеры критической области. Кроме того, эти размеры зависят еще и от формулировки альтернативной гипотезы. Критерий, основанный на использовании заранее заданного уровня значимости, называется критерием значимости.

Возможны три вида расположения критической области ω в зависимости от нулевой и альтернативной гипотез, вида распределения статистики z критерия:

1. Правосторонняя критическая область (рис. 5.2), состоящая из интервала $(z_{\text{пр},\alpha}, +\infty)$, где точка $z_{\text{пр},\alpha}$ определяется из условия $P(z > z_{\text{пр},\alpha}) = \alpha$ и называется правосторонней критической точкой, отвечающей уровню значимости α ;

2. Левосторонняя критическая область (рис. 5.3), состоящая из интервала $(-\infty, z_{\text{лев},\alpha})$, где точка $z_{\text{лев},\alpha}$ определяется из уравнения

$P(z < z_{\text{лев},\alpha}) = \alpha$ и называется левосторонней критической точкой с уровнем значимости α ;

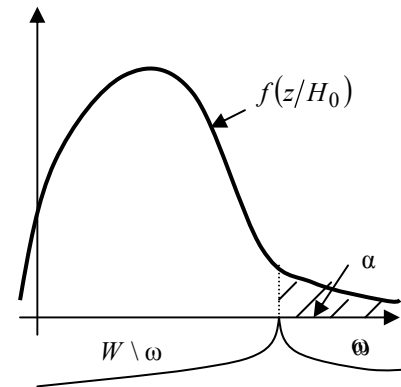


Рис. 5.2. Правосторонняя критическая область

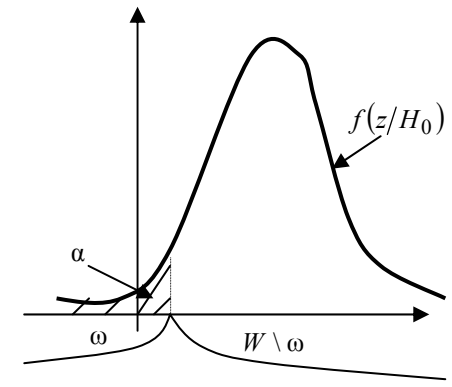


Рис. 5.3. Левосторонняя критическая область

3. Двусторонняя критическая область (рис. 5.4), состоящая из двух интервалов $(-\infty, z_{\text{лев},\alpha/2})$ и $(z_{\text{пр},\alpha/2}, +\infty)$, где точки $z_{\text{лев},\alpha/2}$ и $z_{\text{пр},\alpha/2}$ определяются из условий $P(z < z_{\text{лев},\alpha/2}) = \alpha/2$ и $P(z > z_{\text{пр},\alpha/2}) = \alpha/2$ и называются двусторонними критическими точками. Все сказанное иллюстрируют вышеприведенные рисунки.

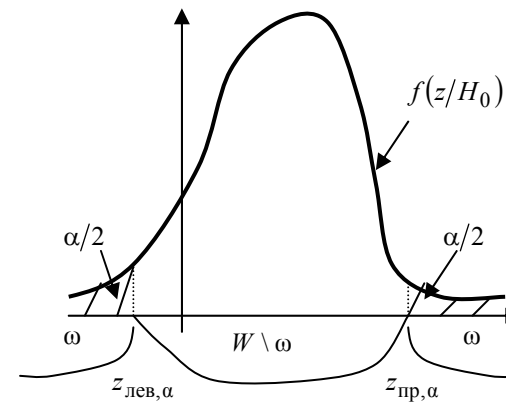


Рис. 5.4. Двусторонняя критическая область

Как правило, статистику z критерия выбирают таким образом, чтобы ее распределения при нулевой гипотезе H_0 и при альтернативной H_1 как можно более различались. При таком выборе статистики z обычно некоторые значения z (например, слишком большие или слишком маленькие) нетипичны при гипотезе H_0 и типичны при альтернативе H_1 .

Проверка параметрической статистической гипотезы при помощи критерия значимости включает в себя следующие этапы.

1. Формулируется проверяемая нулевая гипотеза H_0 и альтернативная H_1 .

2. Назначается уровень значимости α .

3. Выбирается статистика z критерия значимости для проверки гипотезы H_0 .

4. Определяется выборочное распределение статистики z при условии, что верна гипотеза H_0 , т.е. находится функция плотности вероятности $f(z/H_0)$.

5. В зависимости от формулировки альтернативной гипотезы H_1 определяется критическая область ω , область принятия решения $W \setminus \omega$ и вид решения $z > z_{\text{пр},\alpha}$, $z < z_{\text{лев},\alpha}$ или $z < z_{\text{лев},\alpha/2}$ и $z > z_{\text{пр},\alpha/2}$.

6. По имеющейся выборке наблюдений вычисляется выборочное значение статистики $z_{\text{в}}$.

7. Принимается решение о гипотезе H_0 .

Этапы 1-7 обычно используют статистику, квантили которой табулированы. Это либо нормальное, либо χ^2 -распределение, либо распределение Стьюдента.

Однако принимаемое на основе критерия значимости решение тоже может быть ошибочным. Пусть гипотеза H_0 верна, но $z_{\text{в}} \in \omega$, т.е. значение статистики критерия попало в критическую область и, следовательно, H_0 отвергается.

Ошибка, совершаемая при отклонении правильной гипотезы H_0 , называется ошибкой первого рода. Вероятность ошибки первого равна уровню значимости, т.е. $P(z \in \omega/H_0) = \alpha$. Может случиться и другая ситуация. Пусть гипотеза H_0 не верна, но $z \in W \setminus \omega$, т.е. значение статистики критерия попало в область принятия решения. Тогда будет принята неверная гипотеза. Ошибка, совершаемая при принятии неверной гипотезы H_0 , когда верна H_1 , называется ошибкой второго рода. Ее вероятность $P(z \in W \setminus \omega/H_1) = \beta$. Вероятность $1 - \beta = P_1(\omega)$ называется мощностью критерия. Чем выше мощность критерия, тем чаще отвергается неверная гипотеза. Все сказанное иллюстрирует табл. 2 и рис. 5.5.

Т а б л и ц а 2

Решение, принимаемое о гипотезе H_0 по выборке	H_0 отвергается, H_1 принимается	H_0 принимается, H_1 отвергается
Нулевая гипотеза H_0 - верна	$P(z \in \omega / H_0) = \alpha$ - отвергается верная гипотеза. Ошибка первого рода	$P(z \in w \setminus \omega / H_0) = 1 - \alpha$ - принимается верная гипотеза
Гипотеза H_0 неверна, т.е. верна гипотеза H_1	$P(z \in \omega / H_1) = 1 - \beta$ - отвергается неверная гипотеза. Мощность критерия	$P(z \in w \setminus \omega / H_1) = \beta$ - принимается неверная гипотеза. Ошибка второго рода

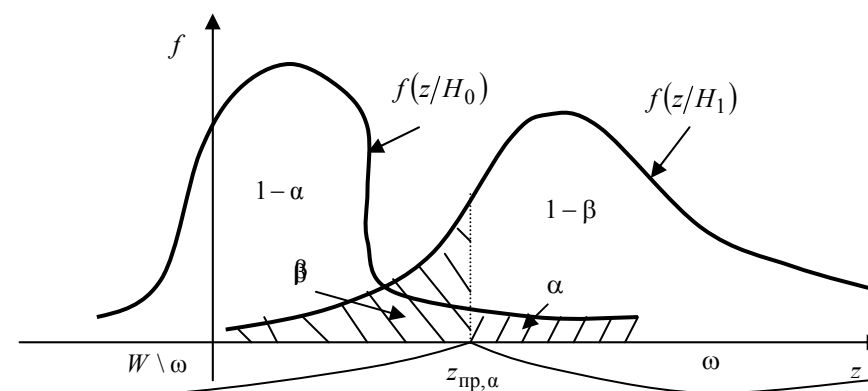


Рис. 5.5. Геометрическая интерпретация уровня значимости и мощности критерия при различных гипотезах

Обратим внимание на то, что в результате проверки нулевой гипотезы H_0 возможно правильное решение двух видов:

- 1) принять гипотезу H_0 , когда в действительности она имеет место; вероятность этого решения $P(z \in w \setminus \omega / H_0) = 1 - \alpha$;
- 2) не принять гипотезу H_0 , когда и на самом деле гипотеза H_0 неверна; вероятность этого решения $P(z \in \omega / H_1) = 1 - \beta$.

Пример. Из продукции автомата, обрабатывающего болты с номинальным значением контролируемого размера $m_0 = 40$ мм, была взята выборка болтов объемом $n = 36$. Выборочное среднее контролируемого

размера $m^* = 40.2$ мм. Результаты предыдущих измерений дают основание полагать, что действительные размеры болтов образуют нормально распределенную совокупность с дисперсией $D = 1 \text{ мм}^2$. Можно ли по результатам проведенного выборочного исследования утверждать, что контролируемый размер в продукции автомата не имеет положительного смещения по отношению к номинальному размеру? Принять $\alpha = 0.01$. Какова критическая область в данном случае?

Решаем задачу по приведенному выше плану из семи пунктов. Сформулируем сначала основную и конкурирующую гипотезы: $H_0 : m_0 = 40$, $H_1 : m_0 > 40$, так как речь идет о положительном смещении контролируемого размера. Уровень значимости задан в условии: $\alpha = 0.01$. В качестве статистики критерия используем оценку математического ожидания $z = m^*$. Так как исходная выборка нормальна, то $m^* \in N(m_0, D/n)$. Альтернативная гипотеза $H_1 : m_0 > 40$ предполагает правосторонний критерий значимости, критическая область определяется неравенством $z > z_{\text{пр}, \alpha}$, где $P(z > z_{\text{пр}, \alpha}) = \alpha$ (см. рис. 5.2). Для того чтобы выбранная статистика критерия была распределена стандартно нормально, ее необходимо центрировать и нормировать. Тогда

$$z = \frac{m^* - m_0}{\sqrt{D/n}} = \frac{m^* - m_0}{\sqrt{1/36}} \in N(0, 1). \text{ Критическую точку порядка } 1 - \alpha \text{ для}$$

нормального распределения $u_{0.99}$, где $P(z > u_{0.99}) = 0.99$, очень легко найти по таблицам: $z_{\text{пр}, \alpha} = u_{0.99} = 2.326$.

Найдем, наконец, выборочное значение статистики

$$z_{\text{в}} = \frac{40.2 - 40}{\sqrt{1/36}} = \frac{0.2}{0.1667} = 1.2. \text{ Теперь можно принять решение о гипотезе}$$

H_0 : так как $z_{\text{в}} < z_{\text{пр}, \alpha}$, т.е. $z_{\text{в}} = 1.2 < u_{0.99} = 2.326$, гипотеза H_0 должна быть принята. Это значит, что по результатам проведенной выборки нельзя утверждать, что автомат при выпуске продукции дает положительный сдвиг.

Найдем в заключение границу критической области, т.е. границу

$$m_{\text{крит}}^*. \text{ Так как } \frac{m_{\text{крит}}^* - 40}{\sqrt{1/36}} = 2.326, \text{ то } m_{\text{крит}}^* = 40.39 \text{ мм. Таким образом,}$$

$$\omega = \{m^* > 40.39\}, \text{ а } W \setminus \omega = \{m^* \leq 40.39\}.$$

5.2. Критерий Неймана – Пирсона

Если имеется некоторая выборка x_1, x_2, \dots, x_n , то с помощью заданных ошибок первого и второго рода α и β можно решать задачу о наилучшем критерии. Именно по заданному значению уровня значимости α ищется такой критерий, чтобы его мощность $1 - \beta$ была максимальна. Введем предварительно несколько обозначений и определений.

Размером α_0 критерия называется максимальное значение вероятности ошибки первого рода при использовании данного критерия, т.е.

$$\alpha_0 = \sup_{F(x) \in F} \alpha(F(x)). \quad (5.2.1)$$

Равномерно наиболее мощным критерием заданного размера α_0 называется критерий, имеющий среди всех критериев размера α_0 наибольшую мощность $1 - \beta = 1 - \beta(F(x))$ при любом распределении $F(x) \in F$. Равномерно наиболее мощные критерии существуют в крайне редких случаях, например, в случае простых гипотез H_0 и H_1 .

Рассмотрим две простые гипотезы на выборке x_1, x_2, \dots, x_n $H_0 : F(x) = F_0(x)$ и $H_1 : F(x) = F_1(x)$, где $F_0(x)$ и $F_1(x)$ - известные функции распределения. В этом случае равномерно наиболее мощный критерий называется критерием отношения правдоподобия и описывается следующим образом. Введем статистику

$$\Lambda(x_1, x_2, \dots, x_n) = \frac{L_1(x_1, x_2, \dots, x_n)}{L_0(x_1, x_2, \dots, x_n)}, \quad (5.2.2)$$

где $L_0(x_1, x_2, \dots, x_n) = f(x_1) \cdot f(x_2) \cdot \dots \cdot f(x_n)$ для непрерывной случайной величины X и $L_0(x_1, x_2, \dots, x_n) = P(x_1) \cdot P(x_2) \cdot \dots \cdot P(x_n)$ для дискретной. Статистика $\Lambda(x_1, x_2, \dots, x_n)$ носит название отношения правдоподобия и является отношением вероятностей (или плотностей распределения) получить выборку x_1, x_2, \dots, x_n при условии справедливости гипотез H_0 и H_1 . Естественно предположить, что чем больше отношение правдоподобия, тем большее предпочтение мы должны оказать гипотезе H_1 . Об этом говорится в лемме Неймана - Пирсона.

Лемма Неймана – Пирсона. Среди всех критериев заданного уровня значимости α , проверяющих две простые гипотезы H_0 и H_1 , критерий отношения правдоподобия является наиболее мощным.

При практической реализации критерия отношения правдоподобия обычно удобно пользоваться не отношением правдоподобия, а его логарифмом.

рифмом. В этом случае мы должны принять гипотезу H_0 , если $\Lambda = \Lambda(x_1, x_2, \dots, x_n) \leq C = \ln C_1$, и отвергнуть ее, т.е. принять H_1 , если $\Lambda > C$. В соответствии с общим правилом уровень значимости α и мощность $1-\beta$ критерия отношения правдоподобия в зависимости от критического значения C определяются по формулам:

$$\begin{cases} \alpha = \alpha(C) = P(\Lambda(x_1, x_2, \dots, x_n) > C/H_0) = \\ \quad = \int \dots \int_{\Lambda(x_1, x_2, \dots, x_n) > C} f_0(x_1) \cdot f_0(x_2) \cdot \dots \cdot f_0(x_n) dx_1 dx_2 \dots dx_n, \\ \beta = \beta(C) = P(\Lambda(x_1, x_2, \dots, x_n) > C/H_1) = \\ \quad = \int \dots \int_{\Lambda(x_1, x_2, \dots, x_n) > C} f_1(x_1) \cdot f_1(x_2) \cdot \dots \cdot f_1(x_n) dx_1 dx_2 \dots dx_n. \end{cases} \quad (5.2.3)$$

Пример. Пусть $x_1, x_2, \dots, x_n \in N(m, D)$ и $H_0 : m = a_0$, $H_1 : m = a_1 > a_0$. Воспользуемся критерием Неймана – Пирсона. Критическая область ω для гипотезы H_0 определена тогда, когда

$$\Lambda(x_1, x_2, \dots, x_n) > C. \quad \text{В нашем случае} \quad f_0(x_i) = \frac{1}{\sqrt{2\pi D}} e^{-\frac{(x_i - a_0)^2}{2D}},$$

$$f_1(x_i) = \frac{1}{\sqrt{2\pi D}} e^{-\frac{(x_i - a_1)^2}{2D}}. \quad \text{Тогда} \quad \Lambda(x_1, x_2, \dots, x_n) = \frac{\left(\frac{1}{\sqrt{2\pi D}}\right)^n e^{-\frac{1}{2D} \sum_{i=1}^n (x_i - a_1)^2}}{\left(\frac{1}{\sqrt{2\pi D}}\right)^n e^{-\frac{1}{2D} \sum_{i=1}^n (x_i - a_0)^2}} > C.$$

Упростим последнее выражение:

$$e^{\frac{1}{2D} \left[\sum_{i=1}^n (x_i - a_0)^2 - \sum_{i=1}^n (x_i - a_1)^2 \right]} > C, \quad \sum_{i=1}^n \left[(x_i - a_0)^2 - (x_i - a_1)^2 \right] > 2D \ln C,$$

$$\begin{aligned} \sum_{i=1}^n \left[(x_i - a_0)^2 - (x_i - a_1)^2 \right] &= \sum_{i=1}^n (x_i^2 - 2x_i a_0 + a_0^2 - x_i^2 + 2x_i a_1 - a_1^2) = \\ &= 2 \sum_{i=1}^n x_i (a_0 - a_1) + n(a_0^2 - a_1^2) > 2D \ln C, \quad \frac{1}{n} \sum_{i=1}^n x_i = m_X^* = \\ &= \frac{1}{n} \frac{2D \ln C - n(a_0^2 - a_1^2)}{2(a_0 - a_1)} = \varphi(C, D, a_0, a_1) = C_1. \quad \text{Итак, } m_X^* > C_1, \text{ а так как} \end{aligned}$$

$m_X^* \in N(a_i, D/n)$, то можно по этому неравенству найти C_1 , зная α , на-

пример, $\alpha = P\left(\frac{m_X^* - a_0}{\sqrt{D/n}} > \frac{C_1 - a_0}{\sqrt{D/n}}\right) = \frac{1}{\sqrt{2\pi}} \int_{\frac{C_1 - a_0}{\sqrt{D/n}}}^{\infty} e^{-\frac{t^2}{2}} dt = 1 - \Phi\left(\frac{C_1 - a_0}{\sqrt{D/n}}\right).$

Таким образом, по α находится C_1 из решения уравнения

$$1 - \Phi\left(\frac{C_1 - a_0}{\sqrt{D/n}}\right) = \alpha. \text{ Кроме того, можно найти и } \beta \text{ из аналогичного равен-$$

$$\text{ства } \beta = P\left(\frac{m_X^* - a_1}{\sqrt{D/n}} \leq \frac{C_1 - a_1}{\sqrt{D/n}}\right) = \Phi\left(\frac{C_1 - a_1}{\sqrt{D/n}}\right).$$

5.3. Проверка гипотез о числовых значениях параметров нормального распределения

Обозначим через X случайную величину, имеющую нормальный закон распределения с параметрами m_X и D_X , т.е. $X \in N(m_X, D_X)$, причем числовые значения либо одного, либо обоих параметров неизвестны. Узнать, каково численное значение неизвестного параметра, можно, обследовав всю генеральную совокупность, что сделать, как правило, нельзя.

Обычно вместо этого проводят выборочные наблюдения, предполагая при этом, что они независимы и проводятся в одинаковых условиях. Тогда

несмещенными оценками m_X и D_X являются $m_X^* = \frac{1}{n} \sum_{i=1}^n x_i$ и

$$\hat{D}_X = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_X^*)^2. \text{ Затем приступают к проверке гипотез.}$$

1. Проверка гипотезы о числовом значении математического ожидания нормального распределения при известной дисперсии

Нулевая гипотеза здесь $H_0: m_X = a_0$, а альтернативная гипотеза может быть сформулирована в трех видах 1) $H_1: m_X = a_1 > a_0$, 2) $H_1: m_X = a_1 < a_0$, 3) $H_1: m_X = a_1 \neq a_0$.

Зададим уровень значимости критерия α , а так как D_X известна, то в качестве статистики критерия можно взять случайную величину

$$z = \frac{m_X - a_0}{\sqrt{D_X/n}}. \text{ Так как } m_X \in N(a_0, D_X/n), \text{ что было уже несколько раз}$$

показано ранее, ибо $x_i \in N(a_0, D_X)$, то $z \in N(0,1)$.

Выделим критическую область ω статистики z , при которой H_0 отвергается. Размер и расположение критической области зависят от форму-

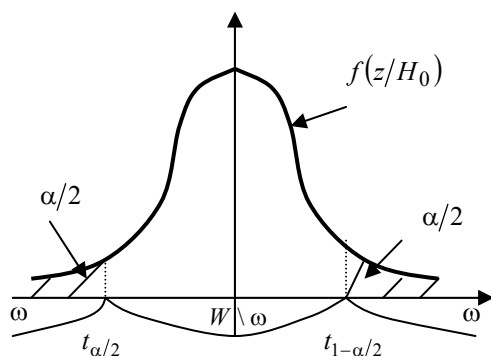


Рис. 5.6. Двусторонняя критическая область для матожидания

ального распределения (см. формулу (1.4.5)), т.е. $z_{\text{лев}, \alpha/2} = t_{\alpha/2} = \Phi^{-1}(\alpha/2)$, а $z_{\text{пр}, \alpha/2} = t_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

Далее по выборке находим выборочное значение статистики критерия

$$z_{\text{в}} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Если $z_{\text{в}} \in \omega$, гипотеза H_0 отвергается с уровнем значимо-

сти α и принимается гипотеза H_1 . Если же $z_{\text{в}} \in W \setminus \omega$, то гипотеза H_0 принимается.

2. Проверка гипотезы о числовом значении математического ожидания нормального распределения при неизвестной дисперсии.

В этом случае отличие от предыдущих формул и предположений будет касаться лишь статистики критерия z и ее распределения. Выберем в

качестве статистики величину $z = \frac{(m_X - a_0)}{\sqrt{\bar{D}_X/n}}$. Как было уже показано

ранее (см. подразд. 4.7, п. 2), эта статистика имеет распределение Стьюдента с $(n-1)$ -й степенью свободы, т.е. $z \in S_{n-1}(t)$. Все остальные пункты проверки остаются без изменений. Например, если выбрана альтернативная гипотеза 2-го вида $H_1 : m_X = a_1 < a_0$ (рис. 5.7), критическая область будет левосторонней, ее образует один интервал $(-\infty, z_{\text{лев}, \alpha})$,

лировки альтернативной гипотезы. Рассмотрим 3-й случай $H_1 : m_X = a_1 \neq a_0$,

здесь целесообразно выбрать двусторонний критерий (рис. 5.6). Критическую область образуют два интервала $(-\infty, z_{\text{лев}, \alpha/2})$ и $(z_{\text{пр}, \alpha/2}, +\infty)$. Критические

точки определяются из условий $P(z < z_{\text{лев}, \alpha/2}) = \alpha/2$ и

$P(z > z_{\text{пр}, \alpha/2}) = \alpha/2$. Так как

$z \in N(0,1)$, то критические

точки – это квантили нор-

мального распределения (см. формулу (1.4.5)), т.е.

$z_{\text{лев}, \alpha/2} = t_{\alpha/2} = \Phi^{-1}(\alpha/2)$, а $z_{\text{пр}, \alpha/2} = t_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

где точка $z_{\text{лев},\alpha}$ есть квантиль распределения Стьюдента. Он определяется из условия $P(z < z_{\text{лев},\alpha} = t_{\alpha,n-1}) = \alpha$

или $\int_{-\infty}^{t_{\alpha,n-1}} s(t) dt = \alpha$, т.е.

$$t_{\alpha,n-1} = S^{-1}(\alpha)$$

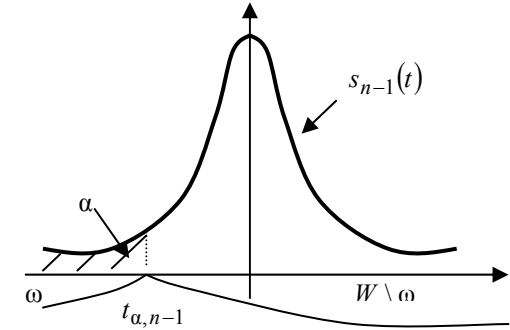


Рис. 5.7. Левосторонняя критическая область для матожидания

3. Проверка гипотезы о числовом значении дисперсии нормального распределения.

Итак, в этом случае известно, что $X \in N(m_X, D)$, но числовое значение дисперсии неизвестно. По выборке наблюдений x_1, x_2, \dots, x_n вычислим

точечные оценки $m_X^* = \frac{1}{n} \sum_{i=1}^n x_i$ и $\hat{D}_X = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_X^*)^2$ и проверим гипотезу $H_0 : D_X = D_0$, где D_0 - заранее заданное число. В качестве статисти

стики такой гипотезы следует взять случайную величину $z = \hat{D}_X(n-1)/D_0$. Ранее (см. подразд. 4.7) было показано, что эта случайная величина имеет χ^2 -распределение с $n-1$ степенью свободы, т.е.

$z \in \chi_{n-1}^2$.

После выбора статистики z и определения ее распределения все остальные вопросы проверки гипотезы носят технический характер. Зададимся уровнем значимости α , сформулируем альтернативную гипотезу и перейдем к построению критической области и проверке H_0 . Рассмотрим правосторонний критерий, т.е. альтернативная гипотеза должна быть сформулирована в виде $H_1 : D_X > D_0$, (рис. 5.8). Критическую область образует один интервал $(z_{\text{пр},1-\alpha}, +\infty)$, где точка $z_{\text{пр},1-\alpha}$ есть $1-\alpha$ - процентный квантиль χ^2 -распределения, определяется из условия

$P(z > z_{\text{пр},1-\alpha}) = \alpha$ или $\int_{z_{\text{пр},1-\alpha}}^{\infty} k_{n-1}(t) dt = \alpha$, т.е. $z_{\text{пр},1-\alpha} = K^{-1}(1-\alpha)$. Далее

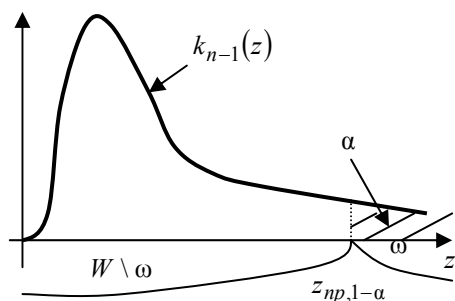


Рис. 5.8. Правосторонняя критическая область для дисперсии

можно вычислить выборочное значение статистики $z_B = \hat{D}_X(n-1)/D_0$ и сравнить ее с критической точкой $z_{np, 1-\alpha}$. Если $z_B \geq z_{np, 1-\alpha}$, гипотезу H_0 следует отклонить, если же $z_B < z_{np, 1-\alpha}$, гипотеза H_0 принимается с уровнем значимости α .

5.4. Проверка гипотез о параметрах двух нормальных распределений

1. Проверка гипотезы о равенстве математических ожиданий двух нормальных распределений с известными дисперсиями.

Проверка гипотезы о равенстве математических ожиданий двух нормальных совокупностей имеет важное практическое значение. Часто возникает вопрос, можно ли отличие двух средних, полученных по двум разным выборкам, объяснить случайной ошибкой экспериментов или оно не случайно? Подобная задача возникает, например, при сравнении качества изделий, изготовленных на разных установках.

Пусть x_1, x_2, \dots, x_{n_1} - первая выборка, y_1, y_2, \dots, y_{n_2} - вторая выборка и $x_i \in N(m_X, D_X)$, $y_j \in N(m_Y, D_Y)$, причем D_X и D_Y должны быть известны. Основная проверяемая гипотеза в этих условиях имеет вид $H_0: m_X = m_Y$.

Вычислим оценки математических ожиданий $m_X^* = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$ и

$m_Y^* = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i$. Очевидно, что $m_X^* \in N(m_X, D_X/n_1)$, а

$m_Y^* \in N(m_Y, D_Y/n_2)$. Тогда из свойств математического ожидания и дисперсии независимых случайных величин следует, что $M(m_X^* - m_Y^*) = M(m_X^*) - M(m_Y^*) = m_X - m_Y$, а $D(m_X^* - m_Y^*) = D(m_X^*) + D(m_Y^*) = D_X/n_1 + D_Y/n_2$.

Таким образом, в силу теоремы о суммировании нормально распределенных случайных величин будем иметь $m_X^* - m_Y^* \in N(m_X - m_Y, D_X/n_1 + D_Y/n_2)$. Тогда нормированная и центрированная случайная величина будет подчинена стандартному нормальному распределению, т.е. $z = \frac{(m_X^* - m_Y^*) - (m_X - m_Y)}{\sqrt{D_X/n_1 + D_Y/n_2}} \in N(0,1)$. Эту статистику

и выбирают за рабочую при проверке нулевой гипотезы $H_0 : m_X = m_Y$. Если H_0 выполняется, то $m_X - m_Y = 0$ и рабочая статистика упрощается $z = \frac{(m_X^* - m_Y^*)}{\sqrt{D_X/n_1 + D_Y/n_2}} \in N(0,1)$.

Дальнейшие действия стандартны и практически совпадают с аналогичными действиями при проверке гипотезы о равенстве математического ожидания выборки определенному значению при известной дисперсии (см. подразд. 5.3, п.1). Задаем α и строим левостороннюю критическую область $(-\infty, z_{\text{лев}, \alpha}) = (-\infty, t_\alpha)$, где t_α - α -процентный квантиль стандартного нормального распределения. Затем находим выборочную статистику

$$z_{\text{в}} = \frac{(m_X^* - m_Y^*)}{\sqrt{D_X/n_1 + D_Y/n_2}}.$$

Если $z_{\text{в}} < t_\alpha$, гипотеза H_0 должна быть отвергнута (рис. 5.9). В случае правостороннего и двустороннего критериев выполняется комплекс аналогичных действий.

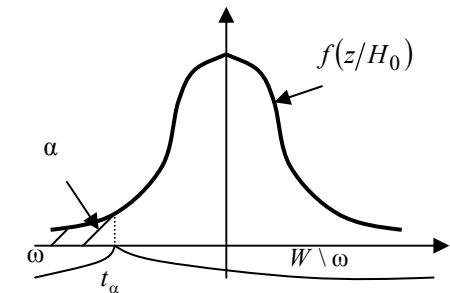


Рис. 5.9. Левосторонняя критическая область статистики z при $H_0 : m_X = m_Y$

2. Проверка гипотезы о равенстве математических ожиданий двух нормальных распределений с неизвестными, но равными дисперсиями.

Пусть имеются две случайные величины $X \in N(m_X, D)$ и $Y \in N(m_Y, D)$ с одинаковыми дисперсиями D , однако числовое значение D неизвестно, неизвестны также и числовые значения математических ожиданий m_X и m_Y . Пусть имеются две выборки этих случайных величин x_1, x_2, \dots, x_{n_1} и

y_1, y_2, \dots, y_{n_2} . Тогда $m_X^* \in N(m_X, D/n_1)$, $m_Y^* \in N(m_Y, D/n_2)$, кроме того $\hat{D}_X(n_1 - 1)/D \in \chi^2_{n_1 - 1}$ и $\hat{D}_Y(n_2 - 1)/D \in \chi^2_{n_2 - 1}$.

Наблюдения организованы так, что результаты и y_1, y_2, \dots, y_{n_2} независимы. Из этого условия следует, что m_X и m_Y независимы, \hat{D}_X и \hat{D}_Y также независимы. Требуется проверить гипотезу $H_1 : m_X > m_Y$.

Подберем подходящую статистику для этого критерия. По предыдущему пункту, очевидно, $\frac{(m_X^* - m_Y^*) - (m_X - m_Y)}{\sqrt{D/n_1 + D/n_2}} \in N(0,1)$. Кроме того, по свойству χ^2 -распределения имеем:

$$\text{если } \frac{\hat{D}_X(n_1 - 1)}{D} \in \chi^2_{n_1 - 1} \quad \text{и} \quad \frac{\hat{D}_Y(n_2 - 1)}{D} \in \chi^2_{n_2 - 1}, \quad \text{то}$$

$$\frac{\hat{D}_X(n_1 - 1)}{D} + \frac{\hat{D}_Y(n_2 - 1)}{D} \in \chi^2_{n_1 + n_2 - 2}.$$

Вспомним способ получения статистики распределения Стьюдента (см. подразд. 2.2): $t = z\sqrt{n}/\sqrt{v}$, где $z \in N(0,1)$, а $v \in \chi^2_n$. В нашем случае

$$z = \frac{(m_X^* - m_Y^*) - (m_X - m_Y)}{\sqrt{D/n_1 + D/n_2}}, \quad \text{а } v_{n_1 + n_2 - 2} = \frac{\hat{D}_X(n_1 - 1)}{D} + \frac{\hat{D}_Y(n_2 - 1)}{D}. \quad \text{Тогда}$$

$$t_{n_1 + n_2 - 2} = \frac{\frac{(m_X^* - m_Y^*) - (m_X - m_Y)}{\sqrt{D/n_1 + D/n_2}} \cdot \sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{\hat{D}_X(n_1 - 1) + \hat{D}_Y(n_2 - 1)}{D}}} =$$

$$= \frac{(m_X^* - m_Y^*) - (m_X - m_Y)}{\sqrt{(1/n_1 + 1/n_2) \frac{\hat{D}_X(n_1 - 1) + \hat{D}_Y(n_2 - 1)}{n_1 + n_2 - 2}}}. \quad \text{Если гипотеза } H_0 : m_X = m_Y$$

выполняется, то $m_X - m_Y = 0$ и вид статистики упрощается

$$t_{n_1 + n_2 - 2} = \frac{(m_X^* - m_Y^*)}{\sqrt{(1/n_1 + 1/n_2) \frac{\hat{D}_X(n_1 - 1) + \hat{D}_Y(n_2 - 1)}{n_1 + n_2 - 2}}}.$$

Итак, рабочая статистика получена. Зададим уровень значимости α и перейдем к построению критической области. Выберем правосторонний критерий, т.е. альтернативная гипотеза будет иметь вид

$H_1 : m_X > m_Y$. Правосторонняя критическая область состоит из интервала

$(z_{\text{пр}, 1-\alpha}, +\infty) = (t_{1-\alpha, n_1+n_2-2}, +\infty)$, где $t_{1-\alpha, n_1+n_2-2}$ - $(1-\alpha)$ - процентный квантиль распределения Стьюдента (рис. 5.10). Он определяется из условия

$P(z > z_{\text{пр}, 1-\alpha} = t_{1-\alpha, n_1+n_2-2}) = \alpha$

или $\int_{t_{1-\alpha, n_1+n_2-2}}^{+\infty} s(t) dt = \alpha$, т.е.

$t_{1-\alpha, n_1+n_2-2} = S^{-1}(1-\alpha)$. Если выборочная статистика

$z_{\text{в}} = \frac{(m_X^* - m_Y^*)}{\sqrt{(1/n_1 + 1/n_2) \frac{\bar{D}_X(n_1-1) + \bar{D}_Y(n_2-1)}{n_1+n_2-2}}}$, где $m_X^* = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$,

$m_Y^* = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i$, $\bar{D}_X = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_i - m_X^*)^2$, $\bar{D}_Y = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (y_i - m_Y^*)^2$ не

превысит $t_{1-\alpha, n_1+n_2-2}$, то гипотезу H_0 следует принять с уровнем значимости α .

3. Проверка гипотезы о равенстве дисперсий двух нормальных распределений.

Задача проверки гипотезы о равенстве дисперсий часто возникает на практике. Дисперсия характеризует точность работы приборов. Убедившись в равенстве двух дисперсий, можно быть уверенным, например, что два прибора, два технологических процесса обеспечивают одинаковую точность.

Пусть x_1, x_2, \dots, x_{n_1} - результаты независимых наблюдений случайной величины X , а y_1, y_2, \dots, y_{n_2} - случайной величины Y . Все наблюдения проводятся в одинаковых условиях и организованы так, что результаты

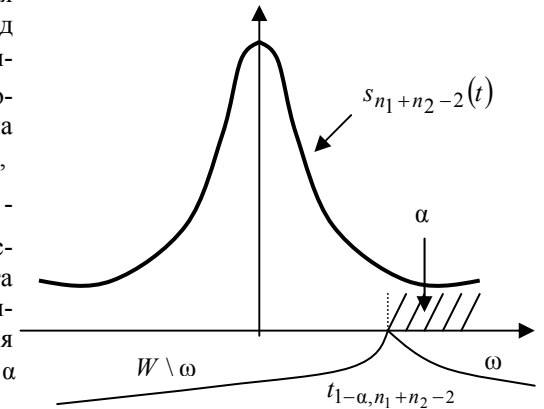


Рис. 5.10. Правосторонняя критическая область статистики z при $H_0 : m_X = m_Y$

обеих выборок независимы. При этих условиях требуется проверить нулевую гипотезу $H_0 : D_X = D_Y$.

Построим критерий для проверки этой гипотезы. Пусть

$$m_X^* = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i, \quad m_Y^* = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i, \quad \hat{D}_X = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - m_X^*)^2,$$

$$\hat{D}_Y = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - m_Y^*)^2$$

- несмещенные оценки матожиданий и дисперсий случайных величин X и Y по используемым выборкам. В подразд. 5.3 уже использовалась статистика $v = \frac{\hat{D}(n-1)}{D} \in \chi_{n-1}^2$.



Рис. 5.11. Двусторонняя критическая область для проверки гипотезы о равенстве дисперсий двух нормальных распределений

В соответствии с определением F -распределения (см. подразд. 2.3) отношение $(\chi_l^2/l)/(\chi_k^2/k)$ имеет F -распределение с l и k степенями свободы. В нашем случае дробь

$$\frac{\frac{\hat{D}_X(n_1-1)}{D_X} / n_1 - 1}{\frac{\hat{D}_Y(n_2-1)}{D_Y} / n_2 - 1} = \frac{\hat{D}_X/D_X}{\hat{D}_Y/D_Y}$$

будет

иметь F -распределение с $n_1 - 1$ и $n_2 - 1$ степенями свободы. Если гипотеза H_0 верна, то $D_X = D_Y$ и для статистики z справедливо соотношение $z = \hat{D}_X/\hat{D}_Y \in F_{n_1-1, n_2-1}$.

Рассмотрим при заданном α двусторонний критерий, т.е. $H_1 : D_X \neq D_Y$ (рис. 5.11). В этом случае критическая область состоит из двух интервалов $(0, z_{\text{лев}, \alpha/2})$ и

$(z_{\text{прав}, \alpha/2}, +\infty)$, где критические точки находятся по следующим схемам:

$$P(0 < z < z_{\text{лев}, \alpha/2}) = \alpha/2, \quad \int_0^{z_{\text{лев}, \alpha/2}} f_F(x) dx = \frac{\alpha}{2}, \quad P(z > z_{\text{прав}, \alpha/2}) = \frac{\alpha}{2},$$

$$\int_{z_{\text{пр}, \alpha/2}}^{\infty} f_F(x) dx = \frac{\alpha}{2}. \text{ Если } z_{\text{в}} = \hat{D}_X / \hat{D}_Y < z_{\text{лев}, \alpha/2} \text{ или } z_{\text{в}} > z_{\text{пр}, \alpha/2}, \text{ гипотеза}$$

за H_0 должна быть отвергнута. Критические точки находятся по таблице F -распределения.

5.5. Лабораторная работа № 5. Проверка статистических гипотез о числовых значениях нормальных распределений в математических пакетах STATGRAPHICS и MATHCAD

Для исследования подчиняющихся нормальному распределению данных математической статистикой выработаны эффективные методы. Строго говоря, эти методы непригодны для данных другой природы. Поэтому перед применением этих методов к имеющимся наблюдениям полезно выяснить, похоже ли их распределение на нормальное. С полной уверенностью сказать этого все равно невозможно, но от грубых ошибок такие проверки могут уберечь.

Так как конкретное нормальное распределение полностью задается значением параметров m_X и D_X , рассмотрим задачу проверки гипотезы о значениях параметров нормального распределения, тесно связанную с построением доверительных интервалов для этих параметров.

В пакете STATGRAPHICS Plus for Windows часть процедур для анализа нормальных выборок собрана в разделе Describe (Описание данных), часть в разделе Compare (Сравнение данных). Проверим все основные гипотезы, описанные в подразд. 5.3 и 5.4. Сначала займемся одной выборкой.

Смоделируем нормальную выборку объемом 50 единиц с параметрами, например, $m_X = 6$, $D_X = 10$, т.е. $\sigma_X = 3.162$. Моделирование выборок с произвольными распределениями в пакете STATGRAPHICS мы занимались в лабораторных работах № 2 и № 4. Опишем кратко последовательность необходимых действий.

В головном меню пакета выбираем пункт Plot→Probability Distribution и помечаем распределение № 17 – нормальное (оно помечено по умолчанию). После щелчка по кнопке ОК открывается заставка нормального распределения. Щелчок правой кнопкой мыши в любом месте этой заставки открывает дополнительное меню, в котором выбираем пункт Analysis Options и задаем лишь одно распределение из возможных пяти, введя значение средней (Mean), равное 6.0, и стандартного отклонения (Std. Deviation), равного 3.162. Затем в меню заставки нормального распределения выберем пункт Tabular Options и поставим галочку в поле

Random Numbers. После щелчка по кнопке ОК будет смоделирована нормальная выборка с заданными параметрами объемом 100 единиц. Уменьшив ее объем в два раза. Для этого щелкнем правой кнопкой мыши в поле заставки Random Numbers и выберем в дополнительном меню пункт Pane Options. В поле Size (Объем) введем число 50.

Сохраним эту выборку под именем NORM. Для этого воспользуемся пунктом дополнительного меню Save Results. В поле Save поставим галочку, а в поле Target Variables (Плановые переменные) наберем имя выборки.

Все описанные действия проделаем еще раз и смоделируем равномерную выборку на интервале (0,0.5). В качестве параметров распределения в поле Lower Limit (Нижний предел) введем ноль, а в поле Upper Limit (Верхний предел) число 0.5. Сохраним равномерное распределение под именем UNIF. Сохраненные данные можно просмотреть. Для этого в нижней части экрана в меню электронных таблиц, комментариев, статконсультанта и составления статистических отчетов нужно развернуть на полный экран пиктограмму untitled.

Для проверки гипотез о числовых значениях математического ожидания и дисперсии засорим нормальную выборку; для этого сложим два полученных распределения NORM и UNIF. Сложение выборок можно провести следующим образом.

В головном меню пакета выберем пункты Describe→Numeric Data→One-Variable Analysis (Анализ одной переменной). В дополнительном меню анализа одной переменной нажмем кнопку Transform (Преобразования). Появится следующая заставка (рис. 5.12). В строке Expression

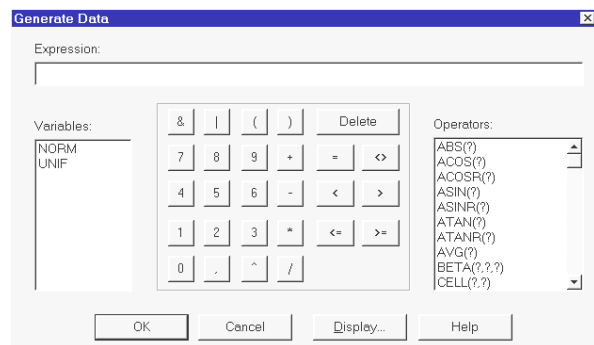


Рис. 5.12. Панель для генерации новых данных

(Выражение) необходимо набрать NORM+UNIF. Появится заставка новой выборки, члены которой представляют собой сумму соответствующих по номеру членов обеих исходных выборок. Заметим, что в появившемся окне диалога можно проводить арифметиче-

ские, логические и другие манипуляции с переменными посредством более ста предоставляемых операторов.

В пункте меню Tabular Options поставим галочки в полях Analysis Summary, Summary Statistics и Hypothesis Tests. После нажатия кнопки OK

на экран будет выведена следующая информация (рис. 5.13). В пункте главного меню Describe, где мы сейчас находимся, проверяется только гипотеза о числовом значении математического ожидания при неизвестной дисперсии, причем по умолчанию задается $\alpha = 0.05$, $H_0 : m_X = 0$, двусторонний критерий.

Чтобы изменить установки гипотезы щелчком правой кнопкой мыши в поле заставки Hypothesis Tests for NORM+UNIF. Появится дополнительное меню (рис. 5.14).

В поле Mean введем проверяемое значение среднего 6.0. Можно при желании изменить вид альтернативной гипотезы (Not Equal - двусторонний критерий, т.е.

$H_1 : m_X \neq 6.0$; Less Than - левосторонний критерий, $H_1 : m_X < 6.0$; Greater Than - правосторонний критерий, $H_1 : m_X > 6.0$). Выберем двусторонний критерий и оставим значение $\alpha = 0.05$. После щелчка по кнопке OK информация в поле заставки Hypothesis Tests for NORM+UNIF изменится (рис. 5.15).

```
Analysis Summary
Data variable: NORM+UNIF
50 values ranging from -1,38362 to 14,6087

Summary Statistics for NORM+UNIF

Count = 50
Average = 6,49235
Variance = 10,0711
Standard deviation = 3,1735
Minimum = -1,38362
Maximum = 14,6087
Std. skewness = -0,313651
Std. kurtosis = 0,0536109
Sum = 324,618

Hypothesis Tests for NORM+UNIF

Sample mean = 6,49235
Sample median = 6,92332

t-test
-----
Null hypothesis: mean = 0,0
Alternative: not equal

Computed t statistic = 14,466
P-Value = 0,0

Reject the null hypothesis for alpha = 0,05.
```

Рис. 5.13. Информация о числовых характеристиках нормального распределения и проверке гипотезы о матожидании

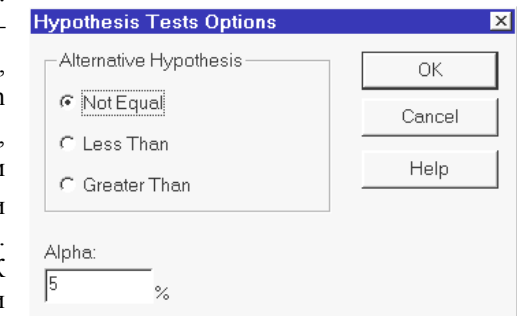


Рис. 5.14. Окно диалога для задания вида критерия значимости

Далее идут данные по непараметрическому критерию знаков для медианы.

```
Hypothesis Tests for NORM+UNIF
Sample mean = 6,49235
Sample median = 6,92332


t-test
-----
Null hypothesis: mean = 6,0
Alternative: not equal

Computed t statistic = 1,09704
P-Value = 0,277985

Do not reject the null hypothesis for alpha = 0,05.
```

Рис. 5.15. Исправленная информация о проверке гипотезы о матожидании

В результате проверки процедура выдает значение t -статистики Стьюдента, ее минимальный уровень и заключение о принятии нулевой гипотезы. При желании часть информации можно сохранить с помощью инструмента составления статистических отчетов – StatGallery. Как это делается, показано в лабораторной работе № 2. При этом следует помнить, что каждая

следующая страница статистического анализа, бывшая когда-то на экране дисплея, открывается нажатием клавиш Ctrl+F6 или щелчком по пиктограмме .

Основная часть процедур проверки гипотез пакета STATGRAPHICS сосредоточена в разделе Describe→Hypothesis Tests. При таком выборе появляется дополнительное меню (рис. 5.16). При выборе в качестве параметра Normal Mean проверяются гипотезы о числовом значении математического ожидания нормальной выборки, Normal Sigma о числовом значении дисперсии, Binomial Proportion – проверяются гипотезы о биномиальных долях в схеме Бернулли, Poisson Rate – об интенсивностях пуассоновских потоков. Эта процедура не имеет отношения к анализу нормальных выборок.

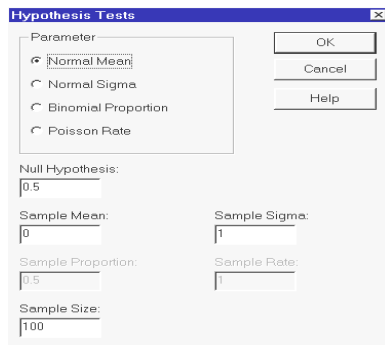


Рис. 5.16. Дополнительное меню проверки гипотез

При задании параметра Normal Mean на экран выводится совершенно аналогичная полученной ранее ин-

формация с единственным отличием: дополнительно выдаются границы 95% доверительного интервала для математического ожидания.

При выборе параметра Normal Sigma проверяются гипотезы о числовом значении дисперсии. Зададим следующие параметры: в поле Null Hypothesis значение 3.0 ($H_0 : \sigma_X = 3.0$), в поле Sample Sigma значение выборочного стандартного отклонения выборки NORM + UNIF, равного 3.1735, в поле

Sample Size объем выборки, равный 50. Кроме того, выберем левосторонний критерий значимости при $\alpha = 0.05$. На экране появится следующая информация (рис. 5.17). Рассмотрим теперь вопросы, связанные с двумя нормальными выборками. В пакете STATGRAPHICS процедуры проверки гипотез о числовых значениях параметров двух нормальных выборок находятся в разделе Compare→Two Samples→Hypothesis Tests. При проверке гипотезы о равенстве средних двух выборок нулевая гипотеза формулируется относительно разности этих средних, т.е. в виде $H_0 : m_X - m_Y = 0$. Дополнительное меню Hypothesis Tests очень похоже на описанное выше.

```

Hypothesis Tests
-----
Sample standard deviation = 3,1735
Sample size = 50

95,0% upper confidence bound for sigma: {3,81367}

Null Hypothesis: standard deviation = 3,0
Alternative: less than
Computed chi-squared statistic = 54,8316
P-Value = 0,736951
Do not reject the null hypothesis for alpha = 0,05.

```

Рис. 5.17. Информация о проверке гипотезы о значении среднеквадратического отклонения

Для проверки оставшихся (см. подразд. 5.4) гипотез смоделируем еще одну нормальную выборку объемом 100 единиц с параметрами $m_X = 7.0$, $D_X = 10.5$, т.е. $\sigma_X = 3.240$, и сохраним ее под именем NORM1.

Затем воспользуемся пунктом головного меню Describe→Numeric Data →Multiple-Variable Analysis (Анализ многих переменных) и в дополнительном меню в пункте Data отметим две имеющиеся нормальные выборки NORM и NORM1. После щелчка по кнопке ОК откроется заставка Analysis Summary для этих выборок, выберем пункт дополнительного меню Tabular Options и отметим галочками разделы Analysis Summary и Summary Statistics.

Для того чтобы выборочные статистики считались по всем данным (у нас выборки разной длины), необходимо, щелкнув правой кнопкой мыши в поле заставки Summary Statistics, выбрать пункт Analysis Options и в открывшемся подменю выбрать пункт All Data (Все данные) (рис. 5.18). В результате будут вычислены некоторые точечные характеристики обеих выборок и информация об этом выведена на экран.

Теперь можно проверить гипотезу о равенстве средних этих двух выборок. Выберем Compare→Two Samples→Hypothesis Tests, в открывшемся подменю по-

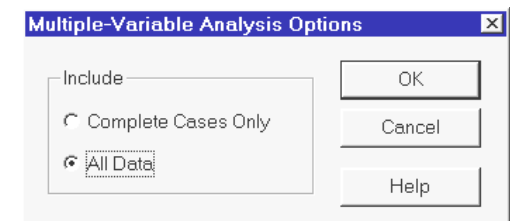


Рис. 5.18. Окно диалога для настройки проводимого анализа

метим пункт Normal Mean, в поле Null Hypothesis for Difference Normal Mean введем ноль, в поля значений средних и средних квадратических отклонений введем числовые значения соответствующих оценок из таблицы Summary Statistics. Наконец, в поле Sample 1 Size введем число 50. После щелчка по кнопке ОК получим информацию, помещенную чуть ниже под заголовком Hypothesis tests. Последняя запись напоминает, что гипотеза проводилась в предположении равенства дисперсий (рис. 5.19). Щелкнув правой кнопкой мыши в поле заставки Hypothesis Tests можно в

```
Summary Statistics
-----
                NORM                NORM1
-----
Count           50                  100
Average         6,23962             7,19464
Variance        10,1913             9,8627
Standard deviation 3,19238          3,14049
Minimum         -1,80021            -1,26979
Maximum         14,4547             15,9236
Std. skewness   -0,314919           0,126714
Std. kurtosis   0,135699            0,30053
Sum             311,981             719,464
-----

Hypothesis Tests
-----
Sample means = 6,23962 and 7,19464
Sample standard deviations = 3,19238 and 3,14049
Sample sizes = 50 and 100

95,0% confidence interval for difference between means: -0,95502 +/- 1,08082  [-2,03584;0,125804]

Null Hypothesis: difference between means = 0,0
Alternative: not equal
Computed t statistic = -1,74611
P-Value = 0,0828664
Do not reject the null hypothesis for alpha = 0,05.

(Equal variances assumed).
```

Рис. 5.19. Числовые характеристики двух выборок и результаты проверки гипотезы о равенстве средних

появившемся подменю сменить вид критерия значимости, установить новое значение α , а также сбросить флажок в поле Assume Equal Sigmas (В предположении равенства дисперсий), т. е. проверить гипотезу о равенстве математических ожиданий двух нормальных выборок с известными дисперсиями. Заметим, что если задать в этом случае $\alpha = 0.1$, нулевая гипотеза будет отвергнута.

Проверим, наконец, последнюю гипотезу о равенстве дисперсий двух нормальных выборок. Выберем Compare→Two Samples→Hypothesis Tests и заполним поля появившегося дополнительного меню. Пометим точкой

поле Normal Sigmas, в поле Null Hypothesis for Ratio of Variance введем 1.0. Здесь нулевая гипотеза формулируется в виде $H_0 : D_X / D_Y = 1$. Очевидно, что если дисперсии выборки равны, то их отношение равно единице. В поля Sample 1 Sigma и Sample 2 Sigma введем соответствующие выборочные значения 3.19238 и 3.14049, а в поле Sample 1 Size число 50. Результаты проверки этой гипотезы приведены на рис. 5.20.

Проверим теперь все уже рассмотренные шесть видов гипотез в пакете MATHCAD. В этом математическом (а не статистическом!) пакете все вспомогательные действия, выполняемые в STATGRAPHICS простым нажатием соответствующей кнопки, придется программировать и выполнять самостоятельно.

Hypothesis Tests

Sample standard deviations = 3,19238 and 3,14049
Sample sizes = 50 and 100
95,0% confidence interval for ratio of variances: [0,646881;1,71924]
Null Hypothesis: ratio of variances = 1,0
Alternative: not equal
Computed F statistic = 1,03332
P-Value = 0,87262
Do not reject the null hypothesis for alpha = 0,05.

Рис. 5.20. Результаты проверки гипотезы о равенстве средних

Смоделируем сначала три выборки: две нормальные и одну равномерную, аналогично тому, как делали в пакете STATGRAPHICS.

```

ORIGIN := 1
mx1 := 6  σx1 := 3.162  Dx1 := 10
mx2 := 7  σx1 := 3.240  Dx1 := 10.5
a := 0  b := 0.5  n := 50  n1 := 100
NORM := rnorm(n,mx1,σx1)  NORM1 := rnorm(n1,mx2,σx2)
UNIF := runif(n,a,b)
i := 1...50  NORMi := NORMi + UNIFi

```

Итак, все три выборки смоделированы, первая нормальная выборка засорена равномерным распределением. Приступим к проверке первой гипотезы о числовом значении матожидания при известной дисперсии, используя двусторонний критерий. Для этого за-

norm =		1
	1	4.612
	2	3.852
	3	4.503
	4	2.991
	5	0.67
	6	6.138
	7	5.619
	8	7.759
	9	12.93
	10	8.557
	11	9.115
	12	8.726
	13	8.895
	14	8.128
	15	2.698

norm1 =		1
	1	8.388
	2	5.46
	3	4.746
	4	16.876
	5	6.072
	6	2.853
	7	9.461
	8	5.728
	9	3.1
	10	3.641
	11	6.937
	12	7.104
	13	7.858
	14	8.593
	15	7.772

unif =		1
	1	0.335
	2	0.312
	3	0.399
	4	0.494
	5	0.315
	6	0.296
	7	0.253
	8	0.225
	9	0.47
	10	0.126
	11	0.365
	12	0.313
	13	0.321
	14	0.278
	15	0.223

программируем формулы подразд. 5.1 п. 1.

$$\alpha := 0.05 \quad xmean := \text{mean}(NORM) \quad xmean = 6.165 \quad xright := \text{qnorm}(1 - \alpha/2, 0, 1)$$

$$xleft := -xright \quad xleft = -1.960 \quad xright = 1.960$$

$$zb := \frac{xmean - mx1}{\sqrt{Dx1/n}} \quad zb = 0.368$$

Функция `qnorm` вычисляет квантили нормального распределения. Гипотеза $H_0 : M(NORM) = 6$ принимается, так как $xleft < zb < xright$, т.е. выборочная статистика критерия находится в области принятия решения.

Для проверки гипотезы о числовом значении матожидания при неизвестной дисперсии вычислим квантили распределения Стьюдента. Зададим правосторонний критерий:

$$Dx := \frac{n}{n-1} * \text{var}(NORM) \quad Dx = 8.135 \quad xright := \text{qt}(1 - \alpha, n - 1) \quad xright = 1.677$$

$$zb := \frac{xmean - mx1}{\sqrt{Dx/n}} \quad zb = 0.408$$

Так же как в предыдущем случае гипотеза $H_0 : M(NORM) = 6$ принимается, так как $zb = 0.408 < xright = 1.677$.

Последняя гипотеза о параметрах одной нормальной выборки – это гипотеза о числовом значении дисперсии. Выберем для нее левосторонний критерий:

$$xleft := \text{qchisq}(\alpha, n - 1) \quad xleft = 33.930 \quad zb := (n - 1) * \frac{Dx}{Dx1} \quad zb = 39.862$$

Гипотеза $H_0 : D_X = 10$ при левостороннем критерии принимается с уровнем значимости $\alpha = 0.05$ против альтернативы $H_1 : D_X < 10$, поскольку $zb = 39.862 > xleft = 33.930$.

Проверим теперь три описанные в подразд. 5.4 гипотезы о параметрах двух нормальных распределений. Первая из них о равенстве матожиданий, если обе дисперсии выборок известны. В нашем случае $Dx1 = 10$, $Dx2 = 10.5$, $mx1 = 6$, $mx2 = 7$, $H_0 : m_X = m_Y$, т.е.

$H_0 : mx1 = mx2$. Зададим опять левосторонний критерий:

$$xleft := \text{qnorm}(\alpha, 0, 1) \quad xleft = -1.645 \quad ymean := \text{mean}(NORM1) \quad ymean = 6.537$$

$$zb := \frac{xmean - ymean}{\sqrt{Dx1/n + Dx2/n1}} \quad zb = -0.675$$

Гипотеза H_0 принимается с уровнем значимости 0.05, так как $zb = -0.675 > xleft = -1.645$.

Следующая гипотеза о равенстве матожиданий двух нормальных выборок с неизвестными, но равными дисперсиями. В нашем случае $Dx1 = 10 \neq Dx2 = 10.5$, но значения очень близки друг к другу. Вычислим несмещенную оценку дисперсии второй нормальной выборки. Примем $H_1 : m_X > m_Y$, т.е. выберем правосторонний критерий:

$$Dy := \frac{n1}{n1-1} * \text{var}(NORM1) \quad Dy = 10.917 \quad xright := \text{qt}(1-\alpha, n+n1-2) \quad xright = 1.655$$

$$zb := \frac{xmean - ymean}{\sqrt{\left(\frac{1}{n} + \frac{1}{n1}\right) * \frac{(n-1)*Dx + (n1-1)*Dy}{n+n1-2}}} \quad zb = -0.681$$

Итак, мы имеем $zb = -0.681 < xright = 1.655$, т.е. основная гипотеза должна быть принята.

Наконец, последняя гипотеза о равенстве дисперсий двух нормальных выборок. Ее рабочая статистика имеет распределение Фишера. Выберем двусторонний критерий, т.е. проверим $H_0 : D_X = D_Y$, $H_1 : D_X \neq D_Y$:

$$xleft := \text{qF}\left(\frac{\alpha}{2}, n-1, n1-1\right) \quad xleft = 0.601 \quad xright := \text{qF}\left(1-\frac{\alpha}{2}, n-1, n1-1\right) \quad xright = 1.577$$

$$zb := \frac{Dx}{Dy} \quad zb = 0.745$$

Так как $xleft < zb < xright$, то гипотезу H_0 следует принять с уровнем значимости $\alpha = 0.05$.

Задание №1. Смоделировать две нормальные выборки со следующими параметрами: m_X равно порядковому номеру месяца вашего дня рождения, $m_Y = m_X + 1.5$; D_X равна номеру вашей фамилии в журнале преподавателя, $D_Y = D_X + 3$, объем первой выборки $n_X = 50$, объем второй выборки $n_Y = 100$. Не засоряя первую выборку, проверить в пакетах STATGRAPHICS или MATHCAD по вашему выбору все шесть описанных в подразд. 5.3 и 5.4 гипотез, приняв уровень значимости $\alpha = 0.1$.

5.6. Критерии согласия

1. Критерий χ^2 -Пирсона*. Во многих практических задачах модель закона распределения заранее не известна и возникает задача выбора модели, согласующейся с результатами наблюдений над случайной величиной. Предположим, что выборка x_1, x_2, \dots, x_n произведена из генеральной совокупности с неизвестной теоретической функцией распределения, относительно которой имеются две непараметрические гипотезы $H_0 : F(x) = F_0(x)$ и $H_1 : F(x) \neq F_0(x)$, где $F_0(x)$ - известная функция распределения. Таким образом, проверяется, согласуются ли эмпирические данные с гипотетическим предположением относительно теоретической функции распределения или нет. Поэтому критерии для проверки H_0 и H_1 носят название критериев согласия.

Критерий χ^2 -Пирсона предполагает, что результаты наблюдений сгруппированы в вариационный ряд. Поскольку при формулировке H_0 почти всегда необходимо оценивать несколько параметров закона, то последовательность действий такова.

1. Формулируют гипотезу о модели закона распределения случайной величины, по результатам наблюдений находят оценки неизвестных параметров этой модели.

2. Подставляют в модель закона оценки неизвестных параметров. В результате предполагаемая модель оказывается полностью определенной.

Пусть наблюдаемая случайная величина X принимает только значения b_1, b_2, \dots, b_k с неизвестными вероятностями p_1, p_2, \dots, p_k . Основная гипотеза H_0 выделяет среди всех распределений случайных величин, принимающих значения b_1, b_2, \dots, b_k , одно фиксированное распределение, для которого значения вероятностей известны и равны p_i . Обозначим через m_i , $i = 1, 2, \dots, k$ число тех элементов выборки x_1, x_2, \dots, x_n , которые приняли значение b_i . В силу закона больших чисел наблюдаемая частота $p_i^* = m_i/n$ с ростом объема n выборки стремится к вероятности p_i , гипотезу H_0 надо признать справедливой, если все p_i^* мало отличаются от p_i .

$$\text{Введем статистику } \chi^2 = \chi^2(x_1, x_2, \dots, x_n) = \sum_{i=1}^k \frac{(m_i - np_i)^2}{np_i}. \quad (5.6.1)$$

* Карл (Чарльз) Пирсон (1857-1936) – английский математик.

Эта статистика является мерой равномерной близости p_i^* к p_i . Кроме того, она соответствует мультиномиальной схеме, в результате которой появляется χ^2 -распределение. Именно пусть $\xi_1, \xi_2, \dots, \xi_n$ - независимые случайные величины, распределенные по нормальному закону с одинаковыми параметрами m и σ^2 . Если $\eta = 1/n(\xi_1 + \xi_2 + \dots + \xi_n)$, тогда $\chi^2 = \frac{1}{\sigma^2}[(\xi_1 - \eta)^2 + (\xi_2 - \eta)^2 + \dots + (\xi_n - \eta)^2]$ имеет χ^2 -распределение с

$n-1$ степенью свободы. Это стандартная схема получения χ^2 -распределения. Она же реализуется в мультиномиальной схеме.

Действительно, если m_i - наблюдаемые частоты, то np_i - теоретические значения соответствующих частот. Дисперсия же в мультиномиальной схеме, как известно, равна np_i . Можно еще добавить, что случайная величина $(m_i - np_i)/\sqrt{np_i}$ имеет распределение, близкое к нормальному (использованы операции центрирования и нормирования). Чтобы это утверждение было достаточно точным, необходимо, чтобы для всех i выполнялось условие $np_i \geq 5$.

Пусть производится n независимых одинаковых испытаний, в каждом из которых с вероятностью p_i может произойти одно из событий A_i , $i = \overline{1, k}$. m_i - число появлений события A_i . Тогда из многомерного аналога теоремы Муавра* - Лапласа следует, что случайная величина $\chi^2 = \frac{(m_1 - np_1)^2}{np_1} + \frac{(m_2 - np_2)^2}{np_2} + \dots + \frac{(m_k - np_k)^2}{np_k}$ при $n \rightarrow \infty$ асимптотиче-

ски распределена по закону χ^2 с $k-l-1$ степенью свободы. Здесь l - число предварительно оцениваемых параметров закона, на их количество понижается число степеней свободы статистики критерия. Таким образом,

введенная статистика $\chi^2 = \sum_{i=1}^k \frac{(m_i - np_i)^2}{np_i}$ при $n \rightarrow \infty$ независимо от ги-

потетических вероятностей p_i имеет χ^2 -распределение с $k-l-1$ степенью свободы. Следовательно, критерий χ^2 предписывает принять гипоте-

* Абрахам Муавр (1667-1754) – французский математик.

зу H_0 , если $\chi^2 < C$ (правосторонний критерий), и отвергнуть, если $\chi^2 \geq C$, где C - критическое значение критерия.

При практической реализации критерия χ^2 нужно следить за тем, чтобы объем выборки был велик, иначе неправомерно аппроксимация χ^2 -распределением распределения статистики критерия. Обычно считается, что достаточным условием этого является выполнение неравенств $m_i \geq 5$ при всех k , в противном случае маловероятные значения b_i объединяются в одно или присоединяются к другим значениям, причем объединенному значению приписывается суммарная вероятность.

В общем случае (непрерывные случайные величины) поступают следующим образом. Всю числовую прямую разбивают на k непересекающихся интервалов $(-\infty, d_1)$, $[d_1, d_2)$, $[d_2, d_3)$, ..., $[d_{k-1}, \infty)$. Затем определяют гипотетические вероятности $p_i = F_0(d_i) - F_0(d_{i-1})$ попадания в интервал $[d_{i-1}, d_i)$ и числа m_i элементов выборки, попавших в эти интервалы.

Затем вычисляют значение статистики $\chi^2 = \sum_{i=1}^k \frac{(m_i - np_i)^2}{np_i}$ и сравнивают его с критическим значением C , являющимся $(1 - \alpha)$ -процентным квантилем χ^2 -распределения. Как и в дискретном случае, маловероятные интервалы объединяются.

Разумеется, для того, чтобы увеличить качество критерия χ^2 (увеличить его мощность), необходимо уменьшить интервалы разбиения, однако этому препятствует ограничение на число попавших в каждый интервал наблюдений.

Пример. В следующей таблице приведен рост (см) 1004 девушек в возрасте 16 лет. Приняв 10% уровень значимости, проверить гипотезу H_0 , что они получены из нормально распределенной генеральной совокупности.

Границы интервалов	134-137	137-140	140-143	143-146	146-149	149-152	152-155
Частоты	1	4	16	53	121	197	229
Границы интервалов	155-158	158-161	161-164	164-167	167-170	170-173	
Частоты	186	121	53	17	5	1	

Решение

Применим критерий χ^2 -Пирсона для проверки нулевой гипотезы $H_0 : F(x) = \Phi(x)$. Поскольку распределение генеральной совокупности будет сравниваться со стандартным нормальным, выбранная статистика критерия будет центрирована и нормирована. Для этого необходимо знать математическое ожидание и дисперсию предполагаемого нормального закона, которые мы заменим их оценками, определенными по выборке. Сведем все данные в таблицу (см. следующую страницу).

Но- мер ин- тер- вала	Грани- цы интер- вала	Сере- дина ин- тер- вала x_i	Час- то- ты m_i	$x_i - m_X^*$	$(x_i - m_X^*)^2$	$z_i = \frac{ x_i - m_X^* }{\sigma^*}$	$f(z_i)$	$\frac{nd}{\sigma^*} f(z_i) = np_i$	np_i	$m_i - np_i$	$\frac{(m_i - np_i)^2}{np_i}$
1	134-137	135.5	1	-17.99	323.64	3.394	0.0013	0.739	4.89	0.11	0.003
2	137-140	138.5	4	-14.99	224.70	2.828	0.0073	4.149			
3	140-143	141.5	16	-11.99	143.76	2.262	0.0309	17.561	17.56	-1.56	0.139
4	143-146	144.5	53	-8.99	80.82	1.696	0.0947	53.818	53.82	-0.82	0.012
5	146-149	147.5	121	-5.99	35.88	1.130	0.2107	119.741	119.74	1.26	0.013
6	149-152	150.5	197	-2.99	8.94	0.564	0.3403	193.393	193.39	3.61	0.067
7	152-155	153.5	229	0.01	0.0	0.002	0.3989	226.696	226.70	2.30	0.023
8	155-158	156.5	186	3.01	9.06	0.568	0.3395	192.938	192.94	-6.94	0.250
9	158-161	159.5	121	6.01	36.12	1.134	0.2097	119.173	119.17	1.83	0.028
10	161-164	162.5	53	9.01	81.18	1.700	0.0940	53.420	53.42	-0.42	0.003
11	164-167	165.5	17	12.01	144.24	2.266	0.0306	17.390	17.39	-0.39	0.009
12	167-170	168.5	5	15.01	225.30	2.832	0.0072	4.092	4.77	1.23	0.317
13	170-173	171.5	1	18.01	324.36	3.398	0.0012	0.682			

$$m_X^* = 153.49, \bar{D}_X = 28.09, \sigma^* = 5.30$$

$$\chi_{\text{выб}}^2 = 0.864$$

В этой таблице первые четыре столбца – исходные данные задачи. Оценкой математического ожидания является выборочное среднее

$m_X^* = \frac{1}{n} \sum_{i=1}^n x_i$. Приведенная формула справедлива для обычной выборки.

Для группированной выборки, такой, какая приведена в задаче, эта формула принимает вид

$$m_X^* = \frac{1}{n} \sum_{i=1}^k m_i x_i, \quad (5.6.2)$$

где k - число первоначальных интервалов группировки ($k = 13$), m_i - наблюдаемые частоты, x_i - середины интервалов группировки. Аналогичная формула для несмещенной оценки дисперсии приобретает вид

$$\hat{D}_X = \frac{1}{n-1} \sum_{i=1}^k m_i (x_i - m_X^*)^2. \quad (5.6.3)$$

Рассчитанные с помощью этих формул по первым шести столбцам таблицы оценки математического ожидания и дисперсии предполагаемого нормального распределения выборки равны $m_X^* = 153.49$ см, $\hat{D}_X = 28.09$ кв. см, $\sigma^* = 5.30$ см.

В седьмом столбце приведены нормированные и центрированные значения x_i , в восьмом ординаты плотности $f(z_i) = \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2}$ стандартного нормального распределения $N(0,1)$, в девятом вычисляются значения $np_i = \frac{nd}{\sigma^*} f(z_i)$, где $d = 3$ - ширина интервала группировки, в десятом столбце значения np_i после объединения двух первых и двух последних интервалов. Наконец, два последних столбца служат для расчета выборочного значения критерия χ^2 -Пирсона $\chi_{\text{выб}}^2 = \sum_{i=1}^{k_1} \frac{(m_i - np_i)^2}{np_i}$, k_1 - число новых интервалов группировки. Так как по выборке определены оценки двух параметров, то $l = 2$ и число степеней свободы равно $k_1 - l - 1 = 11 - 2 - 1 = 8$. По таблице распределения χ^2 находим, что $\chi_{0.9,8}^2 = 13.4$. Так как $\chi_{\text{выб}}^2 = 0.864 < \chi_{0.9,8}^2$, то гипотеза H_0 о нормальном распределении группированных данных не противоречит результатам наблюдений и должна быть принята с уровнем значимости 0.1.

2. Критерий Колмогорова. В силу теоремы Гливленко–Кантелли эмпирическая функция распределения $F^*(x)$ представляет собой состоятельную оценку теоретической функции распределения $F(x)$. Поэтому можно сравнивать $F^*(x)$ с гипотетической $F_0(x)$, и, если мера расхождения между ними мала, считать справедливой гипотезу H_0 . Наиболее естественная и простая мера – это равномерное расстояние между $F^*(x)$ и $F_0(x)$ (рис. 5.21), т.е.

$$D = \sup_{-\infty < x < +\infty} |F^*(x) - F_0(x)|. \quad (5.6.4)$$

Очевидно, что D – случайная величина, поскольку ее значение зависит от случайного объекта $F^*(x)$. Если гипотеза H_0 справедлива и $n \rightarrow \infty$, то $F^*(x) \rightarrow F(x)$ при всяком x . Как всегда при проверке гипотезы, следует рассуждать так, как если бы гипотеза была верна. Ясно, что H_0 должна быть отвергнута, если полученное в эксперименте значение статистики D окажется неоправданно большим. Замечательное свойство D состоит в том, что если гипотетическое распределение указано правильно, то закон распределения статистики D оказывается одним и тем же для всех непрерывных истинных функций распределения.

При малых n для статистики D при гипотезе H_0 составлены таблицы процентных точек. При больших n распределение D (при гипотезе H_0) указывает найденная в 1933 г. А.Н. Колмогоровым предельная теорема (см. подразд. 2.4). Она говорит о статистике $D_n = \sqrt{n}D$ (поскольку сама величина $D \rightarrow 0$ при H_0 , приходится умножать ее на неограниченно растущую величину, чтобы распределение стабилизировалось).

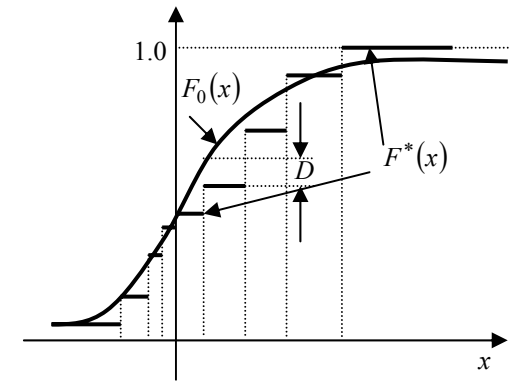


Рис. 5.21. Графики теоретической $F_0(x)$ и эмпирической $F^*(x)$ функций распределения

Рассмотрим статистику $D_n = D_n(x_1, x_2, \dots, x_n) = \sqrt{n} \sup_{-\infty < x < +\infty} |F^*(x) - F_0(x)|$.

Критерий Колмогорова предписывает принять гипотезу H_0 , если $D_n < C$ и отвергнуть ее в противном случае, где C - критическое значение критерия. При $n \rightarrow \infty$ критическое значение C совпадает с $(1 - \alpha)\%$ квантилью распределения Колмогорова.

При практической реализации критерия сначала по выборке x_1, x_2, \dots, x_n составляют вариационный ряд $x_1^*, x_2^*, \dots, x_n^*$. Затем находят значение статистики D_n . Для этого можно использовать несколько формул. Например,

$$D_n = \sqrt{n} \max_{1 \leq i \leq n} \left[\frac{i}{n} - F_0(x_i^*) F_0(x_i^*) - \frac{i-1}{n} \right]. \quad (5.6.5)$$

Другая употребительная формула имеет вид

$$D_n = \sqrt{n} \left[\max_{1 \leq i \leq n} \left| F_0(x_i^*) - \frac{2i-1}{2n} \right| + \frac{1}{2n} \right]. \quad (5.6.6)$$

После этого сравнивают полученное значение D_n с критическим значением C для заданного уровня значимости α и принимают или отвергают гипотезу H_0 .

Пример. Дано следующее распределение успеваемости 100 студентов-заочников, сдавших четыре экзамена:

Число сданных экзаменов	0	1	2	3	4
Число студентов	1	1	3	35	60

Проверить по критериям χ^2 -Пирсона и Колмогорова гипотезу о том, что число сданных экзаменов распределено биномиально. Принять $\alpha = 0.1$.

Решение

Здесь случайной величиной является число сданных экзаменов среди четырех. Обозначим ее X и установим сначала закон распределения этой величины. Для установления закона необходимо сделать некоторые допущения.

1. Процесс сдачи четырех экзаменов представим как четыре испытания. Будем считать эти испытания независимыми, т.е. пусть вероятность сдачи любым студентом любого экзамена не зависит от того, будет сдано или нет любое количество других экзаменов.

2. Вероятность сдачи студентом любого отдельно взятого экзамена одна и та же и равна p , а вероятность не сдачи равна $q = 1 - p$.

Если принять эти допущения, то перед нами схема Бернулли и число сданных экзаменов среди четырех сдаваемых будет иметь биномиальный закон распределения, т.е.

$$P(X = x) = C_4^x p^x q^{4-x}, \quad x = 0, 1, 2, 3, 4. \quad (5.6.7)$$

Для оценки вероятности p^* воспользуемся методом максимального правдоподобия. Получим

$$\begin{aligned} L(x_1, x_2, \dots, x_k) &= P(X = x_1) \times \\ &\times P(X = x_2) \cdot \dots \cdot P(X = x_k) = C_k^{x_1} p^{x_1} q^{k-x_1} \cdot C_k^{x_2} p^{x_2} q^{k-x_2} \cdot \dots \cdot C_k^{x_k} p^{x_k} q^{k-x_k} = \\ &= \left(C_k^{x_1} + C_k^{x_2} + \dots + C_k^{x_k} \right) p^{x_1+x_2+\dots+x_k} q^{k^2-(x_1+x_2+\dots+x_k)} = \\ &= \left(C_k^{x_1} + C_k^{x_2} + \dots + C_k^{x_k} \right) p^{\sum_{i=1}^k x_i} q^{k^2 - \sum_{i=1}^k x_i}. \end{aligned}$$

Найдем логарифм функции правдоподобия

$$\ln L = \ln \left(2^k p^{\sum_{i=1}^k x_i} q^{k^2 - \sum_{i=1}^k x_i} \right) = \left\langle \frac{\text{так как}}{\sum_{i=1}^n C_n^i = 2^n} \right\rangle = k \ln 2 + \sum_{i=1}^k x_i \ln p + \left(k^2 - \sum_{i=1}^k x_i \right) \ln q.$$

Тогда $\frac{\partial \ln L}{\partial p} = \frac{1}{p} \sum_{i=1}^k x_i - \frac{1}{q} \left(k^2 - \sum_{i=1}^k x_i \right) = 0$. Отсюда $\hat{p} = p^* = \sum_{i=1}^k x_i / k^2$.

В рассматриваемой схеме практической случайной величиной является число экзаменов, сданных всеми 100 студентами, и x_i наблюдается m раз, т.е. в задаче задана сгруппированная выборка. Тогда

$$p^* = \frac{\sum_{i=1}^k x_i}{k^2} = \frac{\sum_{i=1}^l x_i m_i}{sn}, \quad \text{где } k^2 = sn - \text{число экзаменов, сдаваемых всеми}$$

100 студентами, $s = 4$ - число сдаваемых экзаменов, $n = 100$ - число студентов, $l = 5$ - число разрядов сгруппированной выборки. Тогда

$$p^* = \frac{\sum_{i=1}^l x_i m_i}{sn} = \frac{0 \cdot 1 + 1 \cdot 1 + 2 \cdot 3 + 3 \cdot 35 + 4 \cdot 40}{4 \cdot 100} = 0.88. \quad \text{Вычислим теперь}$$

теоретические вероятности по формуле $P(X = x) = C_4^x \cdot 0.88^x \cdot 0.12^{4-x}$,

$x = 0, 1, 2, 3, 4$ и относительные частоты $p_i^* = m_i/n$ и поместим их в следующую таблицу:

X	0	1	2	3	4
p	0.00021	0.00608	0.06691	0.32711	0.59969
p^*	0.01	0.01	0.03	0.35	0.60

Проверим теперь сначала с помощью критерия χ^2 -Пирсона гипотезу о соответствии теоретического биномиального распределения фактическим данным исходной таблицы. Итак, $H_0: F(x) = B(p)$, $p = 0.88$. Составим таблицу такой же структуры, как в предыдущем примере.

Но- мер интер- вала	Число сдан- ных экз. x_i	Число студен- тов m_i (частоты)	Относи- тельные частоты p_i^*	Теоре- тиче- ские частоты p_i	np_i	$(m_i - np_i)^2$	$\frac{(m_i - np_i)^2}{np_i}$
1	0	1	0.01	0.00021	0.021		
2	1	1	0.01	0.00608	0.608	5.382	0.735
3	2	3	0.03	0.06691	6.691		
4	3	35	0.35	0.32711	32.711	5.239	0.160
5	4	60	0.60	0.59969	59.969	0.001	0.000

$$\chi_{\text{выб}}^2 = 0.895$$

Число степеней свободы статистики критерия χ^2 -Пирсона равно

$3 - 1 - 1 = 1$. Тогда критиче-

ское значение критерия

$C = \chi_{0.9,1}^2 = 2.71$. Так как

$\chi_{\text{выб}}^2 = 0.895 < \chi_{0.9,1}^2 = 2.71$,

то гипотеза H_0 принимается

с уровнем значимости

$\alpha = 0.1$.

Прделаем то же самое с

помощью критерия Колмо-

рова. Для этого построим

функцию распределения

$F_0(x)$ (рис. 5.22).

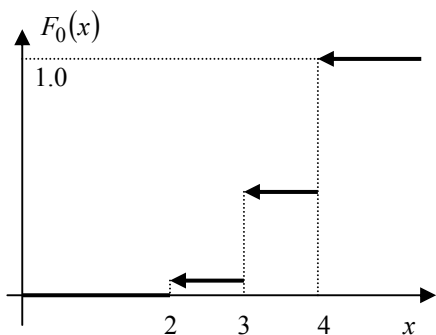


Рис. 5.22. График функции $F_0(x)$

X	0-2	3	4
p	0.0732	0.3271	0.5997

Так как $F_0(x) = \sum_{x_i < x} p_i$, то гипотетическая функция распределения

$$\text{будет равна } F_0(x) = \begin{cases} 0, & x \leq 2, \\ 0.0732, & 2 < x \leq 3, \\ 0.4003, & 3 < x \leq 4, \\ 1, & x > 4. \end{cases}$$

Рассчитаем теперь значение статистики D_n сначала по формуле (5.6.5)

$$D_n = \sqrt{n} \max_{1 \leq i \leq n} \left[\frac{i}{n} - F_0(x_i^*) \right] F_0(x_i^*) - \frac{i-1}{n} \Bigg].$$

$$i = 1: \max \left[\frac{1}{3} - 0, 0 - \frac{1-1}{3} \right] = 0.3333,$$

$$i = 2: \max \left[\frac{2}{3} - 0.0732, 0.0732 - \frac{1}{3} \right] = 0.5935,$$

$$i = 3: \max \left[1 - 0.4003, 0.4003 - \frac{2}{3} \right] = 0.5997.$$

$$\text{Отсюда } D_n = \sqrt{3} \max(0.3333, 0.5935, 0.5997) = 1.0387.$$

При вычислении значения статистики по формуле (5.6.6) получим те

$$\text{же значения. Действительно, } D_n = \sqrt{n} \left[\max_{1 \leq i \leq n} \left| F_0(x_i^*) - \frac{2i-1}{2n} \right| + \frac{1}{2n} \right],$$

$$i = 1: \left| 0 - \frac{2-1}{6} \right| + \frac{1}{6} = 0.3333, \quad i = 2: \left| 0.0732 - \frac{3}{6} \right| + \frac{1}{6} = 0.5935,$$

$$i = 3: \left| 0.4003 - \frac{5}{6} \right| + \frac{1}{6} = 0.5997.$$

$D_n = \sqrt{3} \max(0.3333, 0.5935, 0.5997) = 1.0387$. Найдём критическое значение C критерия Колмогорова. Это 90% квантиль этого распределения, т.е. $C = K^{-1}(0.9) = 1.23$. Так как $D_n < C$, следовательно, $D \in W \setminus \omega$ и гипотеза H_0 должна быть принята с уровнем значимости $\alpha = 0.1$.

5.7. Лабораторная работа № 6. Критерии согласия в статистическом пакете STATGRAPHICS

Критериями согласия называют статистические критерии, предназначенные для проверки согласия опытных данных и теоретической модели. К сожалению, все изложенные в подразд. 5.6 методы без ограничения могут быть применены только к простым гипотезам.

Более трудной, но и более важной для приложений является проверка гипотезы о том, что данная выборка подчиняется определенному параметрическому закону распределения. Параметры этого закона остаются неопределенными, так что гипотеза сложная.

Существуют статистики для проверки таких гипотез, являющиеся функциями неизвестных параметров распределений. Используются и модификации известных нам статистик, например, статистики χ^2 - Пирсона и D_n Колмогорова, их свойства во многом повторяют отмеченные ранее свойства аналогичных статистик для простых гипотез, однако, распределения все же иные. В целом, при справедливости исходной гипотезы модифицированные статистики для сложных гипотез принимают существенно меньшие значения, чем соответствующие статистики для простых. Это приводит к тому, что уровень значимости статистик для сложной гипотезы всегда меньше уровня значимости этих статистик для простой гипотезы. Таким образом, если полученный уровень значимости для простой гипотезы мал, то уровень значимости для сложной гипотезы еще меньше и эту гипотезу следует отвергать.

В пакете STATGRAPHICS процедуры тестов согласия χ^2 и Колмогорова находятся в разделе Describe→Numeric Data→Distribution Fitting (Подбор распределений), причем данные по обоим тестам выводятся одновременно. Встроенных распределений, т.е. таких, функция распределения которых может служить теоретической гипотетической функцией $F_0(x)$, всего пять: экспоненциальное, экстремальных значений, логнормальное, нормальное и Вейбулла.

Опишем этапы и последовательность действий при использовании тестов согласия в пакете STATGRAPHICS. Итак, после выбора пункта основного меню Describe→Numeric Data→Distribution Fitting появляется подменю подбора распределений, в котором надо задать имя файла, содержащего исследуемую выборку. Рассмотрим в качестве исходной нормальную выборку объемом 50 единиц с параметрами $m_X = 6$, $D_X = 10$ под именем NORM из предыдущей лабораторной работы (если ее нет, необходимо смоделировать ее средствами пакета). После набора в графе

DATA подменю подбора распределений имени NORM и щелчка по кнопке OK появляется заставка Analysis Summary (Сводка анализа) с некоторыми общими сведениями о выборке. Здесь же указываются вычисленные выборочные оценки математического ожидания и дисперсии.

Щелчком правой кнопкой мыши в любом месте этой заставки и выберем пункт Analysis

Options. Появится меню выбора гипотетического распределения (рис. 5.23), в котором нормальное распределение помечено по умолчанию. Мы будем проверять гипотезу на соответствие распределения выборки нормальному распределению, поэтому оставим точку в поле Normal и щелкнем по кнопке OK. Числовые

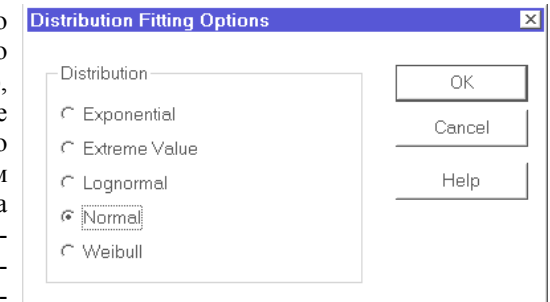


Рис. 5.23. Окно диалога для выбора гипотетического распределения

данные при этом в поле заставки Analysis Summary не изменятся. Они изменились бы, если бы мы выбрали другое гипотетическое распределение.

Вместе с заставкой Analysis Summary появилось дополнительное меню с пунктами Input Dialog, Tabular Options, Graphics Options и Save Results. Выберем пункт Tabular Options и отметим все процедуры этого меню, щелкнув по кнопке All (Все). Укажем назначение всех входящих в это меню процедур. Информация о результатах работы этих процедур по выборке NORM приведена на рис. 5.24, 5.25.

Analysis Summary (Сводка анализа) указывает объем выборки, ее экстремальные значения и оценки математического ожидания и дисперсии.

Test for Normality (Тест на нормальность) содержит вычисленную по критерию согласия χ^2 -Пирсона статистику проверки нулевой гипотезы

$H_0 : F(x) = \Phi\left(\frac{x - m_X}{\sigma_X}\right)$ и ее уровень значимости. Далее следует статистика

теста Шапиро – Уилкса и данные по асимметрии и эксцессу. Эти данные не имеют отношения к рассмотренной нами теории.

Goodness-of-Fit-Tests (Критерии согласия) приводят данные по критериям χ^2 -Пирсона и Колмогорова. В двух первых столбцах таблицы результатов Lower Limit и Upper Limit указаны нижние и верхние границы

Data variable: NORM

50 values ranging from -1,80021 to 14,4547

Fitted normal distribution:
mean = 6,23962
standard deviation = 3,19238

Tests for Normality for NORM

Computed Chi-Square goodness-of-fit statistic = 13,36
P-Value = 0,574513

Shapiro-Wilks W statistic = 0,991189
P-Value = 0,988889

Z score for skewness = 0,240329
P-Value = 0,810071

Z score for kurtosis = 0,358743
P-Value = 0,719784

Goodness-of-Fit Tests for NORM

Chi-Square Test					
	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chisquare
at or below		2,71429	8	6,74	0,24
	2,71429	5,57143	11	14,12	0,69
	5,57143	8,42857	19	16,82	0,28
above	8,42857		12	12,32	0,01

Chi-Square = 1,21618 with 1 d.f. P-Value = 0,270109

Estimated Kolmogorov statistic DPLUS = 0,0456355
Estimated Kolmogorov statistic DMINUS = 0,0786373
Estimated overall statistic DN = 0,0786373
Approximate P-Value = 0,916612

Tail Areas for NORM

area below 4,99169 = 0,347931

area below 5,61565 = 0,422517

area below 6,23962 = 0,500003

area below 6,86358 = 0,577483

area below 7,48754 = 0,652069

Рис. 5.24. Результаты проверки теста на нормальность выборки и критериев согласия

χ^2 -Пирсона и Колмогорова

интервалов группировки. В столбце Observed Frequency представлены наблюдаемые частоты, а в столбце Expected Frequency – частоты подбранного гипотетического распределения. Столбец Chisquare содержит значения слагаемых формулы (5.6.1) для каждого интервала группировки. Нижняя строка включает значение статистики χ^2 , число степеней свободы d.f. (Degree of Freedom) и уровень значимости p-Value.

Число интервалов, нижнюю и верхнюю границы группировки можно задать в специальном подменю Frequency Tabulation Options (Установки таблицы частот) (рис. 5.26), вызываемым щелчком правой кнопки мыши в

Critical Values for NORM

area below	-1,18697	= 0,01
area below	2,14841	= 0,1
area below	6,23962	= 0,5
area below	10,3308	= 0,9
area below	13,6662	= 0,99

Рис. 5.25. Значения квантилей и (или) критических точек распределения

Рис. 5.26. Диалоговое окно установки таблицы частот

поле Goodness-of-Fit-Tests и выбором Pane Options во вспомогательном меню. Следует отметить, что число интервалов группировки, указанное пользователем, корректируется с учетом обеспечения условий применимости аппроксимации распределения статистики с помощью распределения χ^2 . Кроме того, для вычисления частот гипотетического распределения используются оценки матожидания и дисперсии по выборке. Это приводит к тому, что истинный уровень значимости для сложной гипотезы несколько больше, чем вычисленное значение p-Value. Приближенный уровень значимости вычисленной статистики лежит между квантилями χ^2 -распределения с $k-3$ и $k-1$ степенями свободы, где k - число интервалов группировки.

Далее представлены результаты расчетов по критерию Колмогорова. Они включают значения статистик Колмогорова D^+ (Estimated Kolmogorov Statistic DPLUS) и D^- (Estimated Kolmogorov Statistic DMINUS), а также D_n (Estimated Kolmogorov Statistic DN) и минимальный уровень значимости последней статистики в случае простой гипотезы (Approximate p-Value).

Следует правильно интерпретировать большие численные значения уровней значимости в этих тестах. В критериях согласия используется правосторонний критерий значимости.

При этом заданному уровню значимости $\alpha_{\text{зад}}$ соответствует критическая точка $\chi^2_{\text{зад}}$, являющаяся границей критической области ω и области принятия решений $W \setminus \omega$. В пакете STATGRAPHICS в тестах согласия

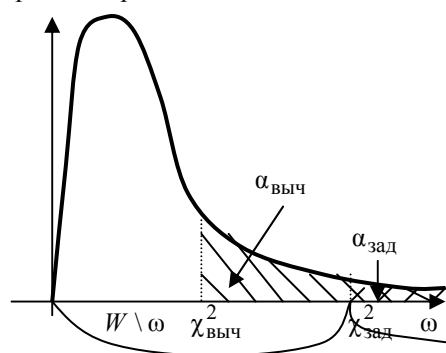


Рис. 5.27. Границы критических областей, определяемые по заданному и вычисленному значениям уровня значимости

решается «обратная» задача: по вычисленному значению статистики критерия $\chi^2_{\text{выч}}$ как критической точке находится соответствующая вероятность (p-Value) события $P(\chi^2 > \chi^2_{\text{выч}})$ (рис. 5.27). Ясно, что если статистика критерия попадает в область принятия решения $\chi^2_{\text{выч}} \in W \setminus \omega$ и гипотезу H_0 надо принять, значение p-Value всегда больше изначально заданного уровня $\alpha_{\text{зад}}$, которое обычно мало (0.1, 0.05, 0.01 и

тому подобное).

Tail Areas (Площади хвостов) содержат значения функции распределения в пяти точках. Эти значения заполнены по умолчанию, но их можно изменить, вызвав щелчком правой кнопки мыши в поле заставки Tail Areas дополнительное меню и выбрав в нем пункт Pane Options. Появится еще одно подменю Tail Areas Options (рис. 5.28). Введем в соответствующие поля этого подменю значения 3, 6, 9, 11 и 13.5. После щелчка по кнопке ОК информация на заставке Tail Areas for NORM сменится на приведенную на рис. 5.29.

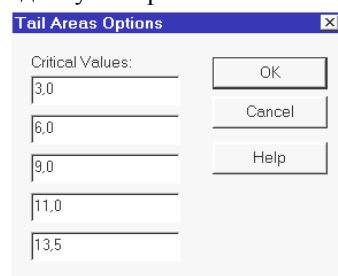


Рис. 5.28. Окно диалога для задания значений квантилей

Tail Areas for NORM

```
area below 3,0 = 0,1551
area below 6,0 = 0,470081
area below 9,0 = 0,806394
area below 11,0 = 0,932042
area below 13,5 = 0,988526
```

Рис. 5.29. Значения квантилей

Critical Values (Критические значения). В этой заставке вычисляются по заданному значению функции распределения (вероятности) p квантили t_p этого распределения.

Эта операция является обратной по отношению к предыдущей процедуре Tail Areas. Необходимые значения вероятностей можно задать вызвав совершенно аналогичным образом подменю Critical Values Options (рис. 5.30). Зададим значения 0.01, 0.1, 0.5, 0.9 и 0.99. Этим вероятностям будут соответствовать квантили, показанные на рис. 5.25.

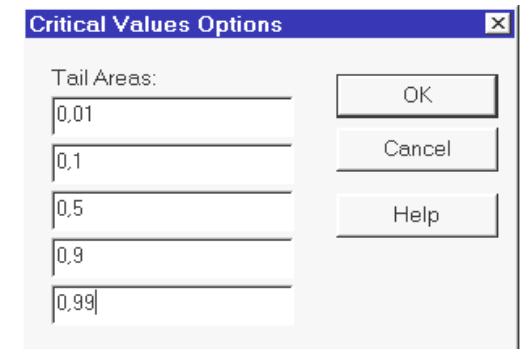


Рис. 5.30. Окно диалога для задания уровней квантилей

Наряду с чисто числовой информацией можно вывести на экран дисплея несколько графиков. Выберем пункт дополнительного меню Graphics Options (рис. 5.31). Для нормальной выборки можно построить шесть графиков, кроме графика распределения Вейбулла, ибо выборка этого распределения не должна содержать отрицательных величин.

Density Trace (График эмпирической функции плотности) строит этот график по данным исходной нормальной выборки NORM (рис. 5.32). Даже на глаз видно, что график имеет отрицательную асимметрию, т.е. более тяжелый левый «хвост» распределения. Действительно, выборочный коэффициент асимметрии этой выборки равен -0.109 .

Symmetry Plot (Симметричный график) содержит точки, являющиеся результатом сглаживания выборки скользящей медианой.

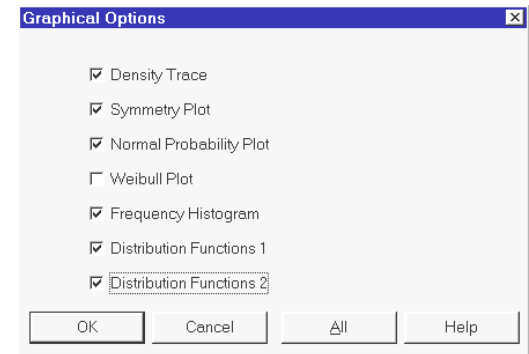


Рис. 5.31. Панель графических параметров при проверке гипотез о распределениях

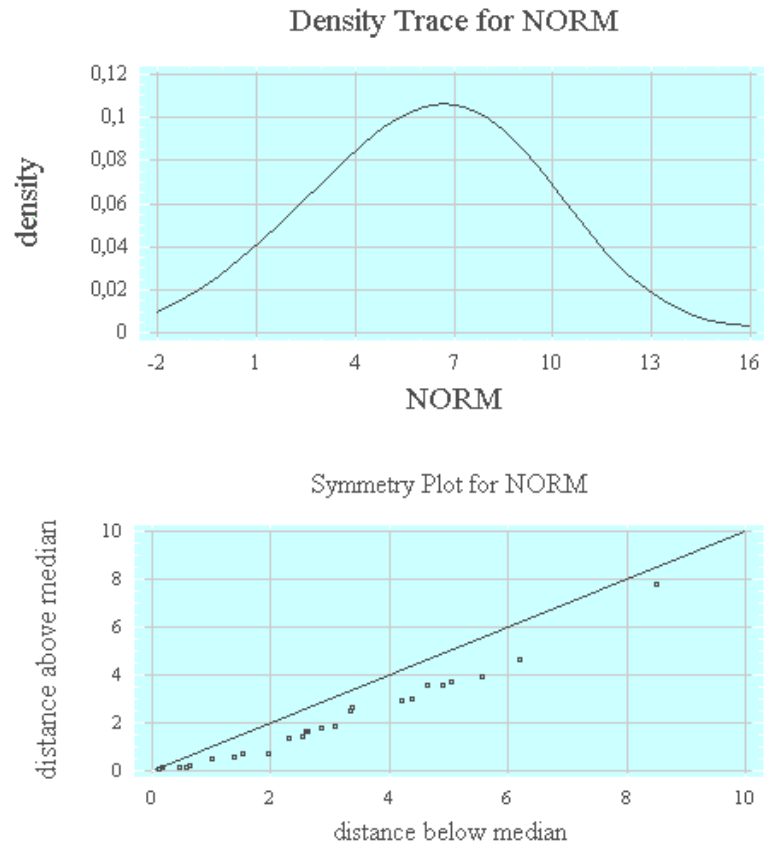


Рис. 5.32. Графики эмпирической функции плотности и скользящей медианы

Normal Probability Plot (График на нормальной вероятностной бумаге) строит график эмпирической функции распределения на нормальной вероятностной бумаге (рис. 5.33). Если $x_i \in N(m_X, D_X)$, то $F(x) = \Phi(x - m_X / \sigma_X)$. Применим к этой зависимости функцию Φ^{-1} и введем переменную $z = \Phi^{-1}(F(x))$. Тогда зависимость превращается в линейную $z = x - m_X / \sigma_X$. Эмпирическая функция распределения $F^*(x)$ в каждой точке вариационного ряда совершает скачок и имеет разрыв первого рода.

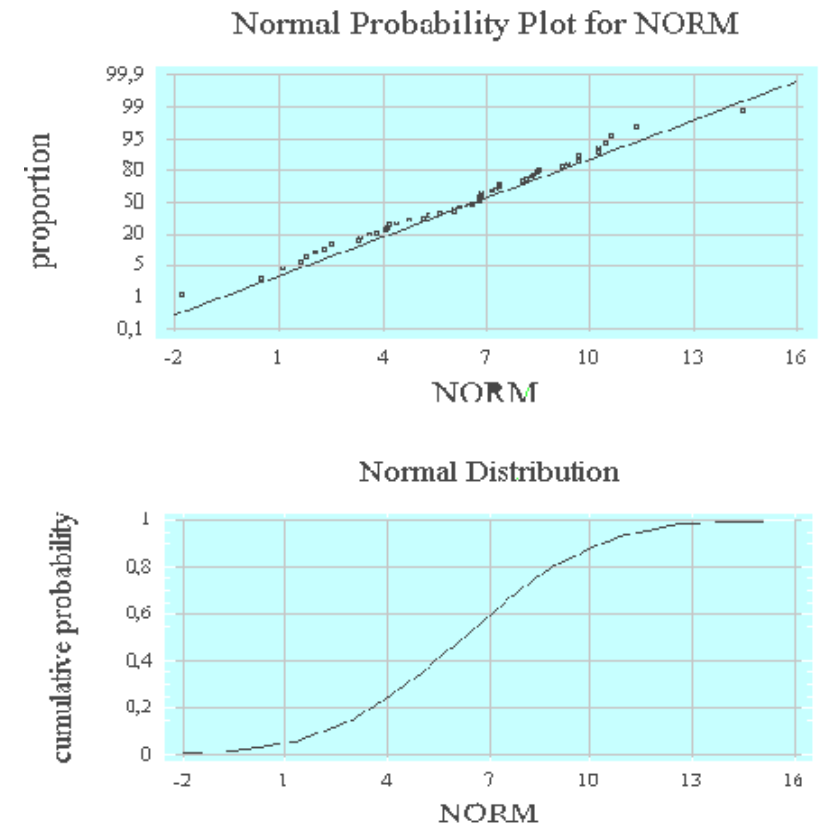


Рис. 5.33. Графики эмпирической функции распределения на нормальной вероятностной бумаге и кумулятивной кривой

Для проверки нормальности выборки применим функцию Φ^{-1} к серединам этих скачков, в результате получим точки $\left(x_i, \Phi^{-1}\left(\frac{2i-1}{2n}\right)\right)$ в плоскости (x, z) . В зависимости от того, насколько хорошо эти точки ложатся на прямую линию, можно судить о нормальности распределения. Это глазомерный метод проверки нормальности. Даже небольшой опыт с реальными выборками позволяет достаточно уверенно выделять среди них отклоняющиеся от нормальных.

Frequency Histogram (Гистограмма частот) в графическом виде представляет таблицу частот выборки после группирования данных на заданном числе интервалов (рис. 5.34).

Distribution Function 1 (Функция плотности распределения) выводит график функции плотности исходного нормального распределения.

Distribution Function 2 (Функция распределения) дает график функции распределения. Все эти шесть графиков приведены здесь для иллюстрации.

Задание 1. Выберите из табл. 3 вид гипотетического распределения и его параметры, смоделируйте соответствующую выборку в пакете STATGRAPHICS и проверьте с помощью критериев согласия пакета соответствие статистического и гипотетического распределений с уровнем значимости $\alpha = 0.1$.

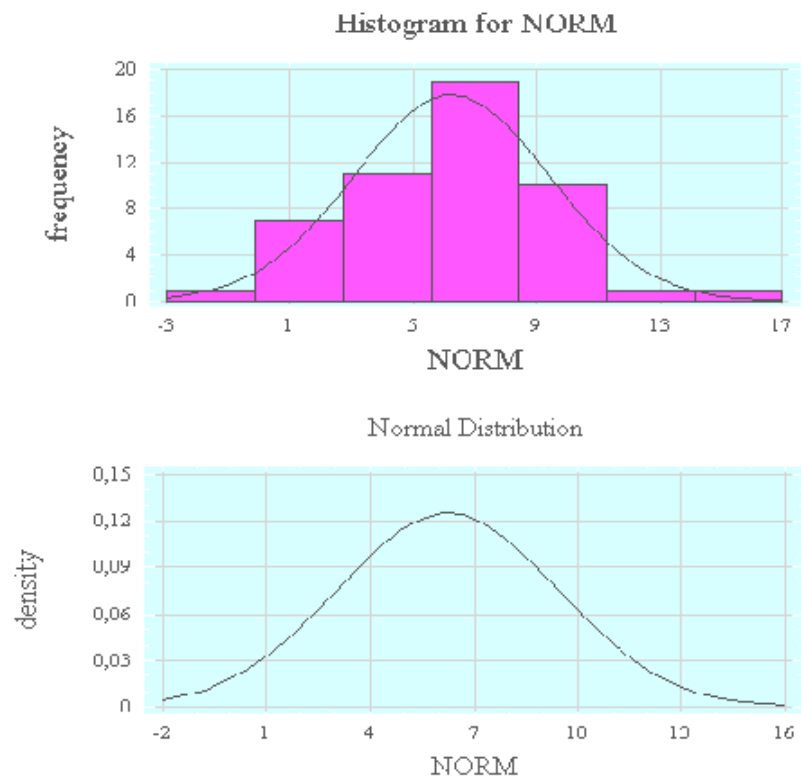


Рис. 5.34. Гистограмма частот и график функции плотности вероятности

Т а б л и ц а 3

Номер фамилии в журнале преподавателя	Вид распределения	Параметры распределения	Объем выборки
1-8	Экспоненциальное	$\lambda = (0.5 \cdot N) \bmod 3 + 1$, где N - порядковый номер дня вашего рождения	100
9-16	Логнормальное	m - порядковый номер месяца рождения, σ - номер фамилии в журнале преподавателя	50
17-24	Нормальное	m - порядковый номер месяца рождения, σ - номер фамилии в журнале преподавателя	50
25-32	Классическое Вейбулла	$\alpha = 2$, $\lambda = (0.5 \cdot N) \bmod 3 + 1$	100

5.8. Лабораторная работа №7. Критерии согласия в математическом пакете MATHCAD

Как и в лабораторной работе №5 в пакете MATHCAD вычисление всех статистик тестов χ^2 -Пирсона и Колмогорова придется программировать. Исходная выборка может быть задана всеми наблюдениями или в виде сгруппированных данных. Для получения конечного результата при использовании этих двух тестов, очевидно, необходимо составить следующие подпрограммы: получения вариационного ряда по исходной выборке, вычисления сгруппированной выборки, исправления разрядов сгруппированной выборки по условию $m_i > 5$, вычисления статистик χ^2 -Пирсона и D_n Колмогорова, наконец, принятия решения о нулевой гипотезе H_0 . Ниже приводятся тексты этих подпрограмм с необходимыми комментариями.

Подпрограмма **str** упорядочивает исходную несгруппированную выборку по возрастанию ее элементов. Используются две встроенные функции

<pre> str(x) := n ← rows(x) l ← cols(x) return x if l=2 for i ∈ 1..n-1 for j ∈ i+1..n a ← x_i if x_j < x_i x_i ← x_j x_j ← a x </pre>	<p>ции пакета MATHCAD rows и cols, которые подсчитывают количество строк и столбцов в матрице-аргументе.</p> <p>Подпрограмма grupvib получает сгруппированную выборку по исходной. Ее параметры: x - вектор исходной выборки, l - первоначальное число разрядов группировки. Выходные параметры: $x1$ - вектор вариационного ряда, первый столбец матрицы $x2$ содержит значения левых концов интервалов группировки, второй столбец – значения правых концов, вектор m содержит частоты</p>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

попадания элементов выборки в образованные интервалы.

Подпрограмма **interval** исправляет сгруппированную выборку, объединяя крайние интервалы, у которых $m_i \leq 5$, в один. Ее параметры: матрица $x_{i,j}$, содержащая первоначальную сгруппированную выборку, и вектор частот m . В результате работы подпрограммы в матрице $x1$ находятся исправленные границы интервалов, вектор $m1$ содержит исправленные частоты, а переменная mnw равна размерности вектора m . В подпрограмме использована встроенная функция пакета MATHCAD **floor**, вычисляющая наибольшее целое, не превосходящее аргумент.

Подпрограмма χ^2 вычисляет статистику критерия согласия χ^2 -Пирсона. Теоретические вероятности считаются по формуле $P(\alpha \leq X < \beta) = F(\beta) - F(\alpha)$, где F - функция распределения генеральной совокупности, откуда получена выборка. В приведенной программе $F \equiv \text{pnorm}$ - функции распределения нормального закона. Кроме того, в теле программы оцениваются два параметра нормального закона m_X и D_X .

При необходимости исследовать выборку, подчиняющуюся другому закону распределения, эти операторы необходимо заменить на операторы,

вычисляющие нужную функцию распределения и требуемые ею неизвестные параметры этого распределения. Подпрограмма χ^2 требует задания исходной выборки в сгруппированной или несгруппированной форме, вектора частот и первоначального числа интервалов группировки. Выходными параметрами подпрограммы являются числовое значение статистики χ^2 -Пирсона и число интервалов исправленной сгруппированной выборки.

$\text{interval}(x, m) :=$ $\begin{aligned} & n \leftarrow \text{rows}(x) \\ & \text{nnow} \leftarrow n \\ & k \leftarrow \frac{n}{2} \\ & k \leftarrow \text{floor}(k) \\ & \text{for } i \in 1..k \\ & \quad \left \begin{array}{l} j \leftarrow n - i + 1 \\ \text{if } m_i \leq 5 \\ \quad \left \begin{array}{l} x_{i+1,1} \leftarrow x_{i,1} \\ m_{i+1} \leftarrow m_{i+1} + m_i \\ m_i \leftarrow 0 \end{array} \right. \\ \text{break if } i=j \\ \text{if } m_j \leq 5 \\ \quad \left \begin{array}{l} x_{j-1,2} \leftarrow x_{j,2} \\ m_{j-1} \leftarrow m_{j-1} + m_j \\ m_j \leftarrow 0 \end{array} \right. \end{array} \right. \\ & k \leftarrow 0 \\ & \text{for } i \in 1..n \\ & \quad \text{if } m_i > 10^{-9} \\ & \quad \quad \left \begin{array}{l} k \leftarrow k + 1 \\ x1_{k,1} \leftarrow x_{i,1} \\ x1_{k,2} \leftarrow x_{i,2} \\ m1_k \leftarrow m_i \end{array} \right. \\ & \text{nnow} \leftarrow k \\ & \left[\begin{array}{c} x1 \\ m1 \\ \text{nnow} \end{array} \right] \end{aligned}$	$\text{grupvib}(x, l) :=$ $\begin{aligned} & n \leftarrow \text{rows}(x) \\ & x1 \leftarrow \text{str}(x) \\ & d \leftarrow \frac{(x1_n - x1_1)}{1} \\ & a \leftarrow x1_1 - 10^{-6} \\ & \text{for } i \in 1..1 \\ & \quad \left \begin{array}{l} x2_{i,1} \leftarrow a \\ x2_{i,2} \leftarrow x2_{i,1} + d - 10^{-5} \\ a \leftarrow x2_{i,2} + 10^{-5} \\ x2_{1,2} \leftarrow a + 10^{-6} \text{ if } i=1 \\ k \leftarrow 0 \\ \text{for } j \in 1..n \\ \quad k \leftarrow \text{if}(x1_j > x2_{i,2}, k, \text{if}(x1_j < x2_{i,1}, k, k+1)) \\ m_i \leftarrow k \end{array} \right. \\ & \left[\begin{array}{c} x1 \\ x2 \\ m \end{array} \right] \end{aligned}$
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------


```

Pirson(x, m, l, α) :=
  |χobs←χ2(x, m, l)1
  |nnow←χ2(x, m, l)2
  |k←nnow−3
  |return "не хватает данных для критерия хи-квадрат Пирсона" if k<1
  |χкр←qchisq(1−α, k)
  |return "гипотеза Н0 принимается с заданным уровнем значимости" if χobs<χкр
  |return "гипотеза Н0 отвергается с заданным уровнем значимости" if χobs≥χкр

```

Подпрограмма **Pirson** принимает решение о нулевой гипотезе. Она очень проста, все ее операторы совершенно понятны и не требуют ком-

<pre> χ2(x, m, l) := k←cols(x) if k=1 x2←grupvib(x, l)₂ m←grupvib(x, l)₃ x2←x if k=2 x1←interval(x2, m)₁ m1←interval(x2, m)₂ nnow←interval(x2, m)₃ k←0 for i∈1..nnow k←k+m1_i x2_{i,1}←$\frac{(x1_{i,2}+x1_{i,1})}{2}$ mx←$\frac{1}{k} \left[\sum_{i=1}^{nnow} (x2_{i,1} \cdot m1_i) \right]$ Dx←$\frac{1}{k} \cdot \sum_{i=1}^{nnow} (x2_{i,1} - mx)^2 \cdot m1_i$ σx←\sqrt{Dx} for i∈1..nnow p1←pnorm(x1_{i,2}, mx, σx)−pnorm(x1_{i,1}, mx, σx) p1←k·p1 χ2←$\sum_{i=1}^{nnow} \frac{(m1_i - p1)^2}{p1}$ $\begin{bmatrix} \chi^2 \\ nnow \end{bmatrix}$ </pre>	<pre> Dn(x, m) := n←rows(x) k←cols(x) if k=1 x1←str(x) mx←mean(x) Dx←var(x) σx←\sqrt{Dx} if k=2 l←0 for i∈1..n l←l+m1_i x1_i←$\frac{(x1_{i,2}+x1_{i,1})}{2}$ mx←$\frac{1}{l} \cdot \sum_{i=1}^n x1_i \cdot m1_i$ Dx←$\frac{1}{l} \cdot \sum_{i=1}^n (x1_i - mx)^2 \cdot m1_i$ σx←\sqrt{Dx} for i∈1..n p1←pnorm(x1_i, mx, σx) x1_i←$\frac{i}{n} - p1$ p1←p1−$\frac{i-1}{n}$ x2←stack(x1, p) Dn←max(x2)·\sqrt{n} Dn </pre>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

ментариев. Встроенная функция **qchisq** вычисляет $(1 - \alpha)\%$ квантили χ^2 - распределения. Ее входные параметры аналогичны параметрам подпрограммы χ^2 , а α - уровень значимости нулевой гипотезы.

Подпрограмма **Dn** вычисляет статистику критерия согласия Колмогорова. Здесь, так же как и в подпрограмме χ^2 для вычисления значений гипотетической функции распределения, использована функция распределения нормального закона **pnorm**. При необходимости ее следует поменять вместе с операторами, оценивающими параметры нормального закона. Кроме того, в теле подпрограммы использованы следующие встроенные функции: **mean** и **var** вычисляют оценки математического ожидания и дисперсии, функция **stack** формирует одну матрицу из двух, располагая первую матрицу над второй, наконец, функция **max** находит наибольший элемент в матрице-аргументе.

Подпрограмма **Kolm**, так же как подпрограмма **Pirson**, принимает решение о принятии или отвержении нулевой гипотезы H_0 с уровнем значимости α . $(1 - \alpha)\%$ квантили распределения Колмогорова вычисляются линейным интерполированием с помощью встроенной функции **lin-terp**.

```

Kolm(x,m,α) :=
  Dobs ← Dn(x,m)
  for i ∈ 1..20
    | argi ← 0.1·i
    | 
$$x1_i ← 1 + 2 \cdot \sum_{j=1}^{100} (-1)^j \cdot \exp\left[-2j^2 \cdot (\arg_i)^2\right]$$

    | Dkp ← lininterp(x1, arg, 1 - α)
  return "гипотеза H0 принимается с заданным уровнем значимости" if Dobs < Dkp
  return "гипотеза H0 отвергается с заданным уровнем значимости" if Dobs ≥ Dkp

```

Сама программа использования тестов согласия χ^2 -Пирсона и Колмогорова в пакете MATHCAD может быть, например, такой.

```

ORIGIN := 1
α := 0.05  α = 0.05  l := 10  l = 10  n := 40  n = 40

```

$$\begin{array}{l}
 x1 := \begin{pmatrix} 17 \\ 21 \\ 8 \\ 20 \\ 23 \\ 18 \\ 22 \\ 20 \\ 17 \\ 12 \end{pmatrix} \quad x2 := \begin{pmatrix} 20 \\ 11 \\ 9 \\ 19 \\ 20 \\ 9 \\ 19 \\ 17 \\ 21 \\ 13 \end{pmatrix} \quad x3 := \begin{pmatrix} 17 \\ 22 \\ 22 \\ 10 \\ 20 \\ 20 \\ 15 \\ 19 \\ 20 \\ 20 \end{pmatrix} \quad x4 := \begin{pmatrix} 13 \\ 21 \\ 21 \\ 9 \\ 14 \\ 11 \\ 19 \\ 18 \\ 23 \\ 19 \end{pmatrix} \\
 x := \text{stack}(x1, x2) \quad x := \text{stack}(x, x3) \\
 x := \text{stack}(x, x4)
 \end{array}$$

$$\begin{array}{l}
 m1 := \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad m2 := \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad m3 := \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad m4 := \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \\
 m := \text{stack}(m1, m2) \quad m := \text{stack}(m, m3) \\
 m := \text{stack}(m, m4)
 \end{array}$$

Далее должны следовать тексты всех семи приведенных подпрограмм.

$a := \text{Pirson}(x, m, l, \alpha)$ $a =$ "гипотеза Н0 отвергается с заданным уровнем значимости"

$b := \text{Kolm}(x, m, l, \alpha)$ $b =$ "гипотеза Н0 принимается с заданным уровнем значимости"

Нулевая гипотеза в этой лабораторной работе должна быть сформулирована следующим образом: $H_0 : F(x) = \Phi(x - m_X / \sigma_X)$. Таким образом, с уровнем значимости $\alpha = 0.05$ исследуемая выборка не удовлетворяет нормальному закону.

Задание № 1. Выбрать из табл. 4 вид гипотетического распределения и его параметры, смоделировать в пакете MATHCAD выборку объемом 100 единиц (см. табл. 1) и проверить с уровнем значимости $\alpha = 0.1$ нулевую гипотезу $H_0 : F = F_{\text{гипотет.}}$, исправив, если это необходимо, соответствующие операторы в подпрограммах χ^2 и **Dn**.

Т а б л и ц а 4

Номер фамилии в журнале препода.	Вид распределения	Функция плотности и параметры закона	Функция распределения в пакете MATHCAD	Оценка параметров методом максимального правдоподобия	Числовые значения параметров
1-5	χ^2 -распределение	$k_n(x) = C \cdot x^{n/2-1} e^{-x/2}$ $C = \frac{1}{2^{n/2} \Gamma(n/2)}, x > 0$	pchisq(x, n)	-	n - порядковый номер дня вашего рождения
6-10	Экспоненциальное	$f(x) = \lambda e^{-\lambda x}, x > 0$	rexp(x, λ)	$\hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i}$	$\lambda = (0.5 \cdot n)$, где n - порядковый номер дня рождения
11-15	Стюдента	$s_n(x) = B_n \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$	pt(x, n)	-	n - порядковый номер дня вашего рождения
16-20	F -распределение	$f(x) = C \frac{\frac{n_1-2}{x^2}}{\left(1 + \frac{n_1 x}{2}\right)^{\frac{n_1+n_2}{2}}}$ $x > 0$	pF(x, n_1, n_2)	-	n_1 - ваш номер в журнале преподавателя, n_2 -пор. номер дня рождения
21-25	Гамма-распределение	$f(x) = \frac{1}{\Gamma(\alpha+1)} x^\alpha e^{-x},$ $\beta = 1, x > 0$	pgamma($x, \alpha + 1$)	$\hat{\alpha} = \frac{(\bar{x}^*)^2}{\tilde{D}_X}$	α - пор. номер месяца рождения
26-30	Нормальное	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$	pnorm(x, m, σ)	$\hat{m} = \bar{x}^* = \frac{1}{n} \sum_{i=1}^n x_i$ σ^2 -смещенная дисперсия	m -пор. номер дня рожд., σ^2 -номер фамилии в журнале

6. ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

6.1. Постановка задачи

Изучение рассеяния наблюдаемых величин в эксперименте - один из главных предметов прикладной статистики. Дисперсионный анализ представляет собой метод разложения общей дисперсии совокупности наблюдений на составляющие. Учитывая, что рассеяние наблюдаемой случайной переменной X причинно обусловлено влиянием множества факторов, дисперсионный анализ можно интерпретировать как метод разделения эффектов влияния на наблюдаемые значения X различных подмножеств в общем множестве факторов. Термин «дисперсионный анализ» впервые ввел Фишер и определил его как отделение дисперсии, приписываемой одной группе причин, от дисперсии, приписываемой другим группам. Используемая при этом модель обобщенно может быть представлена в следующем виде.

$$\begin{array}{lcl} \text{Наблюдаемые} & = & \sum \text{параметров,} \\ \text{значения} & & \text{описывающих} \\ & & \text{определяемые эффекты} \end{array} + \begin{array}{l} \sum \text{случайных величин,} \\ \text{описывающих} \\ \text{неопределяемые (6.1.1)} \\ \text{(остаточные) эффекты} \end{array}$$

Чем больше параметров рассматривается в модели, тем меньше будет неопределяемая (остаточная) изменчивость, остающаяся неучтенной, однако некоторая остаточная изменчивость остается всегда.

При исследовании зависимостей одной из наиболее простых является ситуация, когда можно указать только один фактор, влияющий на конечный результат, и этот фактор может принимать лишь конечное число значений (уровней). Такие задачи, называемые задачами однофакторного анализа, весьма часто встречаются на практике. Типичный пример задач однофакторного анализа – сравнение по достигаемым результатам нескольких различных способов действия, направленных на достижение одной цели.

Для применения дисперсионного анализа необходимо вначале построить соответствующую статистическую модель и выяснить структуру экспериментальных данных. Опыт показывает, что при изменении способа обработки наибольшей изменчивости в первую очередь, как правило, подвержено положение случайной величины, которое можно охарактеризовать медианой или средним значением. Следуя этому эмпирическому правилу, в однофакторных задачах также обычно предполагают, что все наблюдения принадлежат некоторому сдвиговому семейству распределений. Часто в качестве такого семейства рассматривается семейство нормальных

распределений и для обработки данных применяются методы дисперсионного анализа. В других случаях предположение о нормальности не является правомерным, и тогда используют различные непараметрические методы анализа, из которых наиболее разработаны ранговые методы.

Введем некоторые общепринятые термины, позволяющие получить в сжатом виде описание структуры эксперимента. Основным является понятие фактора – это качество или свойство, в соответствии с которым классифицируются данные и которое должно оказывать влияние на конечный результат. Каждый фактор имеет несколько различных уровней. Уровень – конкретная реализация фактора – используется для описания рассматриваемого свойства, определяющего каждую категорию применяемой классификации.

Структура или схема эксперимента, обычно называемая планом эксперимента, описывается входящими в него факторами и способом комбинирования разных уровней различных факторов. Наконец, величину результата часто называют откликом.

Для сравнения влияния факторов на результат необходим определенный статистический материал. Обычно его получают следующим образом: каждый из k способов обработки применяется несколько раз (не обязательно одно и то же число раз) к исследуемому объекту, затем результаты регистрируются. Данные таких испытаний могут быть сведены в табл. 5.

Т а б л и ц а 5

Обработки (соответствуют уровням факторов)	1	3	...	k
Результаты наблюдений	x_{11}	x_{12}	...	x_{1k}
	x_{21}	x_{22}	...	x_{2k}

	$x_{n_1 1}$	$x_{n_2 2}$...	$x_{n_k k}$

6.2. Дисперсионный анализ

Для описания данных табл. 5 в большинстве случаев оказывается приемлемой аддитивная модель. Она предполагает, что значение отклика x_{ij} можно представить в виде суммы вклада (воздействия) фактора и независимой от вклада фактора случайной величины. Обычно модель однофакторного дисперсионного анализа записывается в виде

$$x_{ij} = \mu + T_j + \varepsilon_{ij}, \quad i = 1, 2, \dots, n_j, \quad j = 1, 2, \dots, k. \quad (6.2.1)$$

Здесь μ - математическое ожидание X в генеральной совокупности, $T_j = \mu_j - \mu$ - эффект влияния j -го уровня фактора, т.е. отклонение от общего среднего уровня при j -й обработке, μ_j - матожидание X в j -й группе, ε_{ij} - случайная ошибка наблюдений.

Обычно предполагается только непрерывность закона распределения величин ε_{ij} и их независимость. Однако во многих случаях о распределении ε_{ij} можно сказать больше, например, предполагают, что величины $\varepsilon_{ij} \in N(0, D)$, т.е. имеют нормальное распределение с нулевым средним и общей дисперсией, которая неизвестна. Дополнительная информация о законе распределения случайных величин ε_{ij} позволяет использовать более сильные методы в модели однофакторного анализа как для проверки гипотез, так и для оценки параметров. Совокупность этих методов носит название однофакторного дисперсионного анализа.

Чаще всего дисперсионный анализ основан на следующих правдоподобных допущениях о случайных величинах ε_{ij} .

1. Математическое ожидание каждой остаточной случайной величины равно нулю. Это означает, что вся изменчивость в математических ожиданиях охватывается параметрами. Это очень правдоподобное предположение, ибо влияние второго члена в модели (6.1.1) всегда меньше первого.

2. Остаточные случайные величины взаимно независимы, Это допущение не столь очевидно, как первое. Смысл его состоит в том, что между различными наблюдениями не существует какой-либо связи, которую нельзя было бы объяснить с помощью членов, описывающих определяемые эффекты.

3. Все остаточные случайные величины имеют одинаковое среднеквадратическое отклонение. Это предположение об однородности дисперсий. Во многих случаях это допущение не выполняется, поэтому прежде чем проводить дисперсионный анализ какого-либо набора данных, важно рассмотреть возможные колебания D .

4. Каждая остаточная случайная величина распределена по нормальному закону. В общем случае справедливость этого допущения наименее вероятна, чем трех остальных. Значительная часть дисперсионного анализа может проводиться без принятия этого допущения, необходимого лишь для обоснования использования некоторых формально точных критериев для проверки значимости и формул оценивания.

Допущения, описанные выше, имеют форму:

- 1) $M(\varepsilon_{ij}) = 0$;
- 2) ε_{ij} взаимно независимы;
- 3) $D(\varepsilon_{ij}) = D = \text{const}$;
- 4) $\varepsilon_{ij} \in N(0, D)$.

(6.2.2)

Если уровни исследуемого фактора фиксированы, то эффекты $T_j = \mu_j - \mu$ являются фиксированными постоянными и их сумма равна нулю, так как в эксперименте выбраны все возможные значения уровней.

Обратимся теперь к табл. 5. Изменчивость или вариация наблюдаемых значений x_{ij} может быть вызвана изменчивостью уровней фактора и изменчивостью значения случайных величин, описывающих неопределяемые эффекты.

Вычислим среднее значение для каждой группы и общее среднее всех наблюдений:

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} = \frac{1}{n_j} x_{\bullet j}, \quad j = 1, 2, \dots, k, \quad \bar{\bar{x}} = \frac{1}{n} \sum_{j=1}^k n_j \bar{x}_j = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij}, \quad n = \sum_{j=1}^k n_j,$$

$$\bar{\varepsilon}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \varepsilon_{ij}, \quad \bar{\bar{\varepsilon}} = \frac{1}{n} \sum_{j=1}^k n_j \bar{\varepsilon}_j = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} \varepsilon_{ij}. \quad \text{С учетом этих формул пер-}$$

вое уравнение модели однофакторного дисперсионного анализа (6.2.1) можно упростить. Просуммируем формулу (6.2.1) по i в пределах от еди-

ницы до n_j . Получим $\sum_{i=1}^{n_j} x_{ij} = \sum_{i=1}^{n_j} \mu + \sum_{i=1}^{n_j} T_j + \sum_{i=1}^{n_j} \varepsilon_{ij}$ или

$$n_j \bar{x}_j = n_j \mu + n_j T_j + n_j \bar{\varepsilon}_j. \quad \text{Окончательно}$$

$$\bar{x}_j = \mu + T_j + \bar{\varepsilon}_j. \quad (6.2.3)$$

Продолжим суммирование по j в пределах от единицы до k . Тогда

$$\sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij} = \sum_{j=1}^k \sum_{i=1}^{n_j} \mu + \sum_{j=1}^k \sum_{i=1}^{n_j} T_j + \sum_{j=1}^k \sum_{i=1}^{n_j} \varepsilon_{ij}, \quad \text{т.е.} \quad n \bar{\bar{x}} = \mu \sum_{j=1}^k n_j + \sum_{j=1}^k T_j n_j + \sum_{j=1}^k n_j \bar{\varepsilon}_j.$$

Так как $\mu_j = \mu + T_j$ - отклонение значений μ_j от среднего значения μ , то

$$\mu = \frac{1}{n} \sum_{j=1}^k n_j \mu_j, \quad \text{т.е. средневзвешенное значений } \mu_j.$$

$$\text{Тогда } \mu n = \sum_{j=1}^k n_j \mu_j = \sum_{j=1}^k n_j (\mu + T_j) = \sum_{j=1}^k n_j \mu + \sum_{j=1}^k n_j T_j = \mu n + \sum_{j=1}^k n_j T_j .$$

Отсюда $\sum_{j=1}^k n_j T_j = 0$. Окончательно второе уравнение модели имеет вид

$$\bar{n\bar{x}} = n\bar{\mu} + n\bar{\varepsilon} \text{ или}$$

$$\bar{x} = \bar{\mu} + \bar{\varepsilon} . \quad (6.2.4)$$

Вычтем из уравнения (6.2.3) уравнение (6.2.4), получим $\bar{x}_j - \bar{x} = T_j + \varepsilon_j - \bar{\varepsilon}$. Тогда $M(\bar{x}_j - \bar{x}) = M(T_j + \varepsilon_j - \bar{\varepsilon}) = T_j$, так как $M(\varepsilon_{ij}) = 0$ по допущениям (6.2.2). Аналогично $D(\bar{x}_j - \bar{x}) = D$.

Выведем теперь основное тождество дисперсионного анализа. Рассмотрим

$$\begin{aligned} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 &= \sum_{j=1}^k \sum_{i=1}^{n_j} [(x_{ij} - \bar{x}_j) - (\bar{x} - \bar{x}_j)]^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{x} - \bar{x}_j)^2 - \\ &- 2 \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(\bar{x} - \bar{x}_j) = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{x} - \bar{x}_j)^2 - 2 \sum_{j=1}^k (\bar{x} - \bar{x}_j) \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j). \end{aligned}$$

$$\text{Но } \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j) = \sum_{i=1}^{n_j} x_{ij} - \sum_{i=1}^{n_j} \bar{x}_j = n_j \bar{x}_j - \bar{x}_j \sum_{i=1}^{n_j} 1 = n_j \bar{x}_j - \bar{x}_j n_j = 0 , \quad \text{тогда}$$

последнее выражение примет вид

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{x} - \bar{x}_j)^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \text{ или}$$

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 . \quad (6.2.5)$$

Таким образом, общая сумма квадратов отклонений наблюдений от общего среднего \bar{x} разбивается на сумму квадратов отклонений выборочных средних \bar{x}_j от общего среднего \bar{x} и сумму квадратов отклонений наблюдений x_{ij} от выборочных средних групп \bar{x}_j (внутри групп), т.е.

$$Q = Q_1 + Q_2, \text{ где } Q = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2, \quad Q_1 = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 ,$$

$$Q_2 = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2.$$

В формулу (6.2.5) входят три члена. Рассмотрим их подробнее. Член

$$\frac{Q_1}{n} = \frac{1}{n} \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2$$

равен дисперсии групповых средних, в него входят

k слагаемых, «свобода» изменения которых ограничена одним соотноше-

$$\text{нием } \bar{x} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij} = \frac{1}{n} \sum_{j=1}^k n_j \bar{x}_j = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_k \bar{x}_k}{n_1 + n_2 + \dots + n_k}.$$

Из этой

формулы вытекает единственное уравнение связи $(\bar{x}_1 - \bar{x})n_1 + (\bar{x}_2 - \bar{x})n_2 + \dots + (\bar{x}_k - \bar{x})n_k = 0$. Поэтому говорят, что величина Q_1 имеет $(k - 1)$ степень свободы.

Величина Q_2/n равна средней из групповых дисперсий. В формулу расчета Q_2 входят $n_1 + n_2 + \dots + n_k = n$ слагаемых. Свобода первых n_1

слагаемых ограничена одним соотношением $\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{i,1}$ или

$(x_{11} - \bar{x}_1) + (x_{21} - \bar{x}_1) + \dots + (x_{n_1 1} - \bar{x}_1) = 0$. Таким образом, «свобода» изменения k слагаемых ограничена k условиями. Это означает, что величина Q_2 имеет $(n - k)$ степеней свободы.

Наконец, в формулу Q/n входят $n_1 + n_2 + \dots + n_k = n$ слагаемых.

На них наложено одно ограничение $\bar{x} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij} =$

$$= \frac{(x_{11} + x_{12} + \dots + x_{1n_1}) + (x_{21} + x_{22} + \dots + x_{2n_2}) + \dots + (x_{k1} + x_{k2} + \dots + x_{kn_j})}{n}$$

или

$$(x_{11} - \bar{x}) + (x_{12} - \bar{x}) + \dots + (x_{1n_1} - \bar{x}) + (x_{21} - \bar{x}) + (x_{22} - \bar{x}) + \dots + (x_{2n_2} - \bar{x}) + \dots + (x_{k1} - \bar{x}) + (x_{k2} - \bar{x}) + \dots + (x_{kn_j} - \bar{x}) = 0.$$

Поэтому Q имеет $(n - 1)$ степень свободы.

По 3-му условию (6.2.2) все генеральные групповые дисперсии должны быть равными, т.е. $D_1 = D_2 = \dots = D_k = D$. Найдем несмещенные оценки D .

Во-первых, убедимся в том, что несмещенная оценка дисперсии D равна $Q_2/(n-k)$, т.е. $M\left(\frac{Q_2}{n-k}\right) = D$. Действительно,

$$\begin{aligned} M\left(\frac{Q_2}{n-k}\right) &= \frac{1}{n-k} M\left[\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2\right] = \frac{1}{n-k} \sum_{j=1}^k M\left[\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2\right] = \\ &= \frac{1}{n-k} \sum_{j=1}^k M(n_j D_j^*), \quad \text{где} \quad D_j^* = \frac{1}{n_j} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = \\ &= \frac{(x_{1j} - \bar{x}_j)^2 + (x_{2j} - \bar{x}_j)^2 + \dots + (x_{n_j j} - \bar{x}_j)^2}{n_j} \end{aligned}$$

- выборочная групповая дисперсия, вычисленная по результатам наблюдений при j -м уровне фактора, $j = 1, 2, \dots, k$.

$$\begin{aligned} \text{Далее} \quad M\left(\frac{Q_2}{n-k}\right) &= \frac{1}{n-k} \sum_{j=1}^k M(n_j D_j^*) = \frac{1}{n-k} \sum_{j=1}^k M[(n_j - 1)\tilde{D}_j] = \\ &= \frac{1}{n-k} \sum_{j=1}^k (n_j - 1) M(\tilde{D}_j) = \frac{1}{n-k} \sum_{j=1}^k (n_j - 1) D_j = \frac{1}{n-k} D_j \sum_{j=1}^k (n_j - 1) = \\ &= \frac{1}{n-k} D_j (n-k) = D_j = D, \text{ так как } \tilde{D}_j = \frac{n_j}{n_j - 1} D_j^* \text{ - несмещенные оцен-} \end{aligned}$$

ки групповых дисперсий, т.е. $M(\tilde{D}_j) = D_j$. Последнее равенство верно только в том случае, когда наблюдения в j -й группе независимы и проводятся в одинаковых условиях. Это справедливо по 2-му условию (6.2.2).

Итак, $M\left(\frac{Q_2}{n-k}\right) = D$.

Рассмотрим теперь вопрос о различии обработок (факторов) в табл. 5. Он сводится к выяснению различия между T_1, T_2, \dots, T_k . Гипотеза об однородности данных означает равенства $\mu_1 = \mu_2 = \dots = \mu_k$, т.е. $T_1 = T_2 = \dots = T_k = 0$. Альтернатива об упорядоченности эффектов обработки (о влиянии фактора) превращается в $T_1 \leq T_2 \leq \dots \leq T_k$, а различие между i -м и j -м уровнем фактора, естественно, характеризуется величиной $\mu_i - \mu_j = T_i - T_j$. Итак, пусть $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ и выполняются условия (6.2.2).

Тогда при каждом уровне фактора величина X будет иметь нормальное распределение с одним и тем же математическим ожиданием и одной и той же дисперсией, равной D , т.е. переход от одного уровня фактора к другому не вносит никаких изменений: имеется одна генеральная совокупность, и результаты наблюдений, приведенные в табл. 5 – это выборка объема n из этой генеральной совокупности. А так как наблюдения независимы и проведены в одинаковых условиях, то несмещенная оценка об-

щей дисперсии D и есть $\tilde{D} = \frac{1}{n-1} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$. Таким образом,

$$M \left[\frac{1}{n-1} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 \right] = M \left(\frac{1}{n-1} Q \right) = D.$$

Убедимся теперь в том, что при выполнении условий гипотезы H_0 и условий (6.2.2) величина $\frac{1}{k-1} Q_1$ также является несмещенной оценкой

$$\begin{aligned} \text{общей дисперсии.} \quad M \left(\frac{Q_1}{k-1} \right) &= \frac{1}{k-1} M(Q_1) = \frac{1}{k-1} M(Q - Q_2) = \\ &= \frac{1}{k-1} [M(Q) - M(Q_2)] = \frac{1}{k-1} [(n-1)D - (n-k)D] = D. \end{aligned}$$

Итак, имеются три несмещенные оценки одной и той же дисперсии D , причем оценка $Q_2/(n-k)$ является несмещенной оценкой в любом случае, а оценки $Q/(n-1)$ и $Q_1/(k-1)$ – только при выполнении гипотезы H_0 , т.е. только тогда, когда исследуемый фактор не влияет на результат.

Проверка гипотезы H_0 о равенстве групповых математических ожиданий основывается на сравнении дисперсий $s_1^2 = Q_1/(k-1)$ и $s_2^2 = Q_2/(n-k)$. Вспомним сначала механизм создания случайной величины с χ^2 -распределением (см. подразд. 4.7, п. 2). Поскольку верны допущения (6.2.2), а s_1^2 и s_2^2 являются несмещенными оценками дисперсии D , то

$$\frac{(k-1)s_1^2}{D} \in \chi_{k-1}^2 \text{ и } \frac{(n-k)s_2^2}{D} \in \chi_{n-k}^2. \quad (6.2.6)$$

Величины χ_{k-1}^2 и χ_{n-k}^2 независимы в силу независимости s_1^2 и s_2^2 .

Тогда (см. подразд. 2.3) отношение $\frac{\chi_{k-1}^2/(k-1)}{\chi_{n-k}^2/(n-k)} = \frac{s_1^2}{s_2^2} = \frac{Q_1/(k-1)}{Q_2/(n-k)}$ имеет

F - распределение с числом степеней свободы $k-1$ и $n-k$,

$$\frac{Q_1/(k-1)}{Q_2/(n-k)} \in F_{k-1, n-k}. \quad (6.2.7)$$

Итак, $H_0 : T_1 = T_2 = \dots = T_k = 0$,
 $H_1 : T_i \neq T_j, i \neq j, 1 \leq i \leq k, 1 \leq j \leq k$. Гипотеза H_0 принимается на уровне значимости α , если выборочное значение статистики $F_{k-1, n-k}$ меньше $F_{1-\alpha/2}(k-1, n-k)$ или больше $F_{\alpha/2}(k-1, n-k)$. В этом случае \bar{x} и $s_2^2 = Q_2/(n-k)$ являются несмещенными оценками математического ожидания и дисперсии выборки (наблюдений x_{ij}). Оценка s_2^2 не зависит от вида нулевой гипотезы H_0 . Оценка s_1^2 существенно использует основное предположение гипотезы H_0 . Она дает близкий к D результат только в том случае, когда гипотеза H_0 верна. При нарушении H_0 оценка s_1^2 имеет тенденцию к возрастанию, тем большому, чем больше отклонение от H_0 . Сопоставляя друг с другом две эти оценки, мы можем заключить, что H_0 следует отвергнуть, если они оказываются значительно различны.

Практически вычисление Q, Q_1 и Q_2 удобно проводить по формулам: $Q = A - C$, $Q_1 = B - C$, $Q_2 = A - B$, где

$$A = \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij}^2, B = \sum_{j=1}^k \frac{1}{n_j} \left(\sum_{i=1}^{n_j} x_{ij} \right)^2 = \sum_{j=1}^k \frac{1}{n_j} x_{\bullet j}^2, C = \frac{1}{n} \left(\sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij} \right)^2 = \frac{1}{n} \bar{x}^2. \quad (6.2.8)$$

Для контроля правильности расчетов используют тождество $Q = Q_1 + Q_2$.

Если гипотеза о равенстве средних отклоняется, то требуется определить, какие именно группы средних имеют значимое различие. Для этого часто используются не сами оценки величин T_j , а некоторые линейные комбинации этих величин. Для их определения вводится понятие контраста. Контрастом параметров T в модели аддитивного влияния фактора на отклик называется величина $L_k = \sum_{j=1}^k c_j T_j$, где $c_j, j = 1, 2, \dots, k$ - константы, однозначно определяемые из формулировок основной и альтернативной гипотез, причем $\sum_{j=1}^k c_j = 0$. Ясно, что разность $T_i - T_j$ является про-

стейшим примером контраста, когда $c_i = 1$, $c_j = -1$, $c_l = 0$ при всех $l \neq i$ и $l \neq j$. Оценки контрастов таковы:

$$M(L_k) = \hat{L}_k = \sum_{j=1}^k c_j \bar{x}_j, D(L_k) = \hat{D}_k = \frac{Q_2}{n-k} \sum_{j=1}^k \frac{c_j^2}{n_j}. \quad (6.2.9)$$

Граница доверительного интервала для L_k имеет вид

$$\hat{L}_k \pm \sqrt{\hat{D}_k} \cdot \sqrt{(k-1)F_{1-\alpha}(k-1, n-k)}. \quad (6.2.10)$$

Пример. Предполагается, что выборки получены из нормально распределенных генеральных совокупностей с равными дисперсиями. Проверить гипотезу о равенстве средних. Если H_0 принимается, найти несмещенные оценки среднего и дисперсии. В случае отклонения H_0 провести попарное сравнение средних, используя метод линейных контрастов. Принять $\alpha = 0.05$.

Номер выборки	Наблюдения				
1	6	5	12	9	10
2	14	11	5	6	-
3	12	4	7	-	-

Решение

Быстрее всего задача решается по формулам (6.2.8). Для этого продолжим исходную таблицу еще несколькими столбцами.

Номер выборки	Наблюдения					n_j	$\sum_{i=1}^{n_j} x_{ij}$	$\sum_{i=1}^{n_j} x_{ij}^2$	$\frac{1}{n_j} \left(\sum_{i=1}^{n_j} x_{ij} \right)^2$
1	6	5	12	9	10	5	42	386	352.8
2	14	11	5	6	-	4	36	378	324
3	12	4	7	-	-	3	23	209	176.3

Тогда $n = 12$, $k = 3$, $A = 973$, $B = 851.133$, $C = 850.083$. Отсюда $Q = A - C = 122.917$, $Q_1 = B - C = 1.05$, $Q_2 = A - B = 121.867$. Проверим справедливость расчетов: $Q_1 + Q_2 = 1.05 + 121.867 = 122.917 = Q$.

Итак, $H_0 : T_1 = T_2 = T_3 = 0$, $H_1 : T_i \neq T_j$, $i \neq j$, $1 \leq i \leq 3$, $1 \leq j \leq 3$.

$$F_{\text{теор.}}^{(1)} = F_{\alpha/2}(k-1, n-k) = F_{0.025}(3-1, 12-3) = F_{0.025}(2, 9),$$

$$F_{\text{теор.}}^{(2)} = F_{1-\alpha/2}(k-1, n-k) = F_{0.975}(2, 9). \quad F_{\text{выб.}} = \frac{Q_1/2}{Q_2/9} = \frac{0.525}{13.541} = 0.039.$$

Таблицы F - распределения с квантилями для малых вероятностей очень редки, поэтому квантили $F_{\text{теор.}}^{(1)}$ и $F_{\text{теор.}}^{(2)}$ были вычислены в пакете MATHCAD. $F_{\text{теор.}}^{(1)} = 0.025$, $F_{\text{теор.}}^{(2)} = 5.715$. Поскольку

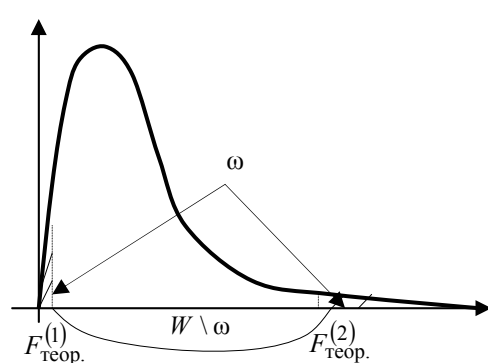


Рис. 6.1. Критическая область статистики для гипотезы о равенстве средних

$$F_{\text{теор.}}^{(1)} < F_{\text{выб.}} < F_{\text{теор.}}^{(2)}, \quad \text{т.е.}$$

$F_{\text{выб.}} \in W \setminus \omega$, то гипотеза H_0 о равенстве средних в исходной общей выборке, состоящей из трех разных подвыборок, принимается (рис. 6.1). Несмещенными оценками среднего и дисперсии здесь будут величи-

$$\bar{x} = \frac{1}{n} \sum_{j=1}^3 \sum_{i=1}^{n_j} x_{ij} = 8.417 \quad \text{и}$$

$$s_2^2 = Q_2 / (n - k) = 13.541.$$

6.3. Ранговый однофакторный анализ

В последние годы очень сильно были развиты методы математической статистики, для которых не требуются никакие предположения о распределении, за исключением предположения о том, что это распределение непрерывно. Эти методы называются непараметрическими или свободными от распределения.

Если мы ничего не знаем о распределении наблюдений, то непосредственно использовать для проверки нулевой гипотезы количественные значения наблюдений x_{ij} становится затруднительно. В этом случае проще всего опираться в своих выводах только на отношение «больше - меньше» между наблюдениями, так как они не зависят от распределения наблюдений.

В этом случае вся полезная информация содержится в рангах. Получим из исходной выборки вариационный ряд, т.е. расположим выборочные значения в порядке возрастания. Каждой величине из этого ряда сопоставим ее ранг, равный порядковому номеру величины в общем вариационном ряду. Заметим, что если наблюдения однородны, т.е. вся выборка взята из одной и той же генеральной совокупности, то любое распределение рангов равновероятно, а общее число способов группировки рангов,

например, при двух подвыборках объемов n и m равно числу способов, которыми можно извлечь m предметов из $N = n + m$, т.е. C_{n+m}^m .

Соответствующие критерии для проверки нулевой гипотезы называются ранговыми, они пригодны для любых непрерывных распределений наблюдений. Более того, они годятся и тогда, когда измерения x_{ij} сделаны в порядковой шкале, например, являются тестовыми баллами или экспертными оценками.

Основные формулы рангового однофакторного анализа выведены в предположении, что среди чисел x_{ij} нет совпадений. При наличии совпадений используются средние ранги, при этом теоретическая схема действует как приближенная, и надежность ее выводов снижается. Для учета совпадений вводятся специальные поправки.

Припишем каждому наблюдению x_{ij} в общем вариационном ряду его ранг r_{ij} . Тогда табл. 5 преобразуется в табл. 6.

Т а б л и ц а 6

Обработки (соответствуют уровням факторов)	1	3	...	k
Ранги результатов наблюдений	r_{11}	r_{12}	...	r_{1k}
	r_{21}	r_{22}	...	r_{2k}

	$r_{n_1 1}$	$r_{n_2 2}$...	$r_{n_k k}$

Общая методика проверки статистических гипотез рекомендует сконструировать некоторую статистику, т.е. функцию от рангов r_{ij} , которая легла бы в основу критерия проверки гипотезы. Основное требование к этой статистике следующее: ее распределение при гипотезе H_0 должно заметно отличаться от ее распределения при альтернативах. Например, часто в качестве статистики берут сумму рангов одной подвыборки. Рациональность такой процедуры состоит в том, что если одно распределение (одной подвыборки) смещено относительно другого, то это должно проявиться в том, что маленькие ранги должны в основном соответствовать одной подвыборке, а большие – другой, вследствие чего соответствующие суммы рангов должны быть маленькими или большими в зависимости от того, какая альтернатива имеет место.

6.4. Критерий Краскела - Уоллиса (H-критерий)

Если нельзя сказать что-то определенное об альтернативах к H_0 , можно воспользоваться для ее проверки свободным от распределения H-критерием. Он был предложен Краскелом и Уоллисом и является обобщением двухвыборочного критерия Вилкоксона*.

Построим общий вариационный ряд, содержащий $n_1 + n_2 + \dots + n_k = n$ элементов, где n_j - число наблюдений в j -й подвыборке (на j -м уровне фактора). Используем обозначения подразд. 6.2. Тогда $R_{\bullet j}$ - сумма рангов каждой обработки, т.е. каждого столбца табл. 6, а \bar{R}_j - среднее арифметическое этих рангов. Формулы для их нахождения таковы:

$$R_{\bullet j} = \sum_{i=1}^{n_j} r_{ij}, \quad \bar{R}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{ij} = \frac{1}{n_j} R_{\bullet j}.$$

$$M(\bar{R}_j) = M\left(\frac{1}{n_j} R_{\bullet j}\right) = \frac{1}{n} \sum_{k=1}^n k = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2} \text{ как среднее арифмети-}$$

ческое всех рангов от единицы до n , а $1 + 2 + \dots + n = \frac{n(n+1)}{2}$. Отсюда

$$M(R_{\bullet j}) = \frac{n_j(n+1)}{2}.$$

Если между столбцами нет систематических различий, средние ранги \bar{R}_j не должны значительно отличаться от среднего ранга, рассчитанного по всей совокупности чисел r_{ij} . Математическое ожидание среднего ранга, очевидно, равно $M(\bar{R}) = (n+1)/2$.

Более сложным образом рассчитывается дисперсия. Для \bar{R}_j она равна $D(\bar{R}_j) = \frac{(n+1)(n-n_j)}{12n_j}$. Если $n \rightarrow \infty$, то дробь $\frac{\bar{R}_j - M(\bar{R}_j)}{\sqrt{D(\bar{R}_j)}}$ имеет в

пределе стандартное нормальное распределение, что и использовали Краскел и Уоллис для построения статистики критерия, которую они обозначили буквой H и которая имеет вид

* Фрэнк Вилкоксон (Уилкоксон) (1892-1965) – американский математик.

$$H = \sum_{j=1}^k \frac{\left[\bar{R}_j - \frac{n+1}{2} \right]^2}{(n+1)(n-n_j)} \left(1 - \frac{n_j}{n} \right). \quad (6.4.1)$$

Краскел и Уоллис показали, что асимптотически статистика H имеет χ^2 -распределение с $(k-1)$ степенью свободы, где k - число подвыборок (уровней фактора). Часто статистика H записывается в одном из следующих двух видов:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^k n_j \left(\bar{R}_j - \frac{n+1}{2} \right)^2, \quad (6.4.2)$$

или

$$H = \frac{12}{n(n+1)} \sum_{j=1}^k \frac{R_{\bullet j}^2}{n_j} - 3(n+1). \quad (6.4.3)$$

Если два или более наблюдений совпадают, то наилучшая процедура состоит в том, что совпавшим наблюдениям нужно приписать один и тот же ранг, равный среднему арифметическому рангов, которые эти наблюдения должны были получить, если бы они не совпали. Эта операция оставляет без изменения сумму рангов и математическое ожидание суммы рангов. Но формула для вычисления дисперсии меняется, так как дисперсия статистики \bar{R}_j зависит от суммы квадратов рангов, которая от такой замены изменится. Изменится и вид статистики H , поэтому ее исправляют соответствующей поправкой.

Если совпадений много, рекомендуется использовать модифицированную форму статистики H' :

$$H' = \frac{H}{1 - \sum_{j=1}^p \frac{T_j}{n^3 - n}}, \quad (6.4.4)$$

где p - число групп совпадающих наблюдений, $T_j = \binom{t_j^3 - t_j}{3}$, t_j - число совпадающих наблюдений в группе с номером j .

Пример. Кислота непрерывным образом концентрируется на некотором типе оборудования, в результате чего часть оборудования ржавеет и со временем разрушается. Потери металла (в сотнях тонн) за период от установки оборудования до момента разрушения некоторой его части зафиксированы в таблице для трех литейных мастерских А, В и С. Прове-

рять нулевую гипотезу, по которой средняя продолжительность службы металла одна и та же для всех трех мастерских.

Мастерская	Потери металла									
	84	60	40	47	34	46				
A	67	92	95	40	98	60	59	108	86	117
B	46	93	100	92	92					
C										

Решение

Никаких правдоподобных предположений о вероятностном распределении потерь металла в этой задаче сделать нельзя. Воспользуемся ранговым методом Краскела – Уоллиса. Надо заметить, что величины, приведенные в исходной таблице, имеют смысл сами по себе, а не только в сравнении с другими величинами. Хотя при переходе от величин потерь металла к их рангам происходит определенная потеря информации, но такая информация, во-первых, не столь значительна, во-вторых, компенсируется тем, что от неизвестного закона распределения величин x_{ij} мы переходим к величинам r_{ij} , распределение которых при гипотезе H_0 известно.

Основная гипотеза H_0 постулирует постоянный срок службы металла во всех трех мастерских, т.е. постоянный уровень потерь, следовательно, однородность исходных выборок. Обозначим потери металла в j -й группе через μ_j . Тогда

$$H_0 : \mu_1 = \mu_2 = \mu_3,$$

$$H_1 : \mu_i \neq \mu_j, i \neq j, 1 \leq i \leq 3, 1 \leq j \leq 3.$$

Сначала получим вариационный ряд и припишем каждому наблюдению его ранг. В связи с наличием в таблице совпадений будем пользоваться средними рангами.

Наблюдения	34	40	40	46	46	47	59	60	60	67	84
Номер наблюдений в вариационном ряду	1	2	3	4	5	6	7	8	9	10	11
Ранг	1	2.5	2.5	4.5	4.5	6	7	8.5	8.5	10	11
Наблюдения	86	92	92	92	93	95	98	100	108	117	
Номер наблюдений в вариационном ряду	12	13	14	15	16	17	18	19	20	21	
Ранг	12	14	14	14	16	17	18	19	20	21	

Общее количество наблюдений $n = 21$. Составим теперь из исходной таблицы таблицу рангов и дополним ее двумя столбцами, содержащими $R_{\bullet j}$ и \bar{R}_j .

Мастер- ская	Ранги потерь металла										$R_{\bullet j}$	\bar{R}_j
А	11	8.5	2.5	6	1	4.5					33.5	5.583
В	10	14	17	2.5	18	8.5	7	20	12	21	130	13.00
С	4.5	16	19	14	14						67.5	13.50

Для вычисления статистики Краскела – Уоллиса удобнее использовать формулу (6.4.3). Тогда $H = \frac{12}{21 \cdot 22} \left(\frac{33.5^2}{6} + \frac{130^2}{10} + \frac{67.5^2}{5} \right) - 3 \cdot 22 = 6.423$. Так как имеются совпадения, скорректируем статистику H .

В нашем случае имеются четыре группы совпадающих наблюдений: 40, 40; 46, 46; 60, 60; 92, 92, 92. Вычислим поправки по формуле (6.4.4): $T_1 = (2^3 - 2) = 6$, $T_2 = 6$, $T_3 = 6$, $T_4 = (3^3 - 3) = 24$. Знаменатель дроби в выражении для H' равен: $1 - \sum_{j=1}^4 \frac{T_j}{(21^3 - 21)} = 1 - \frac{6 + 6 + 6 + 24}{9240} = 0.995$.

Тогда $H' = \frac{H}{0.995} = 6.455$.

Как было указано, величина H асимптотически распределена по закону χ^2 с числом степеней свободы $k - 1$, то есть в данном случае равным двум. Найдем квантиль χ^2 -распределения: $\chi_{0.95,2}^2 = 5.99$. Таким образом, при использовании правостороннего критерия $H' > \chi_{0.95,2}^2$, т.е. $H' \in \omega$, и гипотеза H_0 должна быть отвергнута с уровнем значимости $\alpha = 0.05$.

6.5. Лабораторная работа № 8. Однофакторный ранговый и дисперсионный анализ в статистическом пакете STATGRAPHICS

Дисперсионный анализ применяется для обнаружения влияния выделенного набора факторов на результативный признак. Общая идея дисперсионного анализа состоит в разложении общей дисперсии результативного признака на части, обусловленные влиянием контролируемых факторов, и остаточную дисперсию, вызываемую случайными обстоятельствами.

Известно много моделей дисперсионного анализа. Они классифицируются, с одной стороны, по математической природе факторов (детерминированные, случайные и смешанные), с другой стороны – по числу контролируемых факторов (однофакторные и многофакторные модели). По способу организации исходных данных среди моделей дисперсионного анализа выделяют полные и неполные k -факторные планы, полные и неполные блочные планы и рандомизированные блочные планы. В STATGRAPHICS Plus for Windows реализованы все перечисленные модели дисперсионного анализа.

Решим в пакете STATGRAPHICS следующую задачу однофакторного дисперсионного анализа.

Время химической реакции при различном содержании катализатора распределилось следующим образом (в секундах):

Содержание катализат., %	Номер эксперимента												Сумма
	1	2	3	4	5	6	7	8	9	10	11	12	
5	5.9	6.0	7.0	6.5	5.5	7.0	8.1	7.5	6.2	6.4	7.1	6.9	80.1
10	4.0	5.1	6.2	5.3	4.5	4.4	5.3	5.4	5.6	5.2	-	-	51.0
15	8.2	6.8	8.0	7.5	7.0	7.2	7.9	8.1	8.5	7.8	8.1	-	85.1

	observ	factor	Col_3	Col_4
1	5,9	5		
2	6	5		
...		
11	7,1	5		
12	6,9	5		
13	4	10		
14	5,1	10		
...		
21	5,6	10		
22	5,2	10		
23	8,2	15		
24	6,8	15		
...		
32	7,8	15		
33	8,1	15		

Рис. 6.2. Электронная таблица STATGRAPHICS

Доступ к процедурам анализа осуществляется из пункта

Предполая, что выборки получены из нормально распределенных генеральных совокупностей с равными дисперсиями, проверить нулевую гипотезу H_0 о равенстве средних. Принять $\alpha = 0.1$.

Раскроем электронную таблицу SRATGRAPHICS и введем в нее значения наблюдений (величины x_{ij} - значения результативного признака) и значения градаций фактора (можно вводить закодированные значения, например, 1, 2, 3), так как это пока-

меню Compare→Analysis of Variance→One-Way ANOVA (однофакторный дисперсионный анализ). Сокращение ANOVA происходит от выражения «Analysis of variance». В отечественной литературе вместо термина «анализ вариаций» используется термин «дисперсионный анализ».

Сразу же появляется окно однофакторного дисперсионного анализа (рис. 6.3). В окне Dependent Variable (Зависимая переменная) введем Observ, а в окно Factor (Фактор) имя Factor. Нажмем ОК. На экране появится сводка однофакторного дисперсионного анализа, в которой подтверждается, что введено 33 наблюдения, для которых зафиксировано три уровня фактора. Внизу под этими включено сообщение StatAdvisor с рекомендациями по проведению дальнейшего анализа.

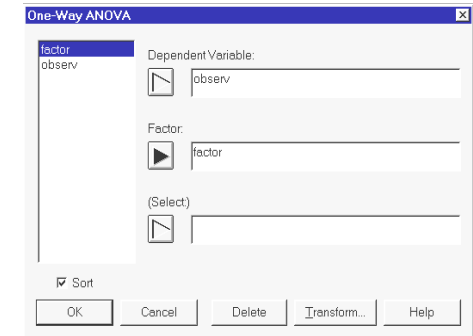


Рис. 6.3. Окно для задания переменных

В появившемся дополнительном меню откроем окно Tabular Options и отметим все процедуры этого меню. Укажем назначение всех входящих в это меню процедур.

Analysis Summary (Сводка анализа). Заставка этого окна уже открыта. На ней указаны самые общие сведения о выборке.

Summary Statistics (Описание данных). Содержание числовой информации, помещенной на этой заставке, понятно из контекста. Сначала анализируется информация о факторе. Приводятся значения уровней фактора, количество наблюдений на каждом уровне, средние, дисперсии и стандартные отклонения на каждом уровне и по всей выборке. Затем приводятся наименьшие и наибольшие значения членов выборки, их стандартные асимметрии и эксцессы, наконец, в последней таблице помещены суммарные значения наблюдений по факторам и в целом по выборке.

ANOVA Table (Таблица дисперсионного анализа). Назначение этой таблицы – дать ответ на вопрос о наличии значимого влияния уровней фактора на исследуемый отклик, т.е. на присутствие эффектов обработки. В первой колонке Source (Источник вариации) указаны две части, на которые разлагается общая дисперсия по формуле (6.2.5) Between groups (Между группами) и Within groups (Внутри групп). Далее приводится общая дисперсия Total (corr.) (Итого (скорректированное значение)). Второй столбец содержит сумму квадратов между группами, внутри групп и общую, т.е. величины Q_1 , Q_2 и Q , третий – соответствующее число степеней свободы. В четвертом столбце находятся значения дисперсий: между

группами величина s_1^2 , внутри групп величина s_2^2 . В столбце F - ratio выводится значение F -статистики, наконец, столбец p-Value содержит уровень значимости этой статистики (рис. 6.4).

Analysis Summary					
Dependent variable: observ					
Factor: factor					
Number of observations: 33					
Number of levels: 3					
Summary Statistics for observ					
factor	Count	Average			
5	12	6,675			
10	10	5,1			
15	11	7,73636			
Total	33	6,55152			
factor	Variance	Standard deviation			
5	0,538409	0,733764			
10	0,411111	0,641179			
15	0,292545	0,540875			
Total	1,53883	1,24049			
factor	Minimum	Maximum			
5	5,5	8,1			
10	4,0	6,2			
15	6,8	8,5			
Total	4,0	8,5			
factor	Stnd. skewness	Stnd. kurtosis			
5	0,423072	-0,0936714			
10	-0,265288	0,0096087			
15	-0,75687	-0,536064			
Total	-0,646217	-1,08188			
factor	Sum				
5	80,1				
10	51,0				
15	85,1				
Total	216,2				
ANOVA Table for observ by factor					
Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	36,6945	2	18,3472	43,87	0,0000
Within groups	12,548	30	0,418265		
Total (Corr.)	49,2424	32			

Рис. 6.4. Результаты однофакторного дисперсионного анализа

Means table (Таблица средних). Некоторые данные из этой таблицы были уже приведены в Summery Statistics. Колонка Stnd. error (pooled s) (Объединенная стандартная ошибка) содержит s_2 . В двух последних столбцах указанной таблицы находятся границы доверительных интервалов для средних из третьего столбца. Обратим внимание на то, что в таблице средних приведены доверительные 95%-ные интервалы, построенные по методике LSD (рис. 6.5). Щелчок правой кнопки мыши в поле

factor	Count	Mean	Stnd. error (pooled s)	Lower limit	Upper limit
5	12	6,675	0,186696	6,40539	6,94461
10	10	5,1	0,204515	4,80466	5,39534
15	11	7,73636	0,194998	7,45477	8,01796
Total	33	6,55152			

Рис. 6.5. Таблица средних

заставки Table of Means открывает следующее дополнительное меню (рис. 6.6), в котором задаются различные способы построения доверительных интервалов. В подразд. 4.7 рассмотрены формулы для построения стандартных доверительных интервалов (Confidence Interval). Сведения о других методах можно найти в [19].

Multiple Range Tests (Множественные сравнения) выдает результаты анализа множественных сравнений средних (рис. 6.7). В столбце Homogeneous

Groups (Однородные группы) вертикальными столбцами звездочек выделены возможные однородные группы наблюдений. В нашем случае таких групп три и каждая из них соответствует одному из трех различных уровней фактора. Таким образом, все группы неоднородны, и объединить их в одну общую группу нельзя.

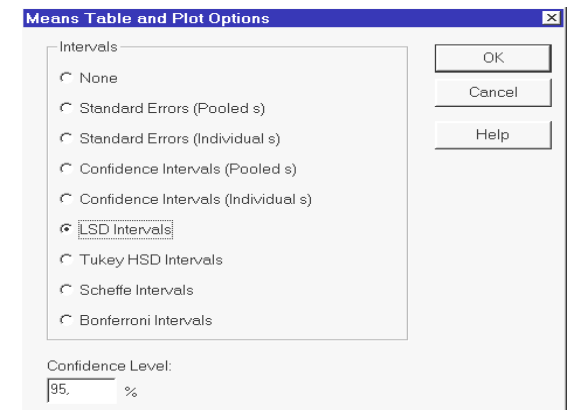


Рис. 6.6. Окно для задания различных способов построения доверительных интервалов

Multiple Range Tests for observ by factor

Method: 95,0 percent LSD			
factor	Count	Mean	Homogeneous Groups
10	10	5,1	X
5	12	6,675	X
15	11	7,73636	X
Contrast		Difference	+/- Limits
5 - 10		*1,575	0,565537
5 - 15		*-1,06136	0,551337
10 - 15		*-2,63636	0,577104

* denotes a statistically significant difference.

(*обозначено статистически значимое различие)

Рис. 6.7. Однородные группы наблюдений и контрасты

Далее в таблице приводятся значения линейных контрастов, вычисленных по формуле $L_k = \sum_{j=1}^k c_j T_j$. В нашем случае $H_0 : m_1 = m_2 = m_3$, где

m_i - средняя i -й подвыборки (уровня обработки). Тогда

$L_{k1} = m_1 - m_2$, $c_1 = 1$, $c_2 = -1$, $c_3 = 0$, $L_{k2} = m_1 - m_3$, $c_1 = 1$, $c_2 = 0$, $c_3 = -1$,

$L_{k3} = m_2 - m_3$, $c_1 = 0$, $c_2 = 1$, $c_3 = -1$, $\bar{x}_1 = 6.675$, $\bar{x}_2 = 5.1$, $\bar{x}_3 = 7.73636$.

Оценки и дисперсии линейных контрастов вычисляются по формуле (6.2.9): $\hat{L}_{k1} = c_1 \bar{x}_1 + c_2 \bar{x}_2 = 6.675 - 5.1 = 1.575$, $\hat{L}_{k2} = c_1 \bar{x}_1 + c_3 \bar{x}_3 =$

$= 6.675 - 7.73636 = -1.06136$, $\hat{L}_{k3} = c_2 \bar{x}_2 + c_3 \bar{x}_3 = 5.1 - 7.73636 = -2.63636$

и так далее. Наконец, в столбце под заголовком $+/- Limits$ приведены границы доверительного LSD интервала, для линейных контрастов, вычисленные по формуле, аналогичной формуле (6.2.10).

После щелчка правой кнопкой мыши в поле заставки Multiple Range Test появляется дополнительное меню, подобное меню в пункте Table of Means, в котором можно задать различные способы построения доверительных интервалов.

Variance Check (Тесты дисперсий). Эта процедура включает в себя результаты трех статистических критериев Кокрена*, Бартлетта и Хартли для сравнения разбросов наблюдений на разных уровнях фактора (рис. 6.8). Критерии Кокрена и Бартлетта проверяют на однородность ряд дисперсий, т. е. нулевую гипотезу вида $H_0 : D_1 = D_2 = \dots = D_k$. В данном случае D_i - дисперсия соответствующей подвыборки на i -м уровне фак-

* Уильям Геммел Кокрен (1909-1990) – английский математик.

тора. По этим двум критериям, кроме значений статистик критериев, приводятся также значения минимальных уровней значимости. Следует заметить, что критерии Кокрена и Бартлетта весьма чувствительны к отклонению модели наблюдений от нормальности, поэтому в интерпретации результатов этих критериев нужна определенная осторожность. Информацию о критериях Кокрена, Бартлетта и Хартли можно найти в [1, 2, 4, 8].

Variance Check		
Cochran's C test:	0,433479	P-Value = 0,601053
Bartlett's test:	1,03227	P-Value = 0,63386
Hartley's test:	1,84043	
Kruskal-Wallis Test for observ by factor		
factor	Sample Size	Average Rank
5	12	17,5833
10	10	5,95
15	11	26,4091
Test statistic = 23,5615 P-Value = 0,00000765048		

Рис. 6.8. Результаты тестов Кокрена, Бартлетта, Хартли и Краскела – Уоллиса

Kruskal – Wallis Tests (Ранговый однофакторный анализ Краскела – Уоллиса) исследует эффект действия одного фактора классификации для сбалансированного или несбалансированного плана.

В колонке factor стоят метки соответствующих способов обработки (факторов), в колонке Sample Size (Размер выборки) – число наблюдений на каждом уровне фактора. В колонке Average Rank (Средний ранг) – соответствующая величина ранга для каждой группы. Под таблицей приведены значения для асимптотической аппроксимации, скорректированной для случая совпадающих наблюдений по формуле (6.4.4), и минимальный уровень значимости этой статистики (p-Value).

Перечисленные выше процедуры довольно слабо затрагивают вопрос о правомерности применения дисперсионного анализа к анализируемым данным. Этот вопрос является определяющим и от него зависит достоверность выводов, полученных в результате анализа. Для более детального рассмотрения исходной выборки в пакете STATGRAPHICS могут быть применены критерии χ^2 и Колмогорова для проверки согласия с нормальным распределением, глазомерный метод проверки нормальности, критерии асимметрии и эксцесса.

Рассмотрим теперь процедуры окна Graphics Options дополнительного меню. В разделе One-Way ANOVA можно строить следующие графики (рис. 6.9). Отметим все пункты за исключением третьего. В результате получим следующие графики:

Scatterplot – это диаграмма рассеивания исходной выборки. Мы имели с ней дело постоянно, начиная с лабораторной работы № 1.

Means Plot реализует графическое представление данных таблицы, выдаваемой процедурой Table of Means (рис. 6.10).

Процедуры Residuals versus Factor Levels, Residuals versus Predicted и Residuals versus Row Number дают графики остатков в одной из трех возможных форм: в зависимости от уровня фактора, в зависимости от предсказанных значений или в зависимости от номера наблюдения в векторе ввода данных

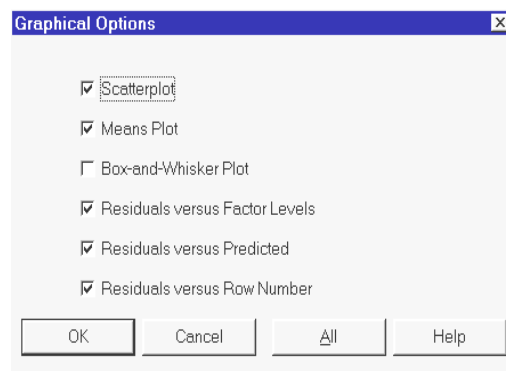


Рис. 6.9. Панель графических процедур однофакторного дисперсионного анализа

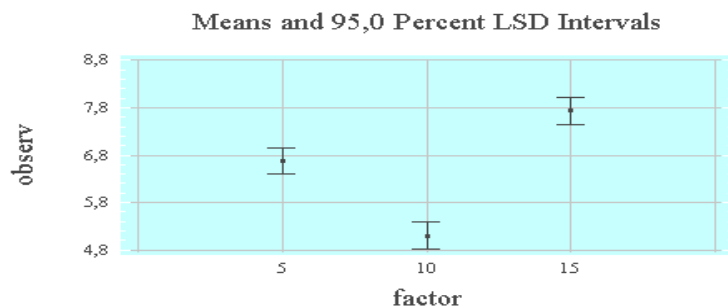
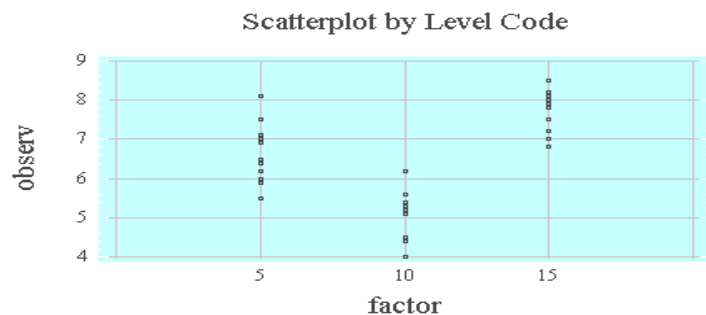


Рис. 6.10. Диаграмма рассеивания выборки и доверительные интервалы для средних по факторам

(рис. 6.11). Каждая из этих форм подчеркивает свой аспект в возможных причинах нарушения однородности распределения остатков.

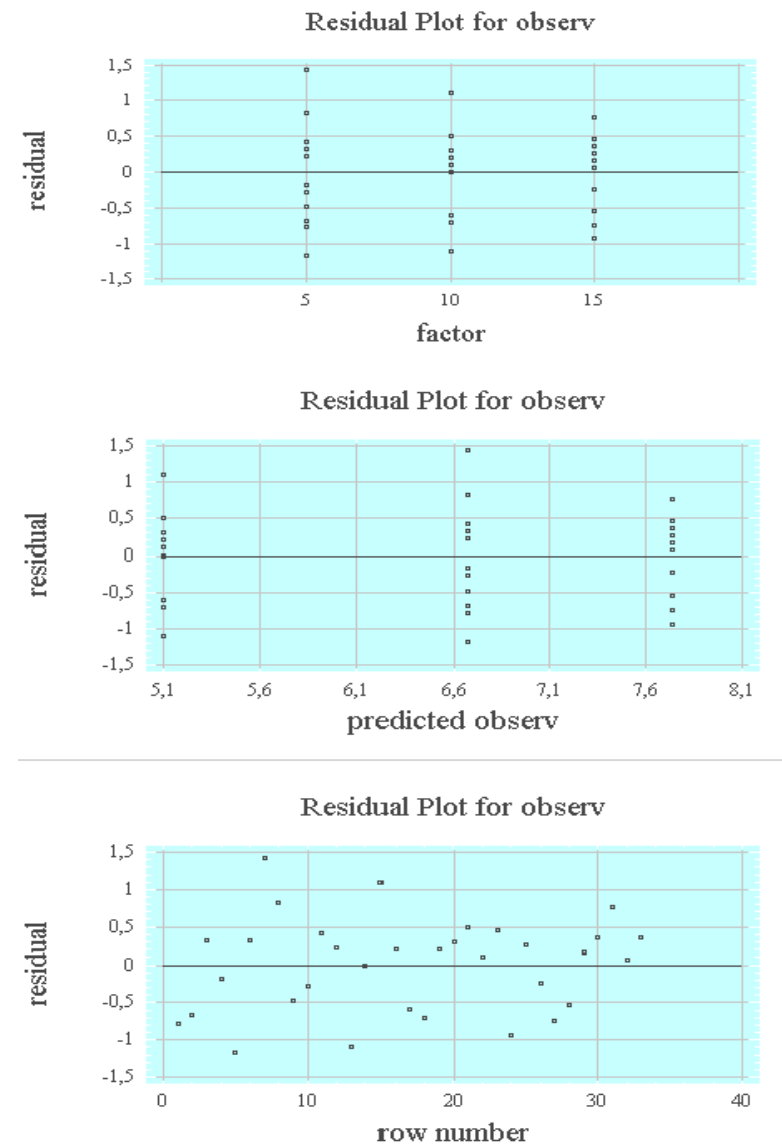


Рис. 6.11. Графики остатков

Задание № 1. С помощью рассмотренных процедур пакета STAT-GRAPHICS решить одну задачу однофакторного дисперсионного анализа. Везде уровень значимости принять равным 0.05. В каждой задаче проверить гипотезу H_0 о равенстве средних. Если гипотеза H_0 принимается, то найти несмещенные оценки среднего и дисперсии. Если же H_0 отклоняется, провести попарное сравнение средних, используя метод линейных контрастов.

1. В трех магазинах, продающих товары одного вида, данные товарооборота за восемь месяцев работы (в тыс. руб.) составили следующую сводку:

Мага- зин	Месяц							
	1	2	3	4	5	6	7	8
I	19	23	26	18	20	20	18	35
II	20	20	32	27	40	23	22	18
III	16	15	18	26	19	17	19	18

2. В следующей таблице приведены результаты обследования 60 работников производства, у которых фиксировалась средняя часовая выработка в натуральных единицах продукции. Принять за фактор – стаж работы.

Стаж	Возраст		
	от 25 до 35 лет	от 35 до 45 лет	от 45 до 55 лет
От 1 до 4 лет	19, 20, 20, 20, 22,	19, 20, 20, 23, 25,	18, 19, 20, 21, 23,
От 4 до 7 лет	30, 31, 32, 32, 34,	20, 29, 30, 31, 31,	19, 25, 25, 26, 26,
От 7 до 10 лет	35, 35, 39, 40, 41,	36, 40, 41, 42, 45,	24, 24, 24, 25, 25,
Свыше 10 лет	40, 40, 41, 41, 42,	28, 31, 35, 36, 40,	20, 24, 25, 31, 32.

3. Решить задачу № 2 с теми же данными, приняв за фактор, влияющий на среднюю часовую выработку, возраст работника.

4.

Номер выборки	Наблюдения								
	1	2	3	4	5	6	7	8	9
1	12	4	7	8	5	9	6	-	-
2	14	11	5	6	3	-	-	-	-
3	6	5	12	9	10	7	11	4	5

5.

Номер выборки	Наблюдения					
	1	2	3	4	5	6
1	4	2	3	4	5	3
2	6	5	4	7	6	8
3	8	9	10	7	8	6

6.

Номер выборки	Наблюдения							
	1	2	3	4	5	6	7	8
1	9	8	8	7	9	-	-	-
2	8	11	8	9	10	12	-	-
3	9	10	7	11	8	10	12	13
4	16	9	12	14	15	17	19	-

7. Приведены данные о содержании иммуноглобулина IgA в сыворотке крови (в мг %) у больных пяти возрастных групп:

Возрастная группа	Содержание IgA (мг %)										
1	83	85	-	-	-	-	-	-	-	-	-
2	84	85	85	86	86	87	-	-	-	-	-
3	86	87	87	87	88	88	88	88	88	89	90
4	89	90	90	91	-	-	-	-	-	-	-
5	90	92	-	-	-	-	-	-	-	-	-

8. На химическом заводе разработаны два новых варианта технологического процесса. Чтобы оценить, как изменится дневная производительность при переходе на работу по новым вариантам технологического процесса, завод в течение десяти дней работает по каждому варианту, включая существующий. Дневная производительность завода (в условных единицах) приводится в таблице:

Технологический процесс	Суточная производительность									
	1	2	3	4	5	6	7	8	9	10
Существующая схема	46	48	73	52	72	44	66	46	60	48
Вариант I	74	82	64	72	84	68	76	88	70	60
Вариант II	52	63	72	64	48	70	78	68	79	54

9. Из большой группы полевых транзисторов с недельным интервалом были получены три выборки. Ниже приводятся результаты измерения емкости затвора-стока у этих транзисторов (в пикофарадах):

Номер выборки	Емкость (пФ)															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	2.8	3.2	2.9	3.5	3.3	3.7	3.9	3.1	3.2	3.1	3.4	3.0	3.6	3.1	3.2	3.2
2	3.1	3.2	3.3	3.4	3.7	3.4	3.0	3.1	2.9	3.5	3.2	3.2	-	-	-	-
3	3.6	2.8	3.0	3.2	3.0	3.7	3.2	3.2	3.6	3.4	3.1	3.2	-	-	-	-

10. Выяснить зависит ли объем работ, выполненных на стройке за смену, от работающей бригады. Данные по четырем бригадам приведены в следующей таблице:

Номер бригады	Объем выполненной работы					
1	140	144	142	145	146	140
2	150	149	152	150	-	-
3	150	149	146	147	148	150
4	150	155	154	152	157	-

11. Приведены два последних десятичных знака константы в эксперименте по определению гравитационной постоянной G . Например, табличное значение 83 соответствует наблюдаемому значению 6.683. Эксперимент ставился с шарами, сделанными из золота, платины и стекла.

Материал	Значение константы					
Золото	83	81	76	78	79	72
Платина	61	61	67	67	64	-
Стекло	78	71	75	72	74	-

12.

Номер выборки	Наблюдения									
1	92	78	60	67	53	66	-	-	-	-
2	83	96	98	60	99	78	77	103	93	107
3	66	97	100	96	96	-	-	-	-	-

13. Представлены пробы долговечности электрических ламп, взятых из четырех партий.

Номер партии	Продолжительность горения в часах							
1	1600	1610	1650	1680	1700	1700	1800	-
2	1580	1640	1640	1700	1750	-	-	-
3	1460	1550	1600	1620	1640	1660	1740	1820
4	1510	1520	1530	1570	1600	1680	-	-

14. Приведены изменения критерия чистоты поверхности металла для трех приборов.

Номер прибора	Отклонения от общей медианы в сотых долях микрона					
1	-4	-2	-21	-4	-4	-35
2	7	11	30	28	27	103
3	19	2	-13	-9	2	1

15. Результаты 22 испытаний на четырех уровнях фактора следующие:

Уровень фактора	Наблюдения						
F_1	1.38	1.45	1.38	1.42	1.42	1.44	1.39
F_2	1.41	1.42	1.44	1.45	1.46	1.43	-
F_3	1.32	1.33	1.34	1.31	1.35	-	-
F_4	1.31	1.33	1.32	1.33	-	-	-

16. Проведено 22 испытания, результаты которых представлены в таблице.

Уро- вень фак- тора	Наблюдения								
F_1	30.56	32.66	34.78	35.50	36.63	40.20	42.28	41.76	35.17
F_2	43.44	47.51	53.80	50.11	46.23	51.19	-	-	-
F_3	31.36	36.20	36.38	42.20	35.13	39.93	34.72	-	-

17. Результаты испытаний на трех уровнях фактора следующие:

Уро- вень фак- тора	Наблюдения											
F_1	37	47	40	60-	52	48	42	-	-	-	-	-
F_2	60	86	67	92	90	95	98	103	89	91	95	97
F_3	69	100	98	75	85	101	94	73	89	96	-	-

18. В следующей таблице приведены уровни поставок сырья (в условных единицах) в серии из пяти партий.

Пар- тии	Уровень поставок сырья																			
1	62	66	64	64	63	62	64	64	66	64	66	63	65	63	63	63	61	56	64	65
2	66	65	65	66	67	66	69	70	68	69	63	65	64	65	64	-	-	-	-	-
3	62	64	62	62	65	64	65	62	62	63	64	-	-	-	-	-	-	-	-	-
4	65	64	63	62	65	63	64	63	-	-	-	-	-	-	-	-	-	-	-	-
5	65	64	67	62	65	62	64	64	64	65	-	-	-	-	-	-	-	-	-	-

19. Таблица данных содержит результаты по определению октанового числа бензина, полученные в четырех округах на северо-востоке США летом 1953 года.

Ок- руг	Октановое число бензина												
A	84.0	83.5	84.0	85.0	83.1	83.5	81.7	85.4	84.1	83.0	85.8	84.0	84.2
	82.2	83.6	84.9	-	-	-	-	-	-	-	-	-	-
B	82.4	82.4	83.4	83.3	83.1	83.3	82.4	83.3	82.6	82.0	83.2	83.1	82.5
C	83.2	82.8	83.4	80.2	82.7	83.0	85.0	83.0	85.0	83.7	83.6	83.3	83.8
	85.1	83.1	84.2	80.6	82.3	-	-	-	-	-	-	-	-
D	80.2	82.9	84.6	84.2	82.8	83.0	82.9	83.4	83.1	83.5	83.6	86.7	82.6
	82.4	83.4	82.7	82.9	83.7	81.5	81.9	81.7	82.5	-	-	-	-

20. Приведены две последние цифры чисел, выражающих скорость света, полученные Майкельсоном* в его опыте с шестью круговыми зеркалами.

Но- мер зер- ка- ла	Наблюдения																		
1	47	47	38	62	29	59	92	44	41	47	44	41	-	-	-	-	-	-	-
2	42	18	36	45	33	30	0	27	18	27	57	66	48	24	15	-	-	-	-
3	3	39	27	67	48	15	3	7	27	27	42	37	69	24	63	15	30	27	42
4	6	21	27	33	9	24	6	39	42	18	12	63	-	-	-	-	-	-	-
5	18	9	12	30	30	27	30	39	18	27	48	24	18	-	-	-	-	-	-
6	30	21	33	18	12	33	24	23	57	39	44	33	30	24	24	30	-	-	-

21. Фруктовый сок хранился в течение нескольких месяцев в цистернах четырех типов, после чего определялось его качество выставлением численной оценки. Ниже приведены результаты испытаний.

Цистерна	Наблюдения							
A	6.14	5.72	6.90	5.80	6.23	6.06	5.42	6.04
B	6.55	6.29	7.40	6.40	6.28	6.26	6.22	6.76
C	5.54	5.61	6.60	5.70	5.31	5.58	5.57	5.84
D	4.81	5.09	6.61	5.03	5.15	5.05	5.77	6.17

22. Лечащий врач рекомендовал своим пациентам, жалующимся на лишний вес, лекарства A, B и C. При этом он каждый раз фиксировал вес

* А.А. Майкельсон (1852-1931) – американский физик.

пациента после лечения в фунтах (1 фунт = 453.6 г), в результате чего получены следующие результаты.

Лекарство	Вес пациента											
A	147	183.5	150	167	180	216.5	127.5	222	132	167	221	203
B	180	161.5	157	155	146	131.5	163.3	160	162	225	159	-
C	216	172	140	154	161	-	-	-	-	-	-	-

23. Следующая таблица содержит специальные оценки в баллах, соответствующие одному из четырех экспериментальных условий.

Условие	Оценки								
1	0	1	3	3	5	10	13	17	26
2	0	6	7	9	11	13	20	20	24
3	0	5	8	9	11	13	16	17	20
4	1	5	12	13	19	22	25	27	29

24. Приведено содержание влаги (в %) в образцах некоторого продукта в зависимости от условий хранения.

Условия хранения	Содержание влаги (в %)														
1	7.8	7.7	7.4	7.9	8.3	8.2	8.0	7.6	7.4	7.7	8.4	8.3	8.5	8.3	8.2
2	5.4	5.3	5.2	5.5	5.6	7.4	7.3	7.5	7.1	7.0	6.9	-	-	-	-
3	8.1	8.0	7.9	8.2	8.3	6.4	6.3	6.5	-	-	-	-	-	-	-
4	7.9	8.0	7.8	8.1	7.9	9.5	9.6	9.4	10.1	10	9.9	-	-	-	-
5	7.1	6.9	7.0	7.3	7.2	-	-	-	-	-	-	-	-	-	-

25. В таблице дано среднее число ошибок при выполнении 12 различных заданий животными трех видов.

Животное	Среднее число ошибок											
Крысы	1.5	1.1	1.8	1.9	4.3	2.0	8.4	6.6	2.4	6.5	2.6	6.5
Кролики	1.7	1.5	8.1	1.3	4.0	4.6	4.0	5.1	2.5	6.9	2.5	6.8
Кошки	0.3	1.0	3.6	0.0	0.6	5.5	1.0	3.1	0.1	1.6	4.3	1.0

26. Приведены результаты исследования дрожания мышц рук (тремор) у шести пациентов в зависимости от веса браслета. Каждое табличное значение – среднее из пяти экспериментальных измерений частоты тремора (в Гц).

Пациент	Частота тремора (в Гц)							
1	2.58	2.63	2.62	2.59	2.85	3.01	2.96	2.78
2	2.70	2.83	3.15	3.43	3.47	-	-	-
3	2.78	2.71	3.02	2.89	3.14	3.01	3.35	-
4	2.36	2.49	2.58	2.86	2.93	3.10	-	-
5	2.67	2.96	3.02	3.08	3.32	3.41	-	-
6	2.43	2.50	2.85	3.06	3.07	-	-	-

27. В следующей таблице приведено количество решенных задач в шести однородных группах из пяти человек. Задачи предлагались каждому испытуемому независимо от всех остальных. Группы отличаются между собой величиной денежного вознаграждения за решаемую задачу.

Группа	Число решенных задач				
1	10	11	9	13	7
2	8	10	16	13	12
3	12	17	14	9	16
4	12	15	16	16	19
5	24	16	22	18	20
6	19	18	27	25	24

28. Приведено количество металлических заготовок определенных формы и размера, изготовленных рабочими трех разных групп, отличающихся различными представлениями о цели работы (I- отсутствие информации, II- общие представления, III- точная информация).

Информация о цели работы	Число обработанных заготовок					
I	40	35	38	43	44	41
II	38	40	47	44	40	42
III	48	40	45	43	46	44

29. Данные таблицы представляют разрывную прочность волокон хлопка (в условных единицах) в зависимости от уровня калийных удобрений, вносимых в почву.

Уровень удобрений	Прочность волокон							
I	7.46	7.68	7.21	7.17	7.57	7.80	7.87	7.34
II	7.76	7.73	7.74	8.14	8.15	7.87	-	-
III	7.62	8.00	7.93	7.54	8.11	-	-	-

30. Исследовалось влияние метронома на плавность (количество ошибок) речи за определенный отрезок времени при следующих условиях: N- испытуемый говорил без помощи метронома, R- испытуемый говорил при ритмичной работе метронома, А- испытуемый говорил под неритмичный метроном. Полученные данные приведены в таблице.

Условия	Количество ошибок в речи											
N	5	3	3	4	2	2	2	3	2	0	4	1
R	3	3	1	5	2	0	0	0	0	1	2	2
A	15	11	18	21	6	17	10	8	13	4	11	17

7. РЕГРЕССИОННЫЙ АНАЛИЗ

7.1. Модели регрессии

Одной из важнейших задач математической статистики является задача о нахождении связи между двумя случайными величинами X и Y . Во многих случаях одна из двух величин может быть и неслучайной. Предположим, что функциональная зависимость между переменными, называемая моделью, известна из предварительных сведений с точностью до параметров $\theta_1, \theta_2, \dots, \theta_k$ и имеет вид

$$y_i = f(x_i, \theta_1, \theta_2, \dots, \theta_k), \quad i = \overline{1, n}. \quad (7.1.1)$$

Требуется по результатам наблюдений (x_i, y_i) , $i = 1, 2, \dots, n$, найти оценки неизвестных параметров $\theta_1, \theta_2, \dots, \theta_k$. Очень часто задача ставится еще проще. Модель в этом случае имеет линейный вид, т.е.

$$y_i = \theta_0 + \theta_1 x_{1,i} + \theta_2 x_{2,i} + \dots + \theta_k x_{k,i}, \quad (7.1.2)$$

где x_i - неслучайные аргументы, а y - случайная величина. Таким образом, здесь аргументы x_i определяют y лишь в среднем, оставляя просторы для случайных колебаний.

Ситуация, в которой экспериментатор может выбирать значения аргументов x_i по своему желанию и таким образом планировать будущие эксперименты, называется активным экспериментом. В этом случае значения аргументов x_i обычно рассматриваются как неслучайные. В отличие от этой ситуации в пассивном эксперименте значения переменных x_i складываются вне воли экспериментатора, под действием других обстоятельств. Поэтому значения x_i приходится толковать как случайные величины, что накладывает особые черты на интерпретацию результатов.

Итак, в регрессионном анализе предполагается, что можно прямо или косвенно контролировать одну или несколько независимых переменных x_1, x_2, \dots, x_n , и их значения вместе с множеством параметров $\theta_1, \theta_2, \dots, \theta_k$ определяют математическое ожидание зависимой переменной Y . Задача состоит в вычислении оценок параметров с помощью выборочных данных.

Возникает вопрос, почему представляет интерес регрессия? Очень часто применение регрессии связано с необходимостью оценить (или предсказать) среднее значение y при конкретных значениях переменных x_i . Иногда требуется установить определенную функциональную связь между x_i и математическим ожиданием Y . В общем случае какая-нибудь форма функциональной связи является полезным источником информации о зависимости переменной Y от x_i .

При попытках аппроксимировать данные кривой или поверхностью сначала предполагается существование функциональной зависимости определенного вида. С помощью данных и соответствующих математических вычислений находят оценки параметров, дающие наилучшее приближение согласно какому-либо критерию. Можно выяснить, насколько хороша данная зависимость, но не исключено, что удастся получить лучшую, выбрав другую функцию и другой критерий.

Здесь стоит подчеркнуть одно существенное обстоятельство. Имея в своем распоряжении мощный компьютер, сравнительно легко перебрать большое количество разных функций, аппроксимирующих данные. Это сильное искушение, так как можно без конца перебирать комбинации и преобразования данных, надеясь получить идеальный вариант. Совершенно неправильно считать, что найденное уравнение будет наилучшим только потому, что оно дает хорошее приближение, если оно несколько не соответствует реальным физическим или техническим связям. В любой регрессионной задаче в первую очередь следует рассматривать физически обоснованную конкретную функциональную форму независимо от того, была ли она получена с помощью аналитических выводов или благодаря какому-нибудь иному предварительному знанию свойств переменных. Вполне возможно, что для аппроксимации этой функции понадобятся другие функциональные связи.

В последнее время регрессионный анализ – очень бурно развивающаяся отрасль вычислительной математики. Благодаря ему возникло целое направление, связанное с решением плохо обусловленных задач. Появилось огромное число подходов, алгоритмов и программ, позволяющих в этих нелегких условиях более или менее рационально организовывать вычислительные процедуры.

При оценивании параметров регрессий приходится прибегать к поисковым методам, имеющим итеративный характер. Для их реализации написаны многочисленные программы, развитие которых вылилось в метод всех возможных регрессий, а затем в шаговый регрессионный анализ. При этом необходимо отметить несколько тенденций, определяющих методы и темпы развития регрессионного анализа.

Первая тенденция заключается в пересмотре довольно жестких базовых предпосылок классического регрессионного анализа. Это касается таких предположений, как нормальность распределения ошибок, однородность, независимость и т.п. Отказ хотя бы от одного из перечисленных предположений фактически приводит к созданию новой модели.

Вторая тенденция состоит в вовлечении в регрессионный анализ более тонких математических методов, таких как функциональный анализ, теория групп, обобщение регрессионной задачи на бесконечномерные пространства.

Третья тенденция – обращение ко все более сложным объектам исследования. Речь может идти о моделях в форме обыкновенных дифференциальных уравнений, интегро-дифференциальных уравнений, уравнений математической физики.

Наконец, четвертая тенденция – одновременный выбор модели и метода оценивания, итеративная обработка результатов и адаптация модели и метода оценивания друг к другу.

Рассмотрим сначала простейшую регрессионную задачу: построим уравнение линейной регрессии в рамках гауссовской модели наблюдений.

Пусть имеется n парных наблюдений $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, причем примем, что переменная x – регрессор – неслучайна и измеряется без ошибок. Если при этом есть признаки связи между наблюдениями, то обычно исследователь спешит построить некоторую кривую, чаще всего прямую линию, связывающую все эти наблюдения. Для нахождения параметров уравнения регрессии обычно используется метод наименьших квадратов или метод максимального правдоподобия. Метод наименьших квадратов при оценке параметров регрессии не требует никаких предположений о нормальности распределения ошибок, но они становятся необходимыми при построении доверительных интервалов и для проверки гипотез о значениях параметров уравнения регрессии.

Рассмотрим одномерную линейную модель вида

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = \overline{1, n}, \quad (7.1.3)$$

где ε_i – ошибки измерений переменной y предполагаются независимыми случайными величинами, распределенными нормально: $\varepsilon_i \in N(0, D_\varepsilon)$.

Наша задача состоит в том, чтобы по наблюдениям найти оценки $a = \hat{a}$, $b = \hat{b}$ и $s^2 = \hat{D}$ для параметров α, β и D соответственно.

Перечислим еще раз все явные и неявные предположения, принимаемые в рамках модели наблюдений. От их выполнения зависит качество получаемых оценок и возможность применения к ним процедур статистического анализа.

1. Значения x задаются или измеряются без ошибок.
2. Регрессия Y на X линейна, т.е. $M(Y/x) = \alpha + \beta x$.
3. Отклонения $y_i - M(Y/x_i)$ взаимно независимы.
4. Эти отклонения имеют одну и ту же дисперсию D , точное значение которой неизвестно, при всех x . Это свойство называется гомоскедастичностью, а сами дисперсии – гомоскедастичными.
5. Отклонения распределены по нормальному закону.
6. Данные действительно были взяты из совокупности, относительно которой должны быть сделаны выводы.
7. Не было посторонних переменных, существенно уменьшающих значения связи между X и Y .

Полезно отметить последствия невыполнения некоторых предположений. Невыполнение третьего предположения может существенно повлиять на характеристики применяемых статистических методов из-за не учета зависимости между переменными, представляющими измерения над разными объектами. Хотя отклонения от нормальности встречаются довольно часто, они имеют значение, только если очень значительны. Отсутствие гомоскедастичности приводит к тому, что метод наименьших квадратов не гарантирует минимальных дисперсий оценок. Невыполнение последних двух предположений также имеет принципиальное значение. Если они нарушены, полезность проведенного исследования незначительна.

7.2. Оценка параметров линейной регрессии методом наименьших квадратов

Перепишем уравнение регрессии в несколько ином виде

$$y = \alpha + \beta(x - \bar{x}), \quad (7.2.1)$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Эта прямая называется теоретической линией регрессии

или прямой отклика. Уравнение

$$\hat{y} = a + b(x - \bar{x}) \quad (7.2.2)$$

определяет кривую, которая является оценкой для прямой регрессии.

Суть метода наименьших квадратов состоит в выборе таких оценок a и b , которые бы минимизировали сумму квадратов отклонений наблюдаемых значений y_i от прогнозируемых величин \hat{y}_i , полученных подстановкой значений x_i в уравнение (7.2.2), т.е.

$$R = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - a - b(x_i - \bar{x})]^2 \Rightarrow \min. \text{ Чтобы найти значения } a$$

и b , минимизирующие R , продифференцируем это уравнение по a и b и приравняем производные нулю:

$$\begin{cases} \frac{\partial R}{\partial a} = -2 \sum_{i=1}^n [y_i - a - b(x_i - \bar{x})] = 0, \\ \frac{\partial R}{\partial b} = -2 \sum_{i=1}^n [y_i - a - b(x_i - \bar{x})](x_i - \bar{x}) = 0. \end{cases}$$

Раскроем здесь члены под знаком суммы: $\sum_{i=1}^n y_i - an - b \sum_{i=1}^n (x_i - \bar{x}) = 0$,

$$\sum_{i=1}^n y_i (x_i - \bar{x}) - a \sum_{i=1}^n (x_i - \bar{x}) - b \sum_{i=1}^n (x_i - \bar{x})^2 = 0. \text{ Но } \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = n\bar{x} - n\bar{x} = 0. \text{ Тогда } na = \sum_{i=1}^n y_i, \quad b \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x}) y_i. \text{ Отсюда легко}$$

получить оценки параметров a и b :

$$a = \hat{a} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}, \quad b = \hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (7.2.3)$$

Вторую оценку часто видоизменяют и переписывают в следующем виде $\sum_{i=1}^n (x_i - \bar{x}) y_i = \sum_{i=1}^n (x_i - \bar{x}) y_i + \bar{y} \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x}) y_i + \sum_{i=1}^n (x_i - \bar{x}) \bar{y} = \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})$. Тогда

$$b = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (7.2.4)$$

Рассмотрим теперь свойства полученных оценок. Они являются несмещенными, состоятельными и эффективными в классе линейных (относительно наблюдений) оценок. Действительно,

$$\begin{aligned}
 M(a) &= M\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n} \sum_{i=1}^n M(y_i) = \frac{1}{n} \sum_{i=1}^n M[\alpha + \beta(x_i - \bar{x}) + \varepsilon_i] = \\
 &= \frac{1}{n} \left[\alpha n + \beta \sum_{i=1}^n (x_i - \bar{x}) \right] = \alpha, \\
 M(b) &= M\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x}) M(y_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n [(x_i - \bar{x})(\alpha + \beta(x_i - \bar{x}))]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \\
 &= \alpha \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta.
 \end{aligned}$$

Здесь учтено, что переменные x_i - не-

случайные, а y_i - случайные величины. Кроме того, математическое ожидание y_i есть теоретическая линия регрессии (7.2.1).

Найдем теперь дисперсии оценок a и b в предположении, что наблюдения y_i независимы и нормально распределены, причем $D(y_i) = D = \sigma^2$ (предположения 3, 4 и 5 предыдущего подраздела).

Имеем: $D(a) = D\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(y_i) = \frac{D}{n^2} n = \frac{D}{n},$

$$\begin{aligned}
 D(b) &= D\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} \sum_{i=1}^n (x_i - \bar{x})^2 D(y_i) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 D}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} = \\
 &= \frac{D}{\sum_{i=1}^n (x_i - \bar{x})^2}.
 \end{aligned}$$

Состоятельность оценок a и b немедленно следует после применения к ним неравенства Чебышева. Например, для оценки a получим

$$P(|a - \alpha| \geq \varepsilon) \leq \frac{D(a)}{\varepsilon^2} = \frac{D}{n\varepsilon^2}. \text{ Отсюда } \lim_{n \rightarrow \infty} P(|a - \alpha| \geq \varepsilon) = 0.$$

В общем случае доказательство того, что метод наименьших квадратов дает оценки с наименьшей дисперсией в классе всех несмещенных оценок, довольно сложно. Приведем его для оценки b параметра β . Предположим, что существует еще одна линейная оценка b' параметра β ,

отличная от оценки b и пусть, например, $b' = \sum_{i=1}^n c_i y_i$. Очевидно, что

$$M(b') = \sum_{i=1}^n c_i M(y_i) = \sum_{i=1}^n c_i [\alpha + \beta(x_i - \bar{x})] = \alpha \sum_{i=1}^n c_i + \beta \sum_{i=1}^n (x_i - \bar{x}) c_i. \text{ Оценка } b'$$

будет несмещенной, если $M(b') = \beta$, т.е.

$$\begin{cases} \sum_{i=1}^n c_i = 0, \\ \sum_{i=1}^n (x_i - \bar{x}) c_i = 1. \end{cases} \quad (7.2.5)$$

$$\text{В этих условиях } D(b') = \sum_{i=1}^n c_i^2 D(y_i) = D \sum_{i=1}^n c_i^2 =$$

$$\begin{aligned} &= D \sum_{i=1}^n \left[c_i - \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 = D \left[\sum_{i=1}^n c_i^2 + \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} - \right. \\ &\quad \left. - 2 \sum_{i=1}^n c_i \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} + \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} + 2 \sum_{i=1}^n c_i \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} - 2 \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \right] = \\ &= \sum_{i=1}^n \left[c_i - \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 + \sum_{i=1}^n \left[c_i - \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} + \sum_{i=1}^n c_i \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned}$$

Но $\sum_{i=1}^n (x_i - \bar{x})^2$ - это константа, т.е. выражение под этой суммой уже не

зависит от индекса внешнего суммирования. Тогда

$$\sum_{i=1}^n c_i \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n c_i (x_i - \bar{x}) = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ с учетом условий}$$

(7.2.5). Аналогично

$$\sum_{i=1}^n \left[c_i - \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \sum_{i=1}^n \left[c_i (x_i - \bar{x}) \sum_{i=1}^n (x_i - \bar{x})^2 - (x_i - \bar{x})^2 \right] =$$

$$= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \left[\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n c_i (x_i - \bar{x}) - \sum_{i=1}^n (x_i - \bar{x})^2 \right] =$$

$$= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \left[\sum_{i=1}^n (x_i - \bar{x})^2 - \sum_{i=1}^n (x_i - \bar{x})^2 \right] = 0. \quad \text{Поэтому}$$

$$D(b') = D \sum_{i=1}^n \left[c_i - \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 + \frac{D}{\sum_{i=1}^n (x_i - \bar{x})^2}. \text{ Последний член в получен-}$$

ном выражении является константой. Следовательно, минимизировать $D(b')$ можно только за счет уменьшения первого члена. Полагая

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ мы обратим первый член в нуль (меньше он не может}$$

быть) и тем самым минимизируем $D(b')$. Но если в формулу $b' = \sum_{i=1}^n c_i y_i$

подставить значения c_i , при которых $D(b')$ минимальна, то альтернатив-

ная оценка b' примет вид $b' = \sum_{i=1}^n c_i y_i = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$, что совпадает с

оценкой наименьших квадратов. Поэтому b - линейная несмещенная оценка параметра β с минимальной дисперсией.

7.3. Интервальные оценки параметров линейной регрессии и кривой регрессии

Построим теперь доверительные границы для параметров α и β и кривой регрессии. Так как $\hat{y} = a + b(x - \bar{x})$ и $D(a) = \frac{D}{n}$, $D(b) = \frac{D}{\sum_{i=1}^n (x_i - \bar{x})^2}$,

то $M(\hat{y}) = M[a + b(x - \bar{x})] = M(a) + (x - \bar{x})M(b) = \alpha + \beta(x - \bar{x}) = y$,

$$D(\hat{y}) = D[a + b(x - \bar{x})] = D(a) + (x - \bar{x})^2 D(b) = \frac{D}{n} + \frac{(x - \bar{x})^2 D}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$

$$= D \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] - \text{выражение для дисперсии } D(\hat{y}) \text{ в текущей точке } x.$$

Очевидно, что \hat{y} - кроме того линейная функция от оценок a и b , которые в свою очередь являются линейными оценками от нормально распределенных наблюдений y_i . Следовательно, \hat{y} - нормально распределенная случайная величина, и для нее может быть построен доверительный интервал стандартным образом. То же можно сказать и об оценках коэффициентов регрессии.

Заметим, что a и b независимы друг от друга, так же как независима от них оценка \hat{D} дисперсии D . Это можно доказать, рассмотрев, например, $M(a \cdot b)$. После непродолжительных вычислений будет видно, что

$M(a \cdot b) = K(a, b) = 0$. Следовательно a и b - некоррелированы, а поскольку мы остаемся в рамках гауссовской модели, то и независимы.

В предыдущих разделах было показано, что дробь $n\hat{D}/D \in \chi_{n-1}^2$, $\hat{D} = D^*$. В нашем случае $\hat{D} = D^* = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - a - b(\bar{x}))^2$. Так как на случайные величины y_i , входящие в

эту формулу, наложены два условия связи вида $\frac{\partial R}{\partial a} = 0$ и $\frac{\partial R}{\partial b} = 0$, то число степеней свободы уменьшается на число связей и $n\hat{D}/D \in \chi_{n-2}^2$.

Составим дроби Стьюдента для a и b . В нашем случае

$$a \in N\left(\alpha, \frac{D}{n}\right), b \in N\left(\beta, \frac{D}{\sum_{i=1}^n (x_i - \bar{x})^2}\right), \text{ а по теории } t = \frac{z\sqrt{n}}{\sqrt{v}}, \text{ где}$$

$z \in N(0,1)$, $v \in \chi_n^2$, причем в этой дроби под корнем в числителе стоит число степеней свободы случайной величины v . Выберем в качестве стандартной нормальной случайной величины z сначала выражение

$$\frac{a - \alpha}{\sqrt{D/n}} = \frac{(a - \alpha)\sqrt{n}}{\sigma} \in N(0,1), \text{ затем } \frac{b - \beta}{\sigma} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \in N(0,1). \text{ Подставляя}$$

эти результаты в дробь Стьюдента, будем иметь

$$t_a = \frac{(a - \alpha)\sqrt{n}\sqrt{n-2}}{\sqrt{D}(\sqrt{n\hat{D}}/\sqrt{D})} = \frac{(a - \alpha)\sqrt{n-2}}{\sqrt{\hat{D}}} \in t_{n-2}. \quad \text{Аналогично}$$

$$t_b = \frac{(b - \beta)\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{n-2}}{\sqrt{D}\sqrt{n\hat{D}/D}} = \frac{(b - \beta)\sqrt{(n-2)\sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{n\hat{D}}} \in t_{n-2}. \quad \text{Нако-}$$

нец, получим в явном виде доверительные интервалы для коэффициентов линейной регрессии. $P(|a - \alpha| < \varepsilon) = \beta'$ по определению, где β' - довери-

$$\text{тельная вероятность. } P\left(\frac{|a - \alpha|\sqrt{n-2}}{\sqrt{\hat{D}}} < \frac{\varepsilon\sqrt{n-2}}{\sqrt{\hat{D}}}\right) = P\left(\left|\frac{(a - \alpha)\sqrt{n-2}}{\sqrt{\hat{D}}}\right| < t_{\beta'}\right) =$$

$= P(|t| < t_{\beta'}) = \beta'$, величина $t_{\beta'}$ может быть найдена из уравнения

$$2 \int_0^{t_{\beta'}} s_{n-2}(t) dt = \beta'. \quad \text{Тогда} \quad \varepsilon = t_{\beta', n-2} \sqrt{\frac{\hat{D}}{n-2}} \quad \text{и}$$

$$I_{\alpha} = \left(a - t_{\beta', n-2} \sqrt{\frac{\hat{D}}{n-2}}, a + t_{\beta', n-2} \sqrt{\frac{\hat{D}}{n-2}} \right).$$

Точно такие же преобразования дают интервал для второго коэффициента.

$$P(|b - \beta| < \varepsilon) = \beta', \quad \text{тогда} \quad P\left(\left| \frac{(b - \beta) \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{n\hat{D}/(n-2)}} \right| < \frac{\varepsilon \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{n\hat{D}/(n-2)}} \right) =$$

$$= P(|t| < t_{\beta'}) = \beta'. \quad \text{Отсюда} \quad \varepsilon = t_{\beta', n-2} \sqrt{\frac{n}{n-2} \frac{\hat{D}}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad \text{и}$$

$$I_{\beta} = \left(\varepsilon - t_{\beta', n-2} \sqrt{\frac{n}{n-2} \frac{\hat{D}}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \varepsilon + t_{\beta', n-2} \sqrt{\frac{n}{n-2} \frac{\hat{D}}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right).$$

На практике часто возникает вопрос об оценке отклонения истинной прямой $y = \alpha + \beta(x - \bar{x})$ от ее оценки $\hat{y} = a + b(x - \bar{x})$ при некотором заданном значении x . Особенно важен этот вопрос при построении прогноза. Оценкой точности здесь также может служить интервальная оценка y .

Используя обычные рассуждения, приводящие к t -статистикам, получаем:

$$M(\hat{y}) = \alpha + \beta(x - \bar{x}) = y, \quad D(\hat{y}) = D \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad \hat{y} \in N(M(\hat{y}), D(\hat{y})).$$

Тогда $z = \frac{\hat{y} - y}{\sqrt{D(\hat{y})}} \in N(0,1)$, а $t = \frac{z\sqrt{n-2}}{\sqrt{n\hat{D}/D}} \in t_{n-2}$. В нашем случае дробь

Стьюдента равна

$$t = \frac{(\hat{y} - y)}{\sqrt{D \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}} \sqrt{n-2} = \frac{(\hat{y} - y)\sqrt{n-2}}{\sqrt{\left[1 + \frac{n(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \hat{D}}} = (\hat{y} - y)d \in t_{n-2}.$$

$$P(|\hat{y} - y| < \varepsilon) = \beta' \quad \text{и} \quad P(|(\hat{y} - y)d| < \varepsilon d) = P(|(\hat{y} - y)d| < t_{\beta'}) = P(|t| < t_{\beta'}) = \beta'.$$

$$\text{Тогда} \quad \varepsilon = \frac{t_{\beta'}}{d} = t_{\beta', n-2} \sqrt{\frac{\hat{D}}{n-2}} \sqrt{\left[1 + \frac{n(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}, \quad I_y = (\hat{y} - \varepsilon, \hat{y} + \varepsilon) \quad \text{для}$$

любого конкретного x , так как $\varepsilon = \varepsilon(x)$. Очевидно, что длина доверительного интервала минимальна в точке $x = \bar{x}$. По мере удаления от \bar{x} точность оценки будет заметно снижаться. Наименее надежная оценка по

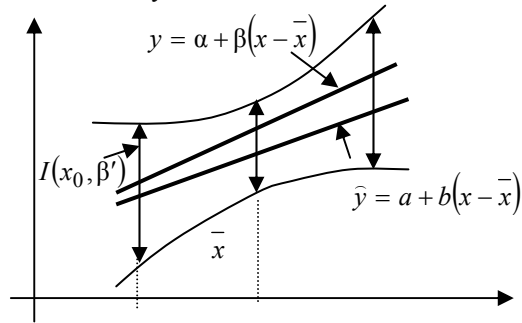


Рис. 7.1. Доверительные границы для линии регрессии

МНК будет получаться для ординат, отвечающим очкам, наиболее удаленным от \bar{x} (рис. 7.1). Вертикальные отрезки на рисунке представляют собой доверительные интервалы в соответствующих точках.

Пример. Дан отрезок временного ряда из средних котировок

Лондонской биржи металлов на свинец: (долл./т.).

1971, ян- варь	фев- раль	март	апр- ель	май	июнь	июль	ав- густ	сен- тябрь	ок- тябрь	но- ябрь	де- кабрь
265	268	270	270	267	268	264	259	139	229	221	231

Подобрать для этих данных параметры линейной регрессионной зависимости и построить доверительные интервалы для кривой регрессии.

Решение. Составим вспомогательную таблицу.

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	265	-5.5	10.75	30.25	-59.125
2	268	-4.5	13.75	20.25	-61.875
3	270	-3.5	15.75	12.25	-55.125
4	270	-2.5	15.75	6.25	-39.375
5	267	-1.5	12.75	2.25	-19.125
6	268	-0.5	13.75	0.25	-6.875
7	264	0.5	9.75	0.25	4.875
8	259	1.5	4.75	2.25	7.125
9	239	2.5	-15.25	6.25	-39.375
10	229	3.5	-25.25	12.25	-88.375
11	221	4.5	-33.25	20.25	-149.625
12	231	5.5	-23.25	30.25	-127.875
$\bar{x} = 6.5$	$\bar{y} = 254.25$			$\sum_{i=1}^{12} (x_i - \bar{x})^2 = 143$	$\sum_{i=1}^{12} (x_i - \bar{x})(y_i - \bar{y}) = -634.75$

Тогда

$$\hat{\alpha} = a = \bar{y} = 254.25, \hat{\beta} = b = \frac{\sum_{i=1}^{12} [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{12} (x_i - \bar{x})^2} = \frac{-634.75}{143} = -4.439. \text{ Таким}$$

образом, уравнение регрессии может быть записано в виде $\hat{y} = 254.25 - 4.439(x - 6.5)$ или $\hat{y} = 283.104 - 4.439x$.

Перейдем к построению доверительных интервалов, задав $\beta' = 0.9$, $\alpha' = 1 - \beta' = 0.1$. Для получения оценок дисперсий параметров a и b вычислим D , заменив ее оценкой \hat{D} . Рассчитаем по полученной линии регрессии значения \hat{y}_i .

x_i	1	2	3	4	5	6
\hat{y}_i	278.67	274.23	269.79	265.35	260.91	256.47
$y_i - \hat{y}_i$	-13.67	-6.23	0.21	4.65	6.09	11.53
$(y_i - \hat{y}_i)^2$	186.87	38.81	0.04	21.62	37.09	132.94

x_i	7	8	9	10	11	12
\hat{y}_i	252.03	247.59	243.15	238.71	234.28	229.84
$y_i - \hat{y}_i$	11.97	11.41	-4.15	-9.71	-13.28	1.16
$(y_i - \hat{y}_i)^2$	143.28	130.19	17.22	94.28	176.36	1.35

$$\hat{D} = \frac{1}{12} \sum_{i=1}^{12} (y_i - \hat{y}_i)^2 = \frac{980.05}{12} = 81.671, \quad \hat{\sigma} = 9.04. \quad \text{Тогда}$$

$$D(a) = \frac{\hat{D}}{12} = \frac{81.671}{12} = 6.81, \quad D(b) = \frac{\hat{D}}{\sum_{i=1}^{12} (x_i - \bar{x})^2} = \frac{81.671}{143} = 0.571,$$

$$D(y) = 81.671 \left(0.083 + \frac{(x - 6.5)^2}{143} \right) = 6.81 + 0.571(x - 6.5)^2, \quad n - 2 = 10, \quad \beta' = 0.9.$$

По таблице распределения Стьюдента находим $t_{0.9,10} = 2.228$. Отсюда

$$\varepsilon_a = t_{0.9,10} \sqrt{\frac{\hat{D}}{10}} = 2.228 \sqrt{\frac{81.671}{10}} = 6.37, \quad I_\alpha = (247.88, 260.62). \quad \text{Для пара-}$$

$$\text{метра } b \text{ все вычисления аналогичны } \varepsilon_b = t_{0.9,10} \sqrt{\frac{12}{10} \cdot \frac{\hat{D}}{\sum_{i=1}^{12} (x_i - \bar{x})^2}} =$$

$$= 2.228 \sqrt{\frac{12}{10} \cdot \frac{81.671}{143}} = 1.84, \quad I_\beta = (-6.28, -2.60).$$

Наконец, получим ε_y и посчитаем доверительные интервалы в нескольких точках:

$$\varepsilon_y = t_{0.9,10} \frac{\hat{\sigma}}{\sqrt{10}} \sqrt{\left(1 + \frac{12(x - \bar{x})^2}{143} \right)} = \frac{2.228 \cdot 9.04}{\sqrt{10}} \sqrt{1 + \frac{12}{143}(x - 6.5)^2} =$$

$$= 6.54 \sqrt{1 + 0.084(x - 6.5)^2}.$$

x_i	1	3	5	$x = \bar{x} = 6.5$	7	9	11	12
ε_y	12.31	9.32	7.13	6.54	6.61	8.08	10.75	12.31

На рис. 7.2 приведена линия регрессии: $\hat{y} = 254.25 - 4.439(x - 6.5)$ и ее 90% доверительные интервалы. Точки, соединенные прямыми – это исходная

выборка. Из графика видно, что линейная модель неудовлетворительно аппроксимирует исходные данные.

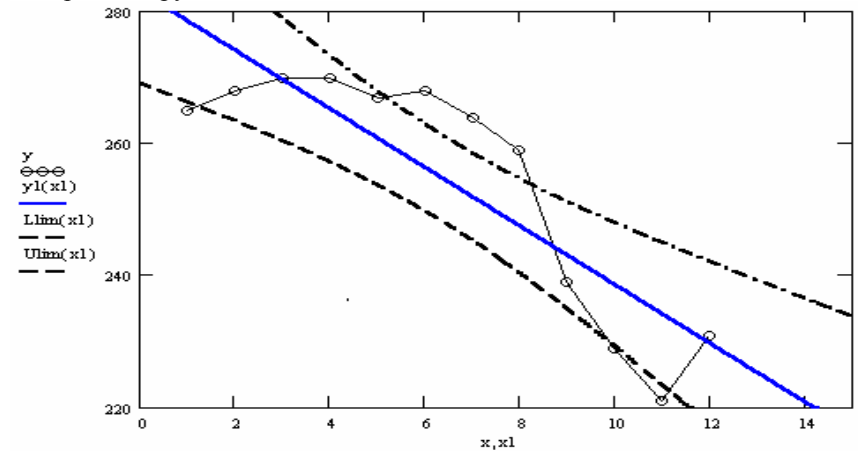


Рис. 7.2. Исходная выборка y , теоретическая линия регрессии $y1$ и ее 90% доверительные границы

7.4. Проверка адекватности линейной регрессии

Основой такой проверки служат взаимные отклонения от установленной закономерности, т.е. величины $y_i - \hat{y}_i$, $i = 1, 2, \dots, n$, где $\hat{y}_i = a + b(x_i - \bar{x})$. Поскольку аргумент x — одномерная переменная, точки $(x_i, y_i - \hat{y}_i)$ можно изобразить на чертеже. Такое наглядное представление наблюдений позволяет иногда обнаружить в поведении остатков какую-либо зависимость от x . Однако глазомерный анализ остатков возможен не всегда и не является правилом с контролируемыми свойствами. Нужны более точные методы.

Один из таких методов основывается на рассмотрении регрессионного анализа с точки зрения дисперсионного анализа. В этом случае общая вариация отклика относительно его среднего распадается на вариацию, обусловленную моделью, и остаточную вариацию, приписываемую случайным ошибкам.

Рассмотрим тождество $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$. Возведем его в квадрат и просуммируем по i от единицы до n . Получим

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}).$$

Но $\bar{y} = a$, тогда $\hat{y}_i - \bar{y} = [a + b(x_i - \bar{x})] - a = b(x_i - \bar{x})$ и

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = b \sum_{i=1}^n (y_i - \hat{y}_i)(x_i - \bar{x}) = 0 \quad \text{в силу условия}$$

$$\frac{\partial R}{\partial a} = -2 \sum_{i=1}^n (y_i - \hat{y}_i)(x_i - \bar{x}) = 0. \text{ Тогда для простой линейной модели будем}$$

иметь следующий вид разложения:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n [a + b(x_i - \bar{x}) - \bar{y}]^2 + \sum_{i=1}^n [a + b(x_i - \bar{x}) - y_i]^2, \quad (7.4.1)$$

где величина в левой части называется общей вариацией или суммой квадратов относительно среднего (Total (Corr.) Sum of Squares), первое слагаемое в правой части – суммой квадратов, обусловленной регрессией или моделью (Model Sum of Squares), второе слагаемое – сумма квадратов относительно модели регрессии или сумма квадратов ошибок (Error Sum of Squares).

При отсутствии повторных наблюдений проверяется гипотеза о равенстве коэффициента b нулю (в общем случае – об адекватности предлагаемой модели) с помощью F -критерия. Разность между наблюдениями y_i и теоретическими значениями y_i^T , определяемыми уравнением регрессии (7.2.1), можно записать в виде

$$\begin{aligned} y_i - y_i^T &= (y_i - \hat{y}_i) + (\hat{y}_i - y_i^T) = (y_i - \hat{y}_i) + [a + b(x_i - \bar{x}) - \alpha - \beta(x_i - \bar{x})] = \\ &= (y_i - \hat{y}_i) + (a - \alpha) + (b - \beta)(x_i - \bar{x}). \end{aligned}$$

Геометрическую интерпретацию последнего соотношения дает рис. 7.3. Возведем это соотношение в квадрат и просуммируем по всем i . Получим

$$\begin{aligned} (y_i - y_i^T)^2 &= (y_i - \hat{y}_i)^2 + (a - \alpha)^2 + \\ &+ (b - \beta)^2 (x_i - \bar{x})^2 + 2(y_i - \hat{y}_i) \times \\ &\times (a - \alpha) + 2(y_i - \hat{y}_i)(b - \beta)(x_i - \bar{x}) + \\ &+ 2(a - \alpha)(b - \beta)(x_i - \bar{x}), \\ \sum_{i=1}^n (y_i - y_i^T)^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \end{aligned}$$

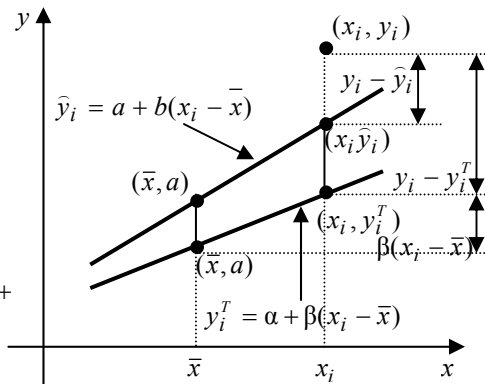


Рис. 7.3. Наблюдаемая и теоретическая линии регрессии

$$+ \sum_{i=1}^n (a - \alpha)^2 + \sum_{i=1}^n (b - \beta)^2 (x_i - \bar{x})^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(a - \alpha) + \\ + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(b - \beta)(x_i - \bar{x}) + 2 \sum_{i=1}^n (a - \alpha)(b - \beta)(x_i - \bar{x}). \text{ Но}$$

$$\sum_{i=1}^n (a - \alpha)(b - \beta)(x_i - \bar{x}) = (a - \alpha)(b - \beta) \sum_{i=1}^n (x_i - \bar{x}) = 0,$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)(b - \beta)(x_i - \bar{x}) = -\frac{b - \beta}{2} \frac{\partial R}{\partial b} = 0,$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)(a - \alpha) = -\frac{a - \alpha}{2} \frac{\partial R}{\partial a} = 0. \text{ Тогда}$$

$$\sum_{i=1}^n (y_i - y_i^T)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + n(a - \alpha)^2 + (b - \beta)^2 \sum_{i=1}^n (x_i - \bar{x})^2. \quad (7.4.2)$$

Дробь $\frac{1}{n} \sum_{i=1}^n (y_i - y_i^T)^2 \cdot n / D$ имеет вид χ_n^2 -статистики (см. под-

разд. 4.7), т.е. сумма $\sum_{i=1}^n (y_i - y_i^T)^2 \in \chi_n^2 \cdot D$. Эта сумма разбита на три ком-

поненты. Вторая и третья из них зависят лишь от a и b соответственно, и, следовательно, каждая имеет одну степень свободы. Первый член в правой части включает n разностей $y_i - \hat{y}_i$, на которые наложены два

ограничения $\frac{\partial R}{\partial a} = 0$ и $\frac{\partial R}{\partial b} = 0$, в силу чего он имеет $n - 2$ степени сво-

боды. Поскольку сумма трех сумм квадратов в правой части (7.4.2) равна сумме квадратов левой части и это же имеет место для степеней свободы, каждый член в правой части распределен как $\chi^2 D$ с соответствующим числом степеней свободы и эти члены независимы между собой.

$$\text{Таким образом, } \frac{n(a - \alpha)^2}{D} \in \chi_1^2, \quad \frac{(b - \beta)^2 \sum_{i=1}^n (x_i - \bar{x})^2}{D} \in \chi_1^2,$$

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{D} \in \chi_{n-2}^2. \text{ Теперь можно построить критерий для проверки}$$

нулевой гипотезы $H_0 : \beta = 0$, составив отношение

$$\frac{\left((b-\beta)^2 \sum_{i=1}^n (x_i - \bar{x})^2 / D \right) / 1}{\left(\sum_{i=1}^n (y_i - \hat{y}_i)^2 / D \right) / (n-2)} = \frac{(b-\beta)^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\tilde{D}} \in F_{1, n-2}, \quad \text{где}$$

$\tilde{D} = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ - несмещенная оценка дисперсии ошибок наблюдений.

Если гипотеза H_0 справедлива, то $\beta = 0$ и

$b^2 \sum_{i=1}^n (x_i - \bar{x})^2 / \tilde{D} \in F_{1, n-2}$. Практически при вычислении этого отношения

величину $\sum_{i=1}^n (y_i - \hat{y}_i)^2$, входящую в \tilde{D} , получают с помощью соотношения (7.4.1).

Пример. Проверим гипотезу $H_0 : \beta = 0$ с $\alpha' = 0.1$ по данным предыдущего примера.

Из таблиц подразд. 7.3 имеем $\sum_{i=1}^{12} (x_i - \bar{x})^2 = 143$, $\sum_{i=1}^{12} (y_i - \hat{y}_i)^2 = 980.05$, $b = -4.439$. Тогда $\tilde{D} = 980.05/10 = 98.01$,

$z = b^2 \sum_{i=1}^{12} (x_i - \bar{x})^2 / \tilde{D} = 28.75$. По таблицам F - распределения находим

$F_{0.9, 1, 10} = 3.29$. Таким образом, $z \in \omega$ и H_0 должна быть отвергнута, т.е. зависимость между x и y значима и значение b отлично от нуля.

Надо заметить, что сформулированный критерий может сигнализировать скорее о наличии зависимости между x и y . О качестве аппроксимации исходных данных данной моделью лучше судить по ошибке \tilde{D} . Ясно, что для данного примера можно попытаться сгладить исходные данные параболой, в пользу чего говорит и глазомерный анализ построенных зависимостей (см. подразд. 7.3).

7.5. Выбор наилучшей регрессии

Одна из основных задач регрессионного анализа состоит в решении вопроса о том, какие именно регрессоры (независимые переменные) следует включать в модель. Пусть x_1, x_2, \dots, x_k - полный набор всех возмож-

ных регрессоров, содержащий такие функции, как квадраты, смешанные произведения и прочие функции, которые кажутся подходящими. Для выбора некоторого подмножества из этой полной совокупности регрессоров есть два противоположных подхода.

С одной стороны, в модель для полноты учета следует включать по возможности наибольшее число регрессоров. С другой - при увеличении числа регрессоров возрастают затраты на построение и использование модели, а также возрастает дисперсия прогноза. Подходящим компромиссом между этими двумя крайностями является процедура, называемая обычно «выбором наилучшего уравнения регрессии». Термин «наилучшее», конечно, субъективен. Нет никакой единой статистической процедуры для выбора соответствующего подмножества, и все статистические методы предполагают необходимость субъективного решения.

Подбор конкретного вида функциональной зависимости – наиболее трудная и творческая часть задачи регрессии.

В пакете STATGRAPHICS Plus for Windows реализована процедура пошаговой множественной регрессии, включающая последовательное увеличение и последовательное уменьшение группы независимых переменных. Другой подход построения всех возможных регрессий состоит в подборе всех возможных уравнений регрессии, которые можно получить, выбирая по $1, 2, \dots, k$ регрессоров из совокупности x_1, x_2, \dots, x_k . Поскольку для каждого регрессора имеются только две возможности: либо он включается в уравнение, либо не включается в него, то всего имеются 2^k возможных уравнений регрессии. Метод применяется, если k не слишком велико.

7.6. Лабораторная работа № 9. Регрессионный анализ в пакетах STATGRAPHICS и MATHCAD

Процедура простой регрессии заключается в нахождении аналитического выражения для связи двух переменных x и y . В пакете STATGRAPHICS предусмотрено определение следующих моделей простой регрессии:

1. Линейная: $y = a + bx$;
2. Экспоненциальная: $y = e^{a+bx} = a_1 e^{bx}$;
3. Обратная по y : $y = 1/(a + bx)$;
4. Обратная по x : $y = a + b/x$;
5. Дважды обратная: $y = 1/(a + b/x)$;
6. Логарифмическая: $y = a + b \ln x$;

7. Мультипликативная: $y = ax^b$;

8. Полиномиальная: $\begin{cases} y = a + b\sqrt{x}, \\ y = (a + bx)^2; \end{cases}$

9. S - кривая: $y = e^{a+b/x}$;

10. Логистическая.

Кроме того, в пакете реализована процедура пошаговой множественной регрессии, в которой количество и вид регрессоров задаются исследователем. Процедура устроена так, что путем последовательного перебора удастся подбирать модели, содержащие гораздо меньше переменных по сравнению с исходным множеством и имеющие лучшие статистические характеристики.

Вычислим в пакете STATGRAPHICS параметры линейной модели для следующих данных:

x	51	32	80	73	64	45	83	44	93
y	52.7	15.2	89.5	94.8	76	39.3	114.8	36.5	137.4
x	28	35	40	29	53	58	65	75	
y	5.3	20.7	21.7	9.2	55.4	64.3	79.1	101	

Эти данные химического производства представляют собой зависимость объема продукта y (кг) от температуры реакции x (°C).

В электронную таблицу STATGRAPHICS занесем сначала значения аргумента – Temp, затем функции – Prod и вызовем процедуру построения простой регрессии: Relate (Отношения данных)→Simple Regression (Простая регрессия). В появившемся окне диалога (рис. 7.4) выделяем сначала переменную Prod и вводим ее в поле анализа y нажатием кнопки со стрелкой, а затем переменную Temp в поле анализа x . Нажимаем ОК. На экран выдается заставка процедуры простой регрессии со статистической сводкой применительно к линейной модели (рис. 7.5).

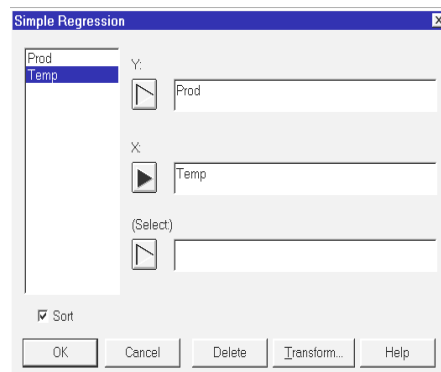


Рис. 7.4. Окно диалога для ввода данных в процедуру построения моделей простой регрессии

Regression Analysis - Linear model: $Y = a + b \cdot X$

Dependent variable: Prod

Independent variable: Temp

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	-48,4705	3,86434	-12,543	0,0000
Slope	1,93766	0,0653622	29,6449	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	24618,6	1	24618,6	878,82	0,0000
Residual	420,197	15	28,0131		
Total (Corr.)	25038,8	16			

Correlation Coefficient = 0,991574

R-squared = 98,3218 percent

Standard Error of Est. = 5,29274

Рис. 7.5. Результаты расчета модели простой линейной регрессии

В первой таблице приведены оценки параметров простой линейной модели $y = a + bx$ и их статистические характеристики. Строка Intercept (Свободный член) относится к параметру a , а строка Slope (Наклон) – к параметру b . Столбец Estimate (Оценки) содержит оценки этих параметров, столбец Standard Error (Стандартная ошибка) дает значения стандартных ошибок указанных коэффициентов. Два последних столбца T Statistic и p-Value содержат значения стьюдентовых отношений t_a и t_b (см. подразд. 7.3) и их минимальные уровни значимости для проверки гипотезы о равенстве значений коэффициентов нулю. Так как p-Value очень малы, то ненулевые значения коэффициентов a и b значимы.

Таблица Analysis of Variance является базовой таблицей дисперсионного анализа и служит для оценки адекватности предлагаемой модели данных. Описание этой таблицы дано в лабораторной работе № 8 (подразд. 6.5, рис. 6.4). Общая дисперсия разлагается здесь на две части по формуле (7.4.1) на дисперсию, обусловленную моделью, и дисперсию ошибок наблюдений.

F-Ratio (F-отношение) служит для проверки гипотезы о равенстве коэффициента b нулю (см. подразд. 7.4).

Еще одним показателем качества подобранной модели является выборочный коэффициент корреляции Пирсона (Correlation Coefficient)

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})[(y_i - \bar{y}) - b(x_i - \bar{x})]}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2}}. \quad (7.6.1)$$

Как известно, если переменные x и y связаны линейной зависимостью, то $r_{xy} = 1$, поэтому близость коэффициента корреляции к единице служит мерой линейной связи между x и y .

Значения R^2 (R-Squared) является отношением суммы квадратов, обусловленных регрессией, к общей сумме квадратов откликов:

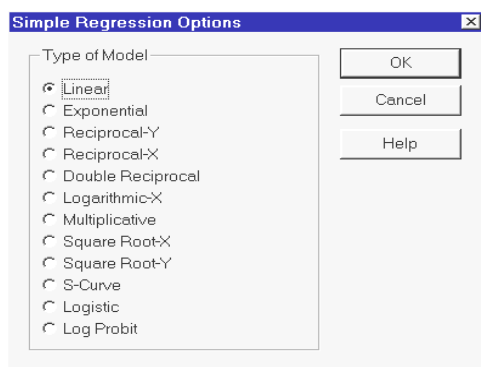
$$R^2 = \frac{24618.6}{25038.8} \cdot 100\% = 98.3218. \text{ Этот показатель дает долю общего раз-}$$

броса функции y относительно \bar{y} , объясняемую регрессией. Величину

R^2 также часто именуют коэффициентом детерминации и измеряют не в долях единицы, а в процентах. Чем ближе значение R^2 к ста процентам, тем лучше подобранная модель описывает данные эксперимента.

Последняя характеристика таблицы Standard Error of Est. (Стандартная ошибка оценки) равна $\sqrt{\tilde{D}}$, где \tilde{D} - несмещенная оценка дисперсии ошибок,

$$\tilde{D} = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (7.6.2)$$



После щелчка правой кнопкой мыши в поле заставки Linear Model и выбора пункта Analysis Options дополнительного меню открывается еще одно меню со списком всех моделей простой регрессии (рис. 7.6). Можно выбрать любую из них. Если исходные данные позволяют, будут вычислены оценки параметров этой модели.

Рис. 7.6. Список реализуемых моделей регрессии

Рассмотрим теперь назначение всех процедур двух дополнительных меню, которые можно использовать при расчете моделей регрессии: Tabular Options (рис. 7.7) и Graphical Options.

Analysis Summary (Сводка анализа). Информация, выводимая этой процедурой, уже описана, так как она выводится пакетом сразу после задания вида рассматриваемой модели по умолчанию.

Lack-of-Fit Test (Тест на адекватность) предназначен для проверки адекватности линейной модели при наличии повторных наблюдений. В этом случае появляется возможность получить еще одну оценку изменчивости случайной составляющей ϵ в модели (7.1.3) и сравнить ее с оценкой дисперсии (7.6.2). Поскольку в нашем примере повторных наблюдений нет, тест работать не может (рис. 7.8).

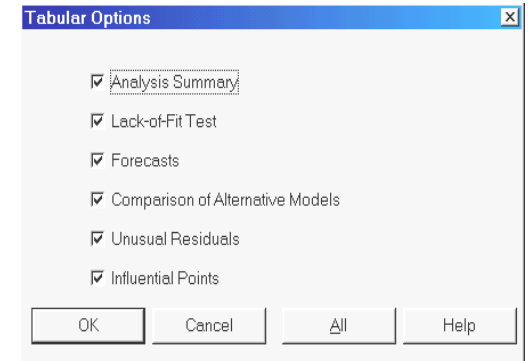


Рис. 7.7. Меню задания процедур регрессии

Analysis of Variance with Lack-of-Fit					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	24618,6	1	24618,6	878,82	0,0000
Residual	420,197	15	28,0131		
Lack-of-Fit	420,197	15	28,0131		
Pure Error(полная ошибка)	0,0	0			
Total (Corr.)	25038,8	16			

Рис. 7.8. Результаты анализа адекватности линейной модели

Forecasts (Предсказания). Эта процедура вычисляет предсказанные по рассмотренной модели величины y_i для ряда значений аргумента x_i . Кроме того, выводятся $(1-\alpha)\%$ доверительные интервалы для текущего y_i и для прогнозируемых значений новых наблюдений. По умолчанию задается наименьший и наибольший аргумент x (рис. 7.9).

Predicted Values					
X	Predicted Y	95,00% Prediction Limits		95,00% Confidence Limits	
		Lower	Upper	Lower	Upper
28,0	5,78388	-6,45191	18,0197	1,04592	10,5218
93,0	131,732	119,017	144,446	125,867	137,596

Рис. 7.9. Предсказанные по умолчанию наблюдения и их доверительные границы

Значения аргумента для предсказываемых y_i и уровень доверитель-

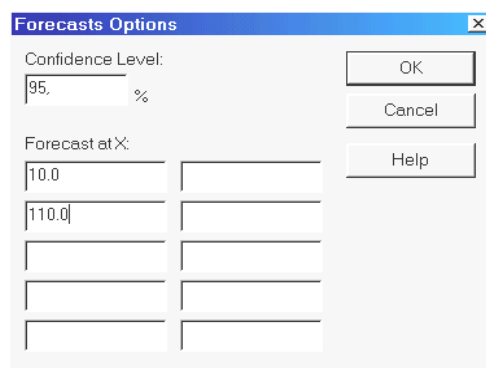


Рис. 7.10. Меню задания аргументов для предсказываемых значений регрессии

ной вероятности $(1 - \alpha)$ можно задать в дополнительном меню (рис. 7.10), которое вызывается щелчком правой кнопки мыши в поле заставки Forecasts и выборе пункта Pane Options. Зададим, например, лишь два значения аргумента $x = 10.0$ и $x = 110.0$. На экран будет выведена следующая информация (рис. 7.11).

Predicted Values					
X	Predicted Y	95,00% Prediction Limits		95,00% Confidence Limits	
		Lower	Upper	Lower	Upper
10,0	-29,0939	-42,3379	-15,85	-36,032	-22,1559
110,0	164,672	150,821	178,523	156,636	172,708

Рис. 7.11. Результаты предсказания заданных наблюдений

Comparison of Alternative Models (Сравнение альтернативных моделей) выводит таблицу (рис. 7.12), в которой представлены результаты анализа для всех типов зависимостей y от x , упорядоченные по убыванию модуля коэффициента корреляции. Оказывается, линейная модель – лучшая по качеству аппроксимации экспериментальных наблюдений. Последние три модели по данным выборки вычислены быть не могут.

Unusual Residuals (Необычные остатки). Эта процедура выводит значения резко выделяющихся наблюдений. Для данной выборки это наблюдение номер три.

Influential Points (Точки влияния). Процедура дает таблицу наблюдений больших некоторого значения по отношению к среднему значению

Comparison of Alternative Models					
Model	Correlation	R-Squared			
Linear	0,9916	98,32%			
S-curve	-0,9894	97,89%			
Square root-X	0,9877	97,55%			
Square root-Y	0,9792	95,89%			
Logarithmic-X	0,9772	95,50%			
Multiplicative	0,9682	93,75%			
Reciprocal-X	-0,9385	88,08%			
Exponential	0,9238	85,33%			
Double reciprocal	0,8571	73,46%			
Reciprocal-Y	<no fit>				
Logistic	<no fit>				
Log probit	<no fit>				

Unusual Residuals					
Row	X	Y	Predicted Y	Residual	Studentized Residual
3	80,0	89,5	106,542	-17,042	-7,77

Рис. 7.12. Результаты анализа альтернативных моделей и резко выделяющиеся наблюдения

всех элементов выборки, влияние которых на определение коэффициентов регрессии выше определенного уровня. В нашей выборке таких точек нет, поэтому таблица пуста.

При анализе каждой регрессионной модели большое значение имеет графическая информация. В пакете STATGRAPHICS можно построить графики пяти видов, которые задаются в меню Graphical Options (рис. 7.13). Опишем коротко особенности всех выводимых графиков.

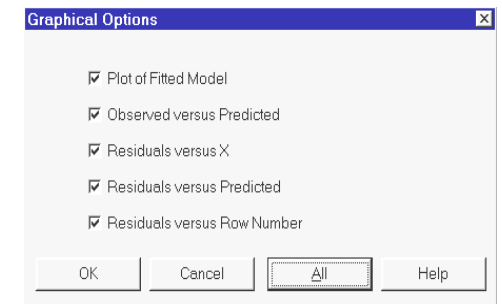


Рис. 7.13. Панель графических параметров в задаче регрессии

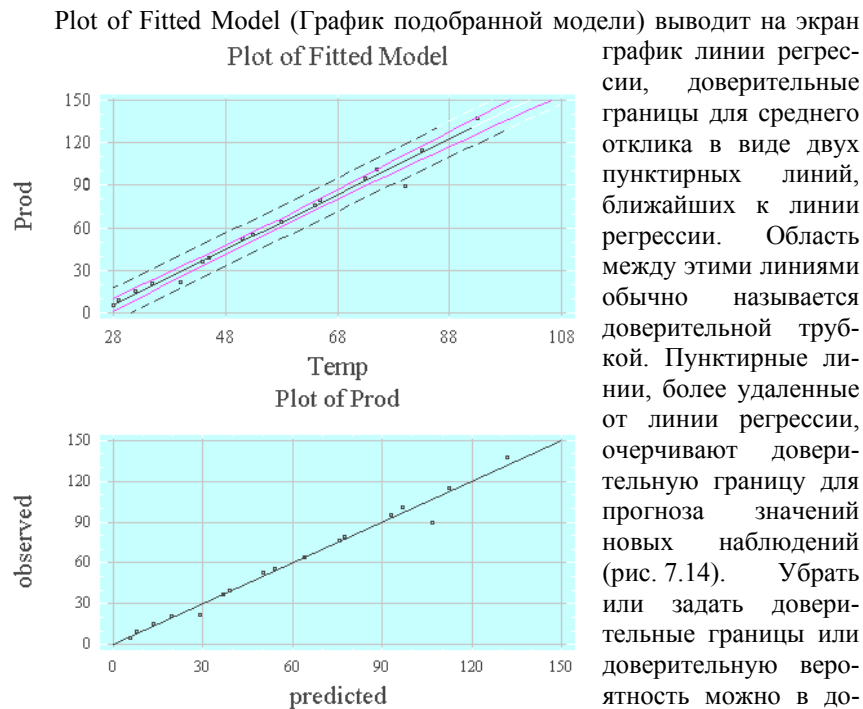


Рис. 7.14. Графики линии регрессии и предсказанных наблюдений

Observed versus Predicted (График предсказанных наблюдений) строит график предсказанных значений в зависимости от наблюдаемых. Эта процедура полезна для выявления случаев, в которых дисперсия зависимых переменных не постоянна. Предсказанные значения вычисляются по формуле $\hat{y}_i = a + bx_i$ для значений аргументов, не входящих в область определения исходной выборки.

Следующие три графика Residual versus x (График остатков), Residual versus Predicted (График остатков в зависимости от предсказаний) и Residual versus Row Number (График остатков в зависимости от номера наблюдения) полезны для представления о том, насколько подобранная модель соответствует исходным данным и насколько выполняются условия применения метода наименьших квадратов (рис. 7.15, 7.16). Подробное описание процедур анализа остатков имеется, например, в [7].

Заметим, что на график можно выводить простые остатки (разности $y_i - \hat{y}_i$) или студентизированные. Для этого нужно открыть дополнительное меню Residual Plot Options щелчком правой кнопки мыши в любом месте графика и выбором пункта Pane Options.

Совершенно аналогично осуществляется процедура полиномиальной и множественной регрессии. Например, полиномиальная регрессия находит аналитическое выражение связи двух переменных x , y в виде степенного многочлена

$$y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n.$$

Пакет STATGRAPHICS предоставляет возможность строить такие многочлены вплоть до восьмой степени, причем степень можно выбирать по желанию пользователя.

Процедура множественной регрессии позволяет осуществлять пошаговой отбор переменных. Для этого в разделе Fit (Аппроксимация) окна диалога Multiple Regression Options, открывающегося щелчком правой кнопки мыши, необходимо установить переключатель в положение Forward Selection (Алгоритм последовательного увеличения группы переменных) или Backward Selection (Уменьшение группы переменных). Кроме того,

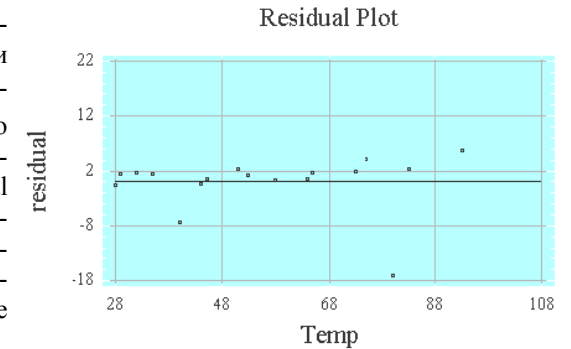


Рис. 7.15. График простых остатков модели линейной регрессии

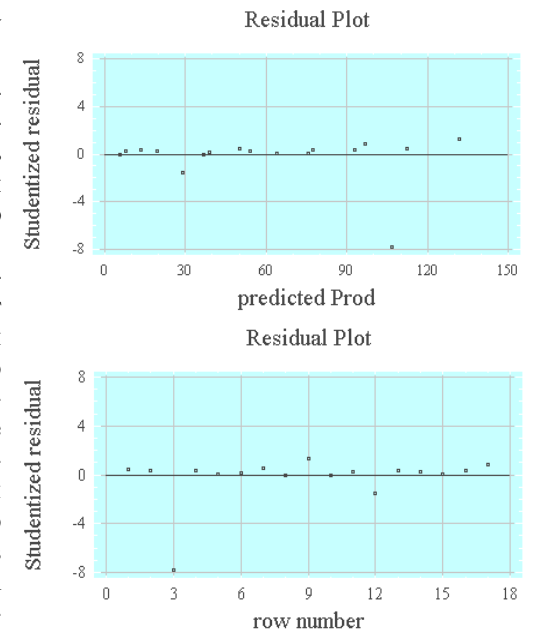


Рис. 7.16. Графики студентизированных остатков модели линейной регрессии

можно снять флажок Constant in Model, отказавшись от свободного члена в подбираемой модели.

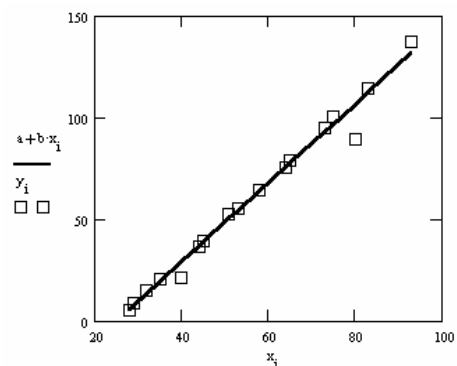
Построение этих моделей и учет их особенностей можно легко освоить самостоятельно. Строение и содержание всех таблиц и графиков, выводимых процедурами этих моделей, аналогично данным, получаемым в моделях простой регрессии.

В имеющемся пакете MATHCAD две встроенные функции **intercept** (to intercept по-английски – отложить отрезок на линии) и **slope** (наклон), решают задачу линейного сглаживания экспериментальных данных методом наименьших квадратов. Доверительные интервалы для параметров модели α и β (см. формулу (7.2.1)), а также для линии регрессии необходимо вычислять отдельно. Запрограммируем необходимые формулы подразд. 7.3 и построим нужные графики.

Процедура **csort(A,j)** производит сортировку матрицы A по столбцу j , т.е. переставляет строки матрицы по возрастанию значений элементов в столбце j . Результат – матрица такого же размера, как A .

$$\text{ORIGIN} := 1 \quad x1 := \begin{pmatrix} 51 \\ 32 \\ 80 \\ 73 \\ 64 \\ 45 \\ 83 \\ 44 \\ 93 \end{pmatrix} \quad x2 := \begin{pmatrix} 28 \\ 35 \\ 40 \\ 29 \\ 53 \\ 58 \\ 65 \\ 75 \end{pmatrix} \quad y1 := \begin{pmatrix} 52.7 \\ 15.2 \\ 89.5 \\ 94.8 \\ 76 \\ 39.3 \\ 114.8 \\ 36.5 \\ 137.4 \end{pmatrix} \quad y2 := \begin{pmatrix} 5.3 \\ 20.7 \\ 21.7 \\ 9.2 \\ 55.4 \\ 64.3 \\ 79.1 \\ 101 \end{pmatrix} \quad x := \text{stack}(x1, x2)$$

$$y := \text{stack}(y1, y2) \quad n := \text{rows}(x) \quad n = 17 \quad a := \text{intercept}(x, y) \quad a = -48.471 \quad b := \text{slope}(x, y) \\ b = 1.938 \quad M^{(1)} := x \quad M^{(2)} := y \quad M := \text{csort}(M, 1) \quad x := M^{(1)} \quad y := M^{(2)}$$



$$i := 1..n \quad y_{r_i} := a + b * x_i$$

Построим исходную выборку и график полученной линейной модели. Определим доверительные интервалы для параметров линейной регрессии. Зададим уровень доверительной вероятности β' равным 0.95:

$$xmean := \text{mean}(x) \quad xmean = 55.765 \quad ymean := \text{mean}(y) \quad ymean = 59.582$$

$$\beta1 := 0.95 \quad \alpha1 := 1 - \beta1 \quad t := \text{qt}\left(1 - \frac{\alpha1}{2}, n - 2\right)$$

$$\alpha1 = 0.05 \quad t = 2.131$$

$$D1 := \frac{1}{n - 2} \sum_{k=1}^n (y_k - yr_k)^2 \quad D1 = 28.013$$

$$aleft := a - t * \sqrt{\frac{D1}{n - 2}} \quad aright := a + t * \sqrt{\frac{D1}{n - 2}} \quad aleft = -51.383 \quad aright = -45.558$$

$$\varepsilon b := \sqrt{\frac{n}{n - 2} * \frac{D1}{\sum_{k=1}^n (x_k - xmean)^2}} \quad \varepsilon b = 0.070 \quad bleft := b - t * \varepsilon b \quad bright := b + t * \varepsilon b$$

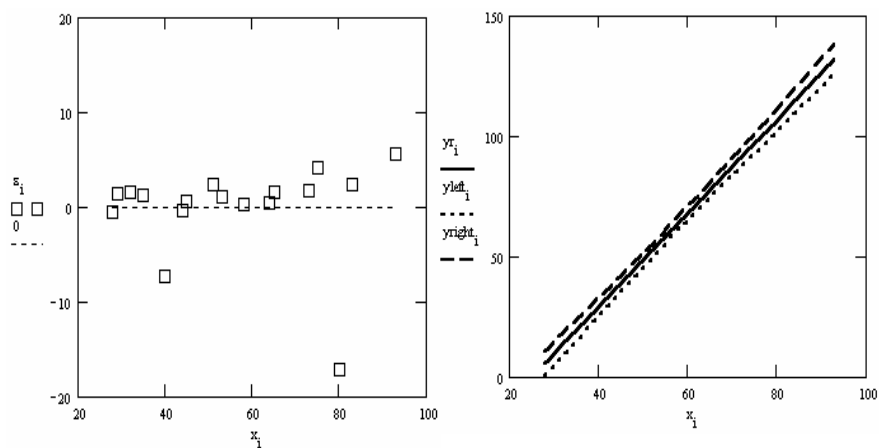
$$bleft = 1.789 \quad bright = 2.086$$

Итак, точные доверительные интервалы для параметров линейной регрессии построены. $\alpha \in (-51.383, -45.558)$, $\beta \in (1.789, 2.086)$ с вероятностью 95 процентов. Найдём теперь доверительный интервал для линии регрессии, построим этот интервал и один график для остатков.

$$i := 1..n \quad \varepsilon_i := y_i - yr_i$$

$$yleft_i := yr_i - t * \sqrt{\frac{D1}{n - 2} * \left[1 + \frac{n * (x_i - xmean)^2}{\sum_{k=1}^n (x_k - xmean)^2} \right]}$$

$$yright_i := yr_i + t * \sqrt{\frac{D1}{n - 2} * \left[1 + \frac{n * (x_i - xmean)^2}{\sum_{k=1}^n (x_k - xmean)^2} \right]}$$



Задание № 1. Найти в пакетах STATGRAPHICS и MATHCAD оценки параметров линейной регрессии y на x , доверительные интервалы для параметров и линии регрессии и проверить согласие линейной регрессии с результатами наблюдений. Принять уровень доверительной вероятности равным 0.90.

1.

x	2	5	8	10	14	15	4	12	3	7	6
y	14.39	9.45	7.05	5.32	16.94	1.97	8.75	3.41	13.37	8.22	9.39

2.

x	2.7	4.6	6.3	7.8	9.2	10.6	12.0	13.4	14.7
y	17.0	16.2	13.3	13.0	9.7	9.9	6.2	5.8	5.7

3.

x	7.9	11.6	12.8	14.9	16.3	18.6	20.3	21.9	23.6
y	13.0	22.8	24.8	28.6	31.6	38.7	40.0	44.9	43.0

4.

x	1	2	3	4	5	6	7	8	9
y	0.21	0.32	0.58	1.02	1.76	2.68	3.75	5.07	6.62
x	10	11	12	13	14	15	16	17	
y	8.32	10.21	12.33	14.58	17.07	19.53	22.72	29.05	

5.

x	2.97	3.56	6.45	1.12	6.66	1.37	6.80	2.19	5.11	7.36
y	2.94	3.54	6.48	1.08	6.73	1.33	6.86	2.34	4.96	7.21

6.

x	19.65	20.01	31.15	32.50	35.95	50.15	59.65
y	3.44	3.93	4.98	5.45	6.40	8.88	11.22

7.

x	0.36	0.56	0.76	0.21	0.44	0.60	0.82	1.12	1.56
y	17	64	62	9	32	71	93	118	163

8.

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2
y	6.00	5.82	5.75	5.83	5.63	5.60	5.69	5.47	5.41	5.23	5.34	5.30

9.

x	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1
y	-0.02	-0.28	-0.06	-0.0	-0.24	-0.11	-0.28	-0.35	-0.47	-0.47	-0.52	-0.68

10.

x	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3
y	0.30	0.50	0.82	1.43	1.49	1.85	2.01	2.56	2.72	2.85	3.12	3.75

11.

x	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55
y	2.43	2.67	2.71	3.15	3.47	3.76	3.91	4.46	4.76	5.15	5.54	5.61

12.

x	1	2	3	4	5	6	7	8	9	10	11
y	6.32	6.52	6.65	7.26	7.49	7.83	8.13	8.40	8.58	9.01	9.05

13.

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2
y	-0.01	-0.20	-0.31	-0.63	-0.73	-0.87	-1.05	-1.39	-1.05	-1.40	-1.74	-1.88

14.

x	2.5	2.75	3.0	3.25	3.5	3.75	4.0	4.25	4.5	4.75
y	3.90	3.83	3.37	3.42	3.25	2.90	2.90	2.76	2.82	2.35

15.

x	1.1	2.2	3.3	4.4	5.5	6.6	7.7	8.8	9.9	11.0
y	-2.96	-2.68	-2.41	-2.37	-1.98	-1.73	-1.39	-1.29	-0.85	-0.69

16.

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1
y	24.3	27.1	34.7	39.1	47.6	55.4	59.3	65.4	72.8	77.9	82.4

17.

x	5.3	5.1	4.9	4.7	4.5	4.3	4.1	3.9	3.7	3.5	3.3	3.1	2.9	2.7
y	4.27	4.45	4.84	5.14	5.55	5.85	6.18	6.38	6.72	7.04	7.26	7.70	7.78	8.33

18.

x	1	2	3	4	5	6	7	8	9	10	11
y	124.9	127.1	134.0	139.1	147.3	155.0	159.8	165.4	172.5	177.4	182.1

19.

x	1	2	3	4	5	6	7	8	9	10	11	12
y	0.00	0.23	0.32	0.24	0.35	0.77	0.68	0.92	0.97	1.08	1.15	1.37

20.

x	6.5	6.7	6.9	7.1	7.3	7.5	7.7	7.9	8.1	8.3	8.5	8.7	8.9
y	5.65	5.43	5.25	5.00	4.79	4.57	4.30	4.07	3.84	3.52	3.28	2.93	2.80

21.

x	0.5	0.9	1.3	1.7	2.1	2.5	2.9	3.3	3.7	4.1	4.5	4.9
y	3.80	3.65	4.52	8.91	9.17	11.10	14.97	17.23	18.32	19.85	23.56	28.29

22.

x	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6
y	0.54	-0.68	-0.60	-0.12	-0.01	0.18	0.48	0.85	1.15	1.18	1.32	1.88

23.

x	1	2	3	4	5	6	7	8
	7.92	8.03	7.64	7.61	7.24	7.08	6.81	6.39
x	9	10	11	12	13	14	15	
y	6.65	6.16	5.89	5.67	5.87	5.36	5.27	

24.

x	7.5	7.25	7.0	6.75	6.5	6.25	6.0	5.75	5.5	5.25	5.0	4.75
y	38.81	38.62	38.40	39.17	37.12	34.95	35.16	36.83	37.49	34.71	36.08	35.10

25.

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3
y	0.93	0.96	1.00	1.14	1.35	1.38	1.22	1.30	1.37	1.23	1.44	1.47	1.66

26.

x	3.14	2.91	2.68	2.45	2.22	1.99	1.76	1.53	1.30	1.07	0.84	0.61
y	2.14	2.19	2.32	2.59	2.56	2.64	2.66	2.84	3.04	2.94	3.23	3.27

27.

x	1.27	1.61	1.95	2.29	2.63	2.97	3.31	3.65	3.99	4.33
y	35.52	34.89	36.41	38.67	40.62	43.95	46.74	49.36	49.51	50.68

28.

x	1	2	3	4	5	6	7	8	9	10	11
y	5.89	6.16	6.65	6.39	6.81	7.08	7.24	7.61	7.64	8.03	7.92

29.

x	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4
y	63.10	63.08	65.46	68.55	70.73	77.70	72.25	77.39	79.95	89.63

30.

x	3.71	3.45	3.19	2.93	2.68	2.42	2.16	1.90	1.64	1.39
y	-16.30	-19.50	-22.31	-25.54	-26.42	-28.57	-33.24	-36.45	-37.80	-38.44

8. НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ СТАТИСТИКИ

8.1. Основные понятия и область применимости непараметрических методов

При рассмотрении статистических гипотез стандартными средствами приходится предполагать вид распределения статистик критерия. Для более тонких и глубоких выводов вводится предположение, что наблюдаемые случайные величины имеют нормальное распределение. На этой основе за многие годы выросла обширная и развитая система статистической обработки регрессионных и факторных экспериментов, в частности дисперсионный анализ. Она позволяла решать основные статистические задачи: получать оценки неизвестных параметров (как точечные, так и интервальные), проверять статистические гипотезы, проводить сравнения и тому подобное. На практике все эти методы приходится применять и в тех случаях, когда наблюдения, возможно, распределены иначе, что превращает точные методы в приближенные. Иногда при этом нарушения, кажущиеся незначительными и поэтому трудно обнаружимые, могут существенно исказить конечные результаты: привести к смещению оценок и доверительных границ.

Один из способов ослабить эти неприятные явления – разработать и применять такие статистические правила, результаты которых были бы устойчивы или малочувствительны к тем или иным отступлениям от предпосылок модели. К сожалению, такие устойчивые (робастные) правила приводят к тому, что если модель полностью справедлива, они имеют меньшую точность, чем традиционные оптимальные процедуры и правила. Такой подход и методы, им реализованные, называются непараметрическими. Точнее эти методы, не предназначенные специально для какого-нибудь параметрического семейства распределений (например, гауссовского) и не использующие его свойства. Благодаря этому, непараметрические методы имеют более широкую область применения, но более низкую точность.

Непараметрические методы используют не сами численные значения элементов выборки, а структурные свойства выборки: отношения порядка между ее элементами. В связи с этим, конечно, теряется часть информации, содержащаяся в выборке, поэтому мощность непараметрических критериев меньше, чем мощность их параметрических аналогов. Однако непараметрические методы могут применяться при более общих предположениях и более просты с точки зрения выполнения вычислений.

Большая группа непараметрических критериев используется для проверки гипотезы о принадлежности двух выборок x_1, x_2, \dots, x_{n_1} и

y_1, y_2, \dots, y_{n_2} к одной и той же генеральной совокупности, т.е. о том, что функции распределения двух генеральных совокупностей $F_X(x)$ и $F_Y(y)$ равны: $H_0 : F_X(x) \equiv F_Y(y) \big|_{x=y}$. Такие генеральные совокупности называются однородными. Необходимое условие однородности состоит в равенстве характеристик положения и (или) рассеивания таких, как средние, медианы, дисперсии и тому подобное. Непараметрические критерии в качестве основного предположения используют только непрерывность распределения генеральной совокупности.

Все выводы статистических методов непараметрического типа основаны на исследовании знаков и рангов. Особенно значимые результаты получены за последние десятилетия. Рассмотрим несколько примеров.

8.2. Критерий знаков

Простейший критерий такого рода, критерий знаков, применяется для проверки гипотезы H_0 об однородности генеральных совокупностей попарно связанным выборкам. Статистикой критерия знаков является число знаков «+» или «-» в последовательности знаков разностей парных выборок $(x_i, y_i), i = \overline{1, n}$. Если сравниваемые выборки получены из однородных генеральных совокупностей, то значения x_i и y_i взаимозаменяемы и, следовательно, вероятности появления положительных и отрицательных разностей $x_i - y_i$ равны, т.е. можно предположить, что $p(x, y) = p(y, x)$. Если же совокупность x в среднем больше или меньше, то $p(x, y) \neq p(y, x)$.

Пусть, к примеру, каждый y будет на θ больше, чем соответствующий x . Тогда $p(x, y - \theta) = p(y - \theta, x)$, т.е. θ является медианой разности $y - x$. Покажем это. Подставим $w = y - \theta$, получим $p(x, w) = p(w, x)$, т.е. совместная плотность симметрична относительно прямой $w = y - \theta$

(рис. 8.1). Тогда $\int_{\Omega_1} p(x, w) d\Omega =$

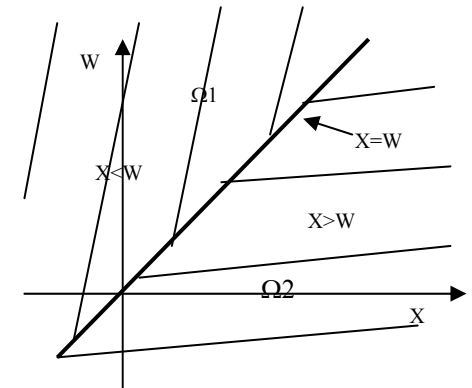


Рис. 8.1. Области интегрирования вероятностей

$p(x < w) = \int_{\Omega} p(x, w) d\Omega = p(w < x)$. Из свойств симметричности следует,

что интегралы численно равны, тогда $p(x < w) = p(w < x)$ или, подставляя $y - \theta$ вместо w , имеем $p(x < y - \theta) = p(y - \theta < x)$. Далее, очевидно, что $p(y - \theta < x) = p(y - x < \theta)$ и $p(x < y - \theta) = p(\theta < y - x) = p(y - x > \theta)$. Так как левые части равны, то равны и правые, следовательно, $p(y - x < \theta) = p(y - x > \theta)$. Наконец, вычисляя вероятности противоположных событий, получим $p(y - x > \theta) = p(\theta > y - x)$, а это и есть определение θ как медианы для совокупности случайных величин $z_i = y_i - x_i$.

Таким образом, проверка нулевой гипотезы $H_0 : \theta = 0$ равносильна проверке гипотезы, согласно которой медиана случайной величины z равна нулю, и, аналогично, при альтернативной гипотезе $H_1 : \theta > 0$ медиана случайной величины z будет больше нуля. Предполагалась непрерывность вероятности $p(x, y)$, поэтому распределение случайной величины z непрерывно, т.е. вероятность совпадения $x_i = y_i$ равна нулю. Реально наблюдается всегда дискретная последовательность случайных величин, и могут быть случайные совпадения. Как поступать в этом случае – вопрос наименее теоретически обоснованный. Простейший выход - отбрасывать совпадающие наблюдения, сокращая при этом выборку.

Обозначим $z_i = y_i - x_i$ и примем модель

$$z_i = \theta + \varepsilon_i, \quad i = \overline{1, n}, \quad (8.2.1)$$

где ε_i - ненаблюдаемая случайная величина, θ - интересующий нас неизвестный параметр. При этом предполагается, что все ε_i - взаимно независимы и извлечены из непрерывной совокупности, имеющей медиану, равную нулю, т.е. $P(\varepsilon_i < 0) = P(\varepsilon_i > 0) = 1/2$, $i = \overline{1, n}$.

Проверим гипотезу $H_0 : \theta = 0$, определив для этого переменную -

счетчик $\psi_i = \begin{cases} 1, & z_i > 0, \\ 0, & z_i < 0. \end{cases}$ Положим $B = \sum_{i=1}^n \psi_i$. Статистика B есть число

положительных величин среди z_i , $i = \overline{1, n}$. Случайные величины ψ_i независимы и, в силу симметричности распределения относительно медианы, с ними можно связать схему последовательных независимых испытаний, в которой вероятность успеха $P(\psi_i = 1) = 0.5$ для каждого испытания. Сле-

довательно, при нулевой гипотезе H_0 их сумма B распределена по биномиальному закону с параметрами $B(n, p) = B(n, 1/2)$.

Пусть b - верхняя α -процентная точка биномиального распределения при объеме выборки n и вероятности p в схеме Бернулли. Введем обозначение $b = b(\alpha, n, p)$. Оно указывает на зависимость b от вероятности ошибки первого рода α . $b(\alpha, n, p)$ есть корень уравнения

$$P(B > b/n, p) = \sum_{i=b}^n C_n^i p^i (1-p)^{n-i} = \alpha. \quad (8.2.2)$$

Тогда процедура проверки гипотезы H_0 при уровне значимости α выглядит следующим образом.

1. Односторонний критерий для H_0 против альтернативы $H_1 : \theta > 0$:
отклонить H_0 , если $B \geq b(\alpha, n, 1/2)$,
принять H_0 , если $B < b(\alpha, n, 1/2)$.

Рис. 8.2 показывает критическую область правостороннего критерия для биномиального распределения.

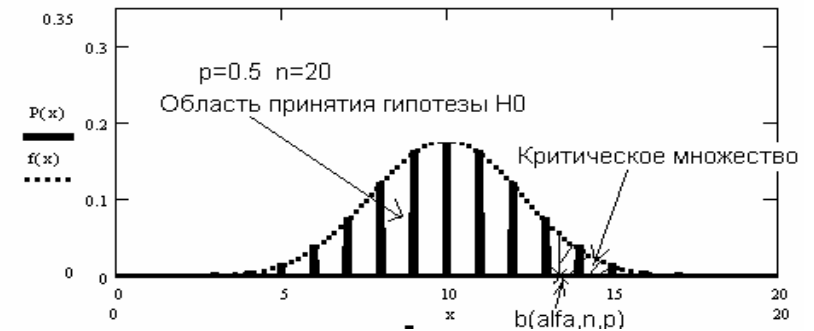


Рис. 8.2. Критическая область и область принятия решения для биномиального распределения

2. Односторонний критерий для H_0 против альтернативы $H_1 : \theta < 0$:
отклонить H_0 , если $B \leq [n - b(\alpha, n, 1/2)]$,
принять H_0 , если $B > [n - b(\alpha, n, 1/2)]$.

3. Двусторонний критерий для H_0 против альтернативы $H_1 : \theta \neq 0$:

отклонить H_0 , если $\begin{cases} B \leq [n - b(\alpha_1, n, 1/2)], \text{ или} \\ B \geq b(\alpha_2, n, 1/2), \end{cases}$

$$\text{принять } H_0, \text{ если } \begin{cases} n - b(\alpha_1, n, 1/2) < B < b(\alpha_2, n, 1/2), \\ \alpha = \alpha_1 + \alpha_2, \end{cases}$$

т.е. левый и правый хвосты распределения могут учитываться несимметрично.

Рассмотрим теперь приближения для большой выборки. Интегральная функция распределения для биномиального закона имеет вид

$$F(m, n, p) = \sum_{i=0}^m C_n^i p^i (1-p)^{n-i}, \quad 0 < p < 1, \quad m = 0, 1, \dots, n. \quad (8.2.3)$$

Именно по этой формуле вычислена функция распределения $P(x)$ на рис. 8.2.

В большинстве статистических приложений желательно иметь достаточно точную аппроксимацию для тех значений функции $F(m, n, p)$, которые принадлежат отрезкам $[0.005, 0.05]$ и $[0.93, 0.995]$. В этом случае условимся говорить, что аппроксимация осуществляется на хвостах распределения. Если же истинные значения аппроксимируемой функции $F(m, n, p)$ принадлежат отрезку $[0.05, 0.93]$, то будем использовать термин аппроксимация между хвостами распределения. При небольших значениях m и n значения функции (8.2.3) легко подсчитать непосредственным образом, но при больших m и n необходимо использовать нормальную аппроксимацию.

Для быстрых прикидочных расчетов рекомендуется следующая простая аппроксимация:

$$F(m, n, p) \approx \begin{cases} \Phi\left(\frac{2\sqrt{(m+1)(1-p)} - 2\sqrt{(n-m)p}}{\sqrt{(4m+3)(1-p)} - \sqrt{(4n-4m-1)p}}\right) & \text{на хвостах,} \\ \Phi\left(\frac{\sqrt{(4m+3)(1-p)} - \sqrt{(4n-4m-1)p}}{\sqrt{(4m+2.5)(1-p)} - \sqrt{(4n-4m-1.5)p}}\right) & \text{между хвостами.} \end{cases} \quad (8.2.4)$$

Здесь Φ - функция Лапласа. Более точная аппроксимация:

$$F(m, n, p) \approx \begin{cases} \Phi\left(\frac{\sqrt{(4m+3)(1-p)} - \sqrt{(4n-4m-1)p}}{\sqrt{(4m+2.5)(1-p)} - \sqrt{(4n-4m-1.5)p}}\right) & \text{на хвостах,} \\ \Phi\left(\frac{\sqrt{(4m+3)(1-p)} - \sqrt{(4n-4m-1)p}}{\sqrt{(4m+2.5)(1-p)} - \sqrt{(4n-4m-1.5)p}}\right) & \text{между хвостами.} \end{cases} \quad (8.2.5)$$

Наконец, для очень большой выборки применима интегральная теорема Муавра – Лапласа. Статистика $B^* = \frac{B - M(B)}{\sqrt{D(B)}} =$

$$= \frac{B - n/2}{\sqrt{n/4}} \in N(0,1) \text{ при } n \rightarrow \infty.$$

Приближение нормальной теории для одностороннего критерия для H_0 против альтернативы $H_1: \theta > 0$ таково:

отклонить H_0 , если $B^* \geq z(\alpha)$;

принять H_0 , если $B^* < z(\alpha)$, где $\int_{z(\alpha)}^{\infty} e^{-t^2/2} dt = \alpha$.

8.3. Критерий знаков для одномерной выборки

Описанная в предыдущем подразделе процедура легко может быть приспособлена для одномерной выборки. Пусть имеется n наблюдений z_1, z_2, \dots, z_n , причем все z_i взаимно независимы и все извлечены из одной и той же непрерывной генеральной совокупности с медианой θ , так что $P(z_i < 0) = P(z_i > 0) = 1/2$, $i = \overline{1, n}$. Для проверки $H_0 : \theta = \theta_0$, где θ_0 - некоторое заданное число, надо модифицировать наблюдения $z'_i = z_i - \theta_0$, $i = \overline{1, n}$. Затем к наблюдениям z'_i применяется вышеописанная процедура критерия знаков.

Очевидно, что метод построения двустороннего критерия знаков легко применяется для получения доверительного интервала для медианы θ с коэффициентом доверия не менее $1 - \alpha$. Действительно, с вероятностью $1 - \alpha$ истинное значение медианы покрывается случайным интервалом $\theta \in [n - b(\alpha_1, n, 1/2), b(\alpha_2, n, 1/2)]$, $\alpha = \alpha_1 + \alpha_2$. Ввиду дискретности биномиального распределения построить доверительный интервал, коэффициент доверия которого в точности равен $1 - \alpha$, в общем случае не удастся. Поэтому границы округляют, а за уровень ошибки первого рода α берут ближайшую к заданному значению вероятность из таблицы биномиального распределения с той или иной стороны, смотря по смыслу решаемой задачи.

Пример. Исследовалась геоморфология большой песчаной отмели пролива Виньярд в штате Массачусетс. Из различных мест отмели отобрали семь проб. Измерялась скорость сегментации (отложения осадка) песка при температуре 22°C. Обычно на пересечении гребней песчаных волн скорость сегментации равна 14 см/с. В следующей таблице даны скорости сегментации песка для семи проб.

Образец	1	2	3	4	5	6	7
z_i , см/с	12.9	13.7	14.5	13.3	12.8	13.8	13.4

Относятся ли эти семь наблюдений к тому месту отмели, где пересекаются гребни песчаных волн?

Решение. Необходимо проверить $H_0 : \theta = 14$ см/с против альтернативы $H_1 : \theta \neq 14$ см/с. Для этого модифицируем наблюдения.

Образец	1	2	3	4	5	6	7
$z'_t = z_i - 14.0$, см/с	-1.1	-0.3	0.5	-0.7	-1.2	-0.2	-0.6

$$\psi_i = \begin{cases} 1, & z'_i > 0, \\ 0, & z'_i < 0, \end{cases} \quad B = \sum_{i=1}^7 \psi_i = 0 + 0 + 1 + 0 + 0 + 0 + 0 = 1. \quad \text{Критерий}$$

здесь двусторонний.

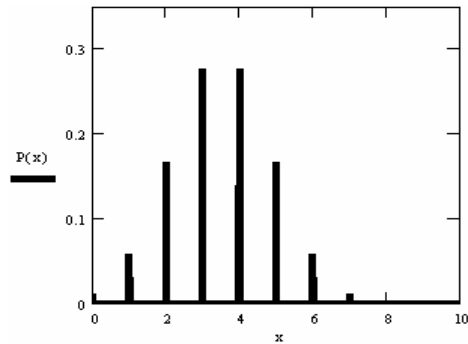


Рис. 8.3. Многоугольник биномиального распределения

При $\alpha_1 = \alpha_2 = 0.0078$
 $b(0.0078, 7, 1/2) = 7$, т.е. процентная точка вычисляется совершенно точно. Покажем это подробнее. Построим многоугольник распределения (рис. 8.3) и $F(m, n, p)$ для биномиального распределения с $n = 7$ и $p = 0.5$.

$$P(m=0) = C_7^0 (1/2)^7 = 0.0078,$$

$$P(m=1) = C_7^1 (1/2)^7 = 0.0055,$$

$$P(m=2) = C_7^2 (1/2)^7 = 0.1641,$$

$$P(m=3) = P(m=4) = C_7^3 (1/2)^7 = 0.2734, \quad P(m=5) = P(m=2) = C_7^5 (1/2)^7 = 0.1641,$$

$$P(m=6) = P(m=1) = C_7^6 (1/2)^7 = 0.0055, \quad P(m=7) = P(m=0) = C_7^7 (1/2)^7 = 0.0078.$$

Вычислим функцию распределения. $F(m) = \sum_{m_i < m} p_i$. Тогда,

например, $F(0) = 0$, $F(1) = P(m=0) = 0.0078$. Аналогично

$F(2) = P(m=0) + P(m=1) = 0.0625$ и так далее.

$$F(m) = \begin{cases} 0, & m \leq 0, \\ 0.0078, & 0 < m \leq 1, \\ 0.0625, & 1 < m \leq 2, \\ 0.2266, & 2 < m \leq 3, \\ 0.5, & 3 < m \leq 4, \\ 0.7734, & 4 < m \leq 5, \\ 0.9375, & 5 < m \leq 6, \\ 0.9922, & 6 < m \leq 7, \\ 1, & m > 7. \end{cases}$$

$P(B \geq b) = \alpha_1$, $P(B < b) = 1 - P(B \geq b) = 1 - \alpha_1$. Итак, критическая точка b выбирается из уравнения $P(B < b) = 1 - \alpha_1 = F(b)$. Мы не можем здесь выбрать, например, $\alpha = 0.05$ и $\alpha_1 = \alpha_2 = 0.025$, так как таких процентных точек нет. Выберем $\alpha = 0.0156$ и $\alpha_1 = \alpha_2 = 0.0078$. Тогда $P(B < b) = 1 - 0.0078 = 0.9922 = F(7)$, т.е. $b = 7$.

Для двустороннего критерия значимости принять или отвергнуть нулевую гипотезу можно, проверив неравенство $n - b(\alpha, n, 0.5) < B < b(\alpha, n, 0.5)$. В нашем случае это равносильно неравенству $7 - 7 = 0 < B < 7$. Так как $B = 1$, то гипотеза H_0 принимается на уровне значимости $\alpha = 0.0156$. Видно, что $b(0.0625, 7, 0.5) = 6$. Следовательно, наименьший уровень значимости, на котором мы могли бы отвергнуть гипотезу H_0 в пользу $H_1: \theta \neq 14$ см/с равен 0.1250.

Построим теперь соответствующий доверительный интервал. Для этого надо определить медиану выборки. Построим вариационный ряд $z^{(1)} \leq z^{(2)} \leq \dots \leq z^{(7)}$. Для данного примера вариационный ряд таков: 12.8, 12.9, 13.3, 13.4, 13.7, 13.8, 14.5. Медиана выборки равна

$$\hat{\theta} = \begin{cases} z^{\left[\frac{n}{2}\right]+1}, & n - \text{нечетное,} \\ \frac{1}{2} \left(z^{\left[\frac{n}{2}\right]} + z^{\left[\frac{n}{2}\right]+1} \right), & n - \text{четное.} \end{cases} \quad \text{В нашем случае}$$

$$\left[\frac{n}{2}\right] = 3, \quad \hat{\theta} = z^{(4)} = 13.4. \text{ Доверительный интервал для медианы определя-}$$

ется как $\theta_{\text{нижн.}} < \hat{\theta} < \theta_{\text{верхн.}}$, $\theta_{\text{нижн.}} = z^{(C_\alpha)}$, $\theta_{\text{верхн.}} = z^{(n+1-C_\alpha)}$, $P(C_\alpha \leq B \leq n - C_\alpha) = 1 - \alpha$. Действительно, последнее равенство $P(C_\alpha \leq B \leq n - C_\alpha) = 1 - \alpha$ определяет нахождение с заданной вероятностью случайной величины, распределенной биномиально в доверительном интервале, включая границы. Вспоминая формулу, определяющую границы двустороннего критерия для B , видим что $n - b(\alpha_1, n, 0.5) < B < b(\alpha_2, n, 0.5)$, $\alpha = \alpha_1 + \alpha_2$. Прибавляя единицу в правую и левую части неравенства для превращения его в равенство, имеем $n - b(\alpha_1, n, 0.5) + 1 = C_\alpha$, т.е. $n - C_\alpha + 1 = b(\alpha_1, n, 0.5)$. В нашем случае $C_\alpha = 7 - 7 + 1 = 1$, $\alpha = 0.0156$. $n + 1 - C_\alpha = 7 + 1 - 1 = 7$.

Итак, $\theta_{\text{нижн.}} = z^{(C_\alpha)} = z^{(1)} = 12.8$, $\theta_{\text{верхн.}} = z^{(n+1-C_\alpha)} = z^{(7)} = 14.5$ с уровнем значимости $\alpha = 0.0156$. Если взять, например, $\alpha = 0.1250$, то $C_\alpha = 7 - 6 + 1 = 2$ и $z^{(2)} < \hat{\theta} < z^{(6)}$, т.е. $12.9 < \hat{\theta} < 13.8$.

Несколько замечаний о свойствах критерия знаков. Асимптотическая эффективность одновыборочных непараметрических методов, основанных на статистике знаков B , по отношению к их соперникам из нормальной

теории, основанным на средней $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$, выражается величиной

$I_\alpha(F)$. Она никогда не бывает меньше $1/3$ и может быть бесконечно большой. Например:

Распределение F	$I_\alpha(F)$
Нормальное	0.637
Равномерное	0.333
Двустороннее экспоненциальное	2.000

8.4. Ранговый критерий (одновыборочный критерий Вилкоксона)

Рассмотрим анализ повторных парных наблюдений с помощью знаковых рангов. В этом случае, как и в предыдущем, проверяется гипотеза о сдвиге. Предположения аналогичны, сделанным в подразд. 8.2.

Пусть мы имеем $2n$ наблюдений, по два наблюдения на каждый из n объектов. Обозначим $z_i = y_i - x_i$ и примем модель $z_i = \theta + \varepsilon_i$, $i = \overline{1, n}$, где что все ε_i взаимно независимы и извлечены из непрерывной совокупности (не обязательно одной и той же), которая симметрична относительно нуля.

Основная гипотеза $H_0 : \theta = 0$, которая может быть сформулирована и в терминах функции распределения. Ведь, если сдвига нет, то $F_1(x) \equiv F_2(y)$, иначе либо $F_1(x) > F_2(y)$, либо $F_1(x) < F_2(y)$. Итак, $H_0 : F_1(x) \equiv F_2(y)$ - аналогичная по смыслу формулировка основной гипотезы. Последовательность действий при проверке этой гипотезы такова.

1. Составим из данных двух выборок общий вариационный ряд из абсолютных значений наблюдений. Каждому члену вариационного ряда припишем ранг R_i , равный порядковому номеру члена в общем вариационном ряду $|z_1|, |z_2|, \dots, |z_n|$.

2. Определим переменную - счетчик $\psi_i, i = \overline{1, n}, \psi_i = \begin{cases} 1, & z_i > 0, \\ 0, & z_i < 0, \end{cases}$

$$r_i = \psi_i R_i.$$

3. Выпишем статистику рангового критерия

$$T^+ = \sum_{i=1}^n \psi_i R_i = \sum_{i=1}^n r_i. \quad (8.4.1)$$

Статистика T^+ равна сумме положительных знаковых рангов. Рациональность предложенной процедуры состоит в том, что если одно распределение смещено относительно другого, то это должно проявиться в том, что маленькие ранги должны в основном соответствовать одной выборке, а большие – другой, вследствие чего соответствующие суммы рангов должны быть маленькими или большими в зависимости от того, какая альтернатива имеет место. Естественно ожидать, что при нулевой гипотезе о симметричности распределения относительно нуля любой ранг может с одинаковым успехом получить как знак «+», так и знак «-», в силу чего существует 2^n разных последовательностей рангов. Кроме того, если нулевая гипотеза справедлива, то в полученной последовательности рангов со знаками количество рангов со знаком «+» не должно значительно отличаться от количества рангов со знаком «-». Напротив, если гипотеза H_1 имеет место, то должно наблюдаться значимое превышение количества рангов со знаком «+» над количеством рангов со знаком «-», что подсказывает выбрать в качестве статистики критерия величину T^+ , равную сумме рангов со знаком «+». p -значение критерия, построенного на статистике T^+ , равно вероятности того, что сумма рангов T^+ примет значение, не меньшее наблюдаемой суммы.

Рассмотрим простейший пример при $n = 3$. Обозначим через B число положительных наблюдений, т.е. $B = \sum_{i=1}^n \psi_i$, а $r_i = \psi_i R_i$.

B	r_1, r_2, \dots, r_B	$P(r_1, r_2, \dots, r_B)$	$T^+ = \sum_{i=1}^n r_i$
0	нет	$\frac{1}{8}$	0
1	$r_1 = 1$	$\frac{1}{8}$	1

B	r_1, r_2, \dots, r_B	$P(r_1, r_2, \dots, r_B)$	$T^+ = \sum_{i=1}^n r_i$
1	$r_1 = 2$	$\frac{1}{8}$	2
1	$r_1 = 3$	$\frac{1}{8}$	3
2	$r_1 = 1, r_2 = 2$	$\frac{1}{8}$	3
2	$r_1 = 1, r_2 = 3$	$\frac{1}{8}$	4
2	$r_1 = 2, r_3 = 3$	$\frac{1}{8}$	5
3	$r_1 = 1, r_2 = 2, r_3 = 3$	$\frac{1}{8}$	6

Вероятность элементарного исхода в этой схеме $1/2^n$. Непосредственный подсчет вероятностей типа $P(T^+ = m)$ в этой схеме, похожей на схему случаев, затруднителен. Поэтому пользуются специальными таблицами или нормальной аппроксимацией. Из приведенной таблицы, например, $P(T^+ \geq 5) = P(T^+ = 5) + P(T^+ = 6) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$.

Путем довольно несложных вычислений можно получить

$$M(T^+) = \frac{n(n+1)}{4}, \quad D(T^+) = \frac{n(n+1)(2n+1)}{24}, \quad (8.4.2)$$

поэтому статистика $T^* = \frac{T^+ - \frac{n(n+1)}{4}}{\left[\frac{n(n+1)(2n+1)}{24} \right]^{1/2}} \in N(0,1)$ при $n \rightarrow \infty$ и если

среди случайных величин $|z_1|, |z_2|, \dots, |z_n|$ не было совпадений. При наличии t совпадений ранги $R_j + 1, R_j + 2, \dots, R_j + t$ совпавших наблюдений следует заменить их средним арифметическим. При такой замене сумма рангов остается без изменений, а следовательно, и первая формула (8.4.2). Сумма же квадратов рангов уменьшится при этом на величину

$(1/12)t(t-1)(t+1)$. Учитывая это, получаем, что в случае наличия t совпадений

$$D(T^+) = \frac{n(n+1)(2n+1)}{24} - \frac{(t-1)t(t+1)}{48}. \quad (8.4.3)$$

Сформулируем теперь три вида критериев.

1. Для одностороннего критерия $H_0 : \theta = 0$ против альтернативы $\theta > 0$ при уровне значимости α :

отклонить H_0 , если $T^+ \geq t(\alpha, n)$,

принять H_0 , если $T^+ < t(\alpha, n)$, где $P(T^+ \geq t(\alpha, n)) = \alpha$, т.е. $t(\alpha, n)$ - $\alpha\%$ -ная критическая точка T^+ -распределения (вероятность верхнего хвоста распределения статистики знаковых рангов Вилкоксона).

2. Для $H_0 : \theta = 0$ против $H_1 : \theta < 0$:

отклонить H_0 , если $T^+ \leq \frac{n(n+1)}{2} - t(\alpha, n)$;

принять H_0 , если $T^+ > \frac{n(n+1)}{2} - t(\alpha, n)$, где $\frac{n(n+1)}{2} - t(\alpha, n) = \max T^+$.

3. Для двустороннего критерия $H_0 : \theta = 0$ против альтернативы $H_1 : \theta \neq 0$ при уровне значимости α :

отклонить H_0 , если $\begin{cases} T^+ \geq t(\alpha_1, n), \\ T^+ \leq \frac{n(n+1)}{2} - t(\alpha_1, n); \end{cases}$

принять H_0 , если $\frac{n(n+1)}{2} - t(\alpha_1, n) < T^+ < t(\alpha_2, n)$, $\alpha = \alpha_1 + \alpha_2$.

Если пользоваться нормальной аппроксимацией, то, например, правосторонний критерий выглядит так:

отклонить H_0 , если $T^+ \geq z_\alpha$,

принять H_0 , если $T^+ < z_\alpha$, где z_α - $\alpha\%$ -ная точка стандартного нормального распределения.

Для проверки гипотезы $H_0 : \theta = \theta_0$, где θ_0 - заданное число, неравное нулю, получаем модифицированные наблюдения $z_i' = z_i - \theta_0$ и далее вычисляем T^+ , используя z_i' вместо z_i . Таким образом, описанная процедура может быть применена к данным одной выборки.

Пример. Семь наблюдений представляют собой семь усредненных значений - измерений θ - отношения массы Земли к массе Луны, полу-

ченные семью различными космическими кораблями. На основании данных, ранее полученных с космического корабля «Рейнджер», специалисты считали θ равным 81.3035. Проверить гипотезу $H_0 : \theta = 81.3035$ против альтернативы $H_1 : \theta \neq 81.3035$.

Космический корабль	Мари-нер-4 (Венера)	Мари-нер-4 (Марс)	Мари-нер-5 (Венера)	Мари-нер-6 (Марс)	Мари-нер-7 (Марс)	Пионер-6	Пионер-7
θ	81.3001	81.3015	81.3006	81.3011	81.2997	81.3005	81.3021

Решение. Модифицируем наблюдения z_i .

i	1	2	3	4	5	6	7
z_i	81.3001	81.3015	81.3006	81.3011	81.2997	81.3005	81.3021
$z_i' = z_i - 81.3035$	-0.0034	-0.0020	-0.0029	-0.0024	-0.0038	-0.0030	-0.0014
ψ_i	0	0	0	0	0	0	0

Поэтому $T^+ = \sum_{i=1}^n R_i \psi_i = 0$, так как все $\psi_i = 0$. Следовательно, ранги

можно не считать. Приведем выписку из таблицы распределения T^+ -статистики. Видно, что если выбрать $\alpha = \alpha_1 + \alpha_2 = 0.078$, то $t(0.039, 7) = 25$, $\alpha_1 = \alpha_2 = 0.039$. В таблице даны значения $P(T^+ \geq x) = \alpha$ (таблица взята из книги [24]).

x	15	16	17	18	19	20	21
α	0.469	0.406	0.344	0.289	0.234	0.188	0.148
x	22	23	24	25	26	27	28
α	0.109	0.078	0.055	0.039	0.023	0.016	0.008

Поскольку критерий двусторонний, то процедура проверки нулевой гипотезы изложена в п. 3.

Отклонить H_0 , если $\begin{cases} T^+ \geq t(\alpha_1, n), \\ T^+ \leq \frac{n(n+1)}{2} - t(\alpha_1, n). \end{cases}$

В нашем случае $\frac{n(n+1)}{2} - t(\alpha, n) = \frac{56}{2} - 25 = 3$, $T^+ = 0 < 3$, т.е. гипотеза $H_0 : \theta = 81.3035$ отклоняется на уровне значимости $\alpha = 0.078$.

Вспользуемся теперь нормальной аппроксимацией

$$T^* = \frac{0 - (7 \cdot 8)/4}{\sqrt{(7 \cdot 8 \cdot 15)/24}} = -2.366.$$

При уровне значимости

$\alpha_1 = 0.0091$ $z_{\alpha_1} = -2.36$, т.е. наименьший уровень значимости, при котором отвергается H_0 , равен $\alpha = \alpha_1 + \alpha_2 = 0.018$. Если же взять $\alpha_1 = 0.039$, то $z_{0.039} = -1.76$. Так как $T^* < z_{0.039}$, т.е. T^* попадает в критическую область на левом хвосте нормального распределения, то гипотеза H_0 отклоняется.

Доверительный интервал, основанный на статистике рангов, строится несколько сложнее. Для этого вводятся дополнительные статистики $W_i = (z_i + z_j)/2$ для всех $i, j = \overline{1, n}$. Число этих статистик $M = n(n+1)/2$. Далее строится вариационный ряд из этих статистик, и именно по нему находится оценка медианы стандартным способом. Проведем эту процедуру. $M = (7 \cdot 8)/2 = 28$. Каждая вторая строка следующей таблицы содержит статистику $W_i = (z_i + z_j)/2$ для соответствующих значений индексов i и j .

$i = 1, j = 1$ 81.3001	$i = 1, j = 2$ 81.3008	$i = 1, j = 3$ 81.3035	$i = 1, j = 4$ 81.3006	$i = 1, j = 5$ 81.2999	$i = 1, j = 6$ 81.3003	$i = 1, j = 7$ 81.3011
	$i = 2, j = 2$ 81.3015	$i = 2, j = 3$ 81.30105	$i = 2, j = 4$ 81.3013	$i = 2, j = 5$ 81.3006	$i = 2, j = 6$ 81.3010	$i = 2, j = 7$ 81.3018
		$i = 3, j = 3$ 81.3006	$i = 3, j = 4$ 81.30085	$i = 3, j = 5$ 81.30015	$i = 3, j = 6$ 81.30055	$i = 3, j = 7$ 81.30135
			$i = 4, j = 4$ 81.3011	$i = 4, j = 5$ 81.3004	$i = 4, j = 6$ 81.3008	$i = 4, j = 7$ 81.3016
				$i = 5, j = 5$ 81.2997	$i = 5, j = 6$ 81.3001	$i = 5, j = 7$ 81.3009
					$i = 6, j = 6$ 81.3005	$i = 6, j = 7$ 81.3013
						$i = 7, j = 7$ 81.3021

Составим из полученных статистик W_i вариационный ряд.

81.2997	81.2999	81.3001	81.3001	81.30015	81.3003	81.30035
81.3004	81.3005	81.30055	81.3006	81.3006	81.3006	81.3008
81.3008	81.30085	81.3009	81.3010	81.30105	81.3011	81.3011
81.3013	81.3013	81.30135	81.3015	81.3016	81.3018	81.3021

Здесь $n = 28$, $\left[\frac{n}{2}\right] = 14$, так как n - четное, то $\hat{\theta} = \frac{1}{2}[W^{(14)} + W^{(15)}] =$
 $= \frac{1}{2}(81.3008 + 81.3008) = 81.3008$. Далее идут стандартные действия, т.е.
 действия, аналогичные тем, какие производились при построении довери-
 тельного интервала по критерию знаков.

$P_{\theta}\left(C_{\alpha} \leq \theta \leq \left[\frac{n(n+1)}{2} - C_{\alpha}\right]\right) = 1 - \alpha$ - доверительный интервал для ме-
 дианы и $\theta_{\text{нижн.}} = W^{(C_{\alpha})}$, $\theta_{\text{верхн.}} = W^{(M+1-C_{\alpha})}$, причем
 $M+1-C_{\alpha} = t\left(\frac{\alpha}{2}, n\right)$. Если оставить то же $\alpha = 0.078$, то $t(0.039, 7) = 25$ и
 $C_{\alpha} = M+1 - t\left(\frac{\alpha}{2}, n\right) = 28+1-25 = 4$, $M+1-C_{\alpha} = 28+1-4 = 25$, т.е.
 $W^{(4)} \leq \theta \leq W^{(25)}$ или $81.3001 \leq \theta \leq 81.3015$. Это $(1-0.078) \cdot 100\%$ - довери-
 тельный интервал.

8.5. Двухвыборочный ранговый критерий Вилкоксона

Этот критерий предназначен для проверки нулевой гипотезы H_0 , со-
 гласно которой двум независимым выборкам объемов n и m отвечают
 одинаковые функции распределения $F_1(x) \equiv F_2(y)$, против односторонней
 альтернативы H_1 , по которой либо $F_1(x) < F_2(y)$, либо $F_1(x) > F_2(y)$, или
 против двусторонней альтернативы $F_1(x) \neq F_2(y)$.

Нулевая гипотеза может быть сформулирована в терминах сдвига од-
 ной выборки относительно другой, так же как в предыдущем подразделе.
 При проверке нулевой гипотезы следует выполнить следующие действия.

1. Расположить выборочные значения обеих выборок в порядке воз-
 растания, т.е. образовать общий вариационный ряд, и каждой величине из
 этого ряда сопоставить ее ранг R_i , равный порядковому номеру величины
 в общем вариационном ряду. Заметим, что если H_0 справедлива, то лю-
 бое распределение по этим двум выборкам равновероятно, а общее число
 способов группирования рангов равно C_{n+m}^m .

2. В качестве статистики критерия берут сумму рангов W одной (на-
 пример, второй) выборки, т.е.

$$W = \sum_{j=1}^m R_j. \quad (8.5.1)$$

3. Подсчитываются все различные способы группирования рангов, при которых статистика W принимает значения, равные или меньшие наблюдаемого, после чего вычисляется отношение этого числа к общему числу возможных распределений рангов по двум выборкам C_{n+m}^m . Полученное отношение дает одностороннее p -значение критерия.

При малых значениях n и m относительно легко вычислить p -значение, но для выборок большого объема строят приближенный критерий, основанный на асимптотическом распределении статистики W . Именно

$$M(W) = \frac{m(n+m+1)}{2}, \quad D(W) = \frac{nm(n+m+1)}{12}. \quad \text{Тогда статистика}$$

$$W^* = \frac{W - M(W)}{\sqrt{D(W)}} = \frac{W - \frac{m(n+m+1)}{2}}{\left[\frac{nm(n+m+1)}{12} \right]^{\frac{1}{2}}} \in N(0,1) \text{ при } n, m \rightarrow \infty.$$

Это приближение не дает хорошей точности при $n, m \leq 50$. По этой причине следует пользоваться аппроксимацией Имана [27]:

$$J = \frac{W^*}{2} \left[1 + \left(\frac{n+m-2}{n+m-1 - (W^*)^2} \right)^{\frac{1}{2}} \right], \quad (8.5.2)$$

$\alpha\%$ -ные точки для которой равны $J_{\alpha, n+m-2} = \frac{1}{2} z_{\alpha} + \frac{1}{2} t_{\alpha, n+m-2}$. Здесь z_{α} - $\alpha\%$ -ная точка стандартного нормального распределения, $t_{\alpha, n+m-2}$ - $\alpha\%$ -ная точка распределения Стьюдента с $n+m-2$ степенями свободы.

Если среди наблюдений есть одинаковые, то надо работать со средними рангами. В этом случае при использовании нормальной аппроксимации в формулу (8.5.2) должна быть введена поправка. Эта поправка, как показано в подразд. 8.4, изменит только оценку дисперсии статистик W или J .

При наличии t совпадений формула для $D(W)$ имеет следующий вид:

$$D(W) = \frac{nm}{12} \left[n+m+1 - \frac{\sum_{j=1}^g t_j(t_j^2-1)}{(n+m)(n+m-1)} \right], \quad (8.5.3)$$

где g - число групп совпадений, t_j - объем j -й группы. В формуле (8.5.3), если наблюдение не совпадает ни с каким другим, оно рассматривается как отдельная группа. Поэтому если в ранжировке нет совпадений, то $g = n + m$, $t_j = 1$, $j = 1, 2, \dots, n + m$, и правая часть (8.5.3) сводится к $\frac{nm(n+m+1)}{12}$.

Три основных вида критериев значимости для данного критерия можно сформулировать в следующей форме.

1. Для одностороннего критерия $H_0 : F_1(x) \equiv F_2(y)$ против альтернативы $H_1 : F_1(x) < F_2(y)$ на уровне значимости α :

отклонить H_0 , если $W \geq w(\alpha, m, n)$;

принять H_0 , если $W < w(\alpha, m, n)$, где константа $w(\alpha, m, n)$ удовлетворяет условию $P[W \geq w(\alpha, m, n)] = \alpha$. Значения $w(\alpha, m, n)$ табулированы. Обширные таблицы критических точек распределения статистики W опубликованы в [28].

2. Для одностороннего критерия $H_0 : F_1(x) \equiv F_2(y)$ против альтернативы $H_1 : F_1(x) > F_2(y)$:

отклонить H_0 , если $W \leq m(n+m+1) - w(\alpha, m, n)$;

принять H_0 , если $W > m(n+m+1) - w(\alpha, m, n)$.

3. Для двустороннего критерия $H_0 : F_1(x) \equiv F_2(y)$ против альтернативы $H_1 : F_1(x) \neq F_2(y)$:

отклонить H_0 , если $\begin{cases} W \geq w(\alpha_2, m, n) \text{ или} \\ W \leq m(n+m+1) - w(\alpha_1, m, n), \end{cases}$

принять H_0 , если $m(n+m+1) - w(\alpha_1, m, n) < W < w(\alpha_2, m, n)$, $\alpha = \alpha_1 + \alpha_2$.

Пример. В биохимическом исследовании, проведенном методом меченых атомов, по результатам изучения 8 препаратов опытной серии и 5 препаратов контрольной серии получены следующие показания счетчика импульсов (в импульсах в минуту):

Опыт	340	343	322	349	332	320	313	304
Контроль	318	321	318	301	312	-	-	-

Можно ли считать, что полученные значения опытной и контрольной серий различны? Принять $\alpha = 0.1$.

Решение. Составим вариационный ряд, отмечая принадлежность элемента к контрольной серии чертой снизу.

Эле- мент	<u>301</u>	304	<u>312</u>	313	<u>318</u>	<u>318</u>	320	<u>321</u>	322	332	340	343	349
Ранг	1	2	3	4	5.5	5.5	7	8	9	10	11	12	13

$W = \sum_{j=1}^5 R_j = 1 + 3 + 5.5 + 5.5 + 8 = 23$. Имеется одна группа совпаде-

ний, т.е. $g = 1, t_1 = 2$. Тогда $M(W) = \frac{5(5+8+1)}{2} = 35$,

$$D(W) = \frac{5 \cdot 8}{12} \left[5 + 8 + 1 - \frac{2(4-1)}{(5+8)(5+8-1)} \right] = 46.538.$$

Воспользуемся аппроксимацией Имана, так как n и m малы. При этом

$$W^* = \frac{23 - 35}{\sqrt{46.538}} = -1.759, J = \frac{-1.759}{2} \left[1 + \left(\frac{5 + 8 - 2}{5 + 8 - 1 - (-1.759)^2} \right)^{1/2} \right] = -1.857.$$

По таблицам нормального распределения и распределения Стьюдента находим: $z_{0.1} = 1.280, t_{0.1,11} = 1.363$. Тогда $\frac{1}{2}(z_\alpha + t_{\alpha, n+m-2}) = 1.322$.

Так как при упорядочении двух выборок, все наблюдения второй оказались сильно сдвинуты в начало общего вариационного ряда, проверим:

$$H_0 : F_1(x) \equiv F_2(y) \text{ против альтернативы}$$

$$H_1 : F_1(x) > F_2(y).$$

Таким образом, выбран левосторонний критерий значимости. Учитывая симметричность нормального распределения и распределения Стьюдента, получим $J_{0.1,11} = -1.322$. Тогда $J = -1.857 < J_{0.1,11}$ и, следовательно, $J \in \omega$. Таким образом, нулевая гипотеза H_0 должна быть отвергнута с уровнем значимости $\alpha = 0.1$, т.е. полученные значения показаний счетчиков в опытной и контрольной партиях различны.

8.6. Лабораторная работа № 10. Критерии знаков и рангов в пакете MATHCAD

Одно из главных достоинств критерия знаков – его простота и очень скромные требования к первоначальному статистическому материалу. Критерий знаков чаще всего используется для проверки гипотезы об однородности наблюдений внутри каждой пары в парных выборках, однако его можно применять и к одномерной выборке для проверки гипотезы о положении медианы $H_0 : \theta = \theta_0$.

Запрограммируем критерий знаков в пакете MATHCAD, решив с его помощью следующую задачу.

В эксперименте по искусственному стимулированию дождя были измерены дождевые осадки в течение 16 пар дней, причем в каждой паре один день облака засеивали стимулятором, а в другой день нет. Для каждой пары день засеивания выбирали случайным образом. В следующей таблице приведены количества выпавших осадков, измеренные специальным прибором за эти 16 пар дней.

Номер пары	1	2	3	4	5	6	7	8
Засеивание	0	2.09	0.07	0.30	0	2.55	1.62	0
Без засеивания	1.37	0	0	0.10	0.44	0	1.01	0.54
Номер пары	9	10	11	12	13	14	15	16
Засеивание	0	1.87	2.50	3.15	0.15	2.96	0	0
Без засеивания	0	0.62	0	5.54	0.01	0	0	0.75

Проверить нулевую гипотезу, согласно которой засеивание не оказывает эффекта.

Перейдем к одномерной выборке. Модернизируем наблюдения по формуле $z_i = x_i - y_i$ и вычислим статистику $\psi_i = \begin{cases} 1, & z_i > 0, \\ 0, & z_i < 0, \end{cases}$ так как мы, очевидно, будем проверять нулевую гипотезу вида $H_0 : \theta = 0$.

z_i	-1.37	2.09	0.07	0.20	-0.44	2.55	0.51	-0.54
ψ_i	0	1	1	1	0	1	1	0
z_i	0	1.25	2.50	-2.39	0.14	2.96	0	-0.75
ψ_i	-	1	1	0	1	1	-	0

Два наблюдения совпадают, следовательно, для них статистика ψ_i не определена. Отбросим эти совпадающие наблюдения, уменьшив объем выборки до $n = 14$.

ORIGIN:=1

$n := 14$ $p := 0.5$ $B := 9$ $\alpha := 0.05$ $\alpha_1 := \frac{\alpha}{2}$ $\beta := 1 - \alpha_1$ $n = 14$ $p = 0.5$ $\alpha_1 = 0.025$
 $\beta = 0.975$ $b := \text{qbinom}(\beta, n, p)$ $b = 11$ $b_1 := \text{qbinom}(\alpha_1, n, p)$ $b_1 = 3$

Альтернативную гипотезу сформулируем в виде $H_1 : \theta \neq 0$, тогда, так как $b_1 < B < b$, нулевую гипотезу следует принять по двустороннему критерию с уровнем значимости $\alpha = 0.05$.

Воспользуемся теперь аппроксимацией для приближения к нормальной теории. В случае двустороннего критерия будем иметь

$$zright := \text{qnorm}(\beta, 0, 1) \quad zleft := -zright \quad zleft = -1.96 \quad zright = 1.96$$

Поскольку опять $zleft < B < zright$, гипотезу H_0 нельзя отвергнуть, т.е. искусственная стимуляция дождя не оказывает эффекта.

Одностороннее p -значение критерия знаков определим по формуле (8.2.5).

$$arm := \sqrt{(4 * B + 3) * (1 - p)} - \sqrt{(4 * n - 4 * B - 1) * p} \quad arm = 1.334$$

$$pValue := 1 - \text{pnorm}(arm, 0, 1) \quad pValue = 0.091$$

Для вычисления рангов элементов выборки и расчета статистики критерия T^+ по формуле (8.4.1) воспользуемся следующей подпрограммой.

```

statT(x) :=
  n ← rows(x)
  for i ∈ 1..n
    yi ← xi
  for i ∈ 1..n-1
    for j ∈ i+1..n
      a ← yi
      if |yj| < |yi|
        yi ← yj
        yj ← a
      j ← 0
    for i ∈ 1..n
      continue if |yi| < 10-5
      j ← j + 1
      ψj ← if(yi < 0, 0, 1)
      zj ← |yi|
    T ← 0
    for i ∈ 1..j
      T ← T + ψi * i
  [ T ]
  [ z ]

```

x :=

-1.37
2.09
0.07
0.20
-0.44
2.55
0.51
-0.54
0
1.25
2.50
-2.39
0.14
2.96
0
-0.75

Распределение $t(\alpha, n)$ статистики рангов T^+ найдем с помощью нормальной аппроксимации. Для этого вычислим математическое ожидание и дисперсию T^+ :

$$T := \text{statT}(x) \quad z := \text{statT}(x)_2 \quad n1 := \text{rows}(z) \quad n1 = 14 \quad T = 68$$

$$MT := n1 * \frac{n1 + 1}{4} \quad Mt = 52.5 \quad DT := MT * \frac{2 * n + 1}{6} \quad DT = 253.75$$

$$T1 := \frac{(T - MT)}{\sqrt{DT}} \quad T1 = 0.973 \quad pValue := 1 - \text{pnorm}(T1, 0, 1)$$

$$pValue = 0.165$$

	1
1	0.07
2	0.14
3	0.2
4	0.44
5	0.51
6	0.54
7	0.75
8	1.25
9	1.37
10	2.09
11	2.39
12	2.5
13	2.55
14	2.96

Итак, поскольку статистика T_1 опять находится в пределах 95%-й области принятия решений двустороннего критерия $z_{left} = -1.96 < T_1 = 0.973 < z_{right} = 1.96$, гипотезу H_0 следует принять.

Задание № 1. Решить следующие задачи с помощью критерия знаков и одновыборочного рангового критерия Вилкоксона. Везде принять $\alpha = 0.05$.

1. Предполагается, что один из двух приборов, определяющих скорость автомобиля, имеет систематическую ошибку. Для проверки этого предположения определили скорость 10 автомобилей, причем скорость каждого фиксировалась одновременно двумя приборами. В результате получены следующие данные:

V_1 км / ч	70	85	63	54	65	80	75	95	52	55
V_2 км / ч	72	86	62	55	63	80	78	90	53	57

Позволяют ли эти результаты утверждать, что второй прибор действительно дает завышенные значения скорости?

2. Приводится время (в секундах) решения контрольных задач одиннадцатую учащимися до и после специальных упражнений по устному счету. Можно ли считать, что эти упражнения улучшили способности учащихся в решении задач?

До упражнений	87	61	98	90	93	74	83	72	81	75	83
После упражнений	50	45	79	90	88	65	52	79	84	61	52

3. Для 10 человек была предложена специальная диета. После двухнедельного питания по этой диете масса их тела изменилась следующим образом:

Масса до диеты (кг)	68	80	92	81	70	79	78	66	57	76
Масса после диеты (кг)	60	84	87	79	74	71	72	67	57	60

Можно ли рекомендовать эту диету для людей, желающих похудеть?

4. Сравнивалось действие двух экстрактов вируса табачной мозаики. Для этого каждая из половин листа натиралась соответствующим препаратом. Число мест приводится в таблице.

Экстракт А	20	39	43	13	28	26	17	49	36
Экстракт В	31	22	45	6	21	13	17	46	31

Можно ли считать, что действие этих экстрактов различно?

5. Изучалось влияние черного и апрельского пара на урожай ржи. Опыт длился шесть лет. Учитывалась масса 1000 зерен в граммах. Результаты опыта следующие:

Год посева	1	2	3	4	5	6
По черному пару	31.1	24.0	24.6	28.6	29.1	30.1
По апрельскому пару	31.6	24.2	24.8	19.1	29.9	31.0

Можно ли считать, что урожай ржи по апрельскому пару значимо выше, чем по черному?

6. Проверить предположение о том, что предлагаемый лечебный препарат не меняет состав крови, если препарат испытывался на десяти особях, а текущий анализ крови дал следующие результаты: 0.97, 1.05, 1.09, 0.88, 1.01, 1.14, 1.03, 1.07, 0.94, 1.02. Числа выражают отношение числа лейкоцитов в опыте к числу лейкоцитов в норме.

7. Изменение урожайности при применении одного из видов предпосевной обработки семян характеризуется следующими данными (в центнерах с гектара):

Год	1972	1973	1974	1975	1976	1977	1978	1979	1980
Необработанные семена	20.0	17.9	20.6	22.0	21.4	23.8	21.4	19.8	18.4
Обработанные семена	22.1	18.5	19.4	22.1	21.7	24.9	21.6	20.3	18.3

Можно ли считать, что предпосевная обработка увеличивает урожайность?

8. Измерялось напряжение пробоя у диодов, отобранных случайным образом из двух партий. Результаты измерения (в вольтах) следующие:

1-я партия	39	50	61	67	40	40	54
2-я партия	60	53	42	41	40	54	63

Можно ли считать, что у диодов из второй партии напряжение пробоя выше, чем у диодов из первой партии?

9. Двум группам испытуемых предлагалось провести опознание трех начертаний цифры 5. Результаты эксперимента (в секундах) следующие:

1-я группа	25	28	27	29	26	24	28	23	30	25	26
2-я группа	18	19	31	32	17	15	41	35	38	13	14

Можно ли считать, что время опознания для первой и второй групп различны?

10. В течение некоторого времени суточная производительность двух автоматов характеризуется следующими данными:

1-й автомат	105	60	83	111	138	71	87	130	93	105
2-й автомат	172	45	51	155	117	103	82	93	31	51

Можно ли считать, что суточная производительность этих двух автоматов различна?

11. Контролируемый размер нескольких деталей был проверен до и после наладки станка. В результате получены следующие данные (в мм):

До наладки	36.4	37.5	36.9	37.6	38.1	35.5	37.8	38.3	36.6
После наладки	36.8	39.2	37.6	39.9	39.6	34.2	36.5	36.3	39.8

Изменилась ли измеряемая величина контролируемого размера после наладки станка?

12. Для контроля настройки двух станков-автоматов, производящих детали по одному чертежу, определили отклонения от номинальных размеров у нескольких деталей, изготовленных на обоих станках. В результате получили следующие данные (в мкм):

Станок А	44	-14	32	8	-50	20	-35	15	10	-8	-20	5
Станок В	52	-49	61	-35	-48	18	-45	35	28	21	-59	-19

Различно ли отклонение от номинальных размеров у этих двух станков-автоматов?

13. Изучалось влияние пищевой добавки на увеличение массы тела кроликов. Опыт длился 7 недель. Исходная масса особей находилась в пределах от 500 до 600 грамм. За время опыта у животных наблюдались следующие прибавки в весе (за одну неделю):

Контрольные	560	580	600	420	530	490	580
Опытные	692	700	621	640	561	680	630

Можно ли утверждать, что пищевая добавка дает прибавку массы тела?

14. По выборкам из двух партий микросхем после операции легирования поликремния измерялось удельное сопротивление. Результаты замеров следующие:

1-я партия	52.2	33	76	32.5	49.5	32.5	191.5	112.5	52.9	114.8	33.7	69.1
2-я партия	119	17.5	43.5	43.5	90.5	40	50	108	62.4	16.5	97.5	96

Одинаково ли удельное сопротивление в обеих партиях?

15. У двух партий приборов измерялась глубина слоя диффузии (в мкм) после напыления рабочей поверхности. Можно ли считать, что глубина слоя диффузии у приборов из обеих партий различна?

1-я партия	9.8	9.8	8.6	8.6	9.2	9.2	9.8	9	10	9.4	9	11.2	10.8
2-я партия	8.6	9.2	10.4	9	9.8	9.2	9.6	10	9.8	9	9.8	8.7	8.6

16. Длина тела личинок шелкуна, обитающих в посевах ржи и проса (в мм), варьируется следующим образом:

В посевах ржи	7	10	14	15	12	16	12
В посевах проса	11	12	16	13	18	15	13

На основании этих проб создается впечатление о более крупных размерах личинок шелкунов, обитающих в просе. Проверить это предположение.

17. У полевых транзисторов измерялась характеристика: емкость затвор-сток. Увеличилась ли величина емкости затвор-сток у транзисторов, изготовленных по технологии В, если измерения дали следующие результаты (в пикофарадах):

Технология А	2.8	3.0	3.1	3.2	3.3	3.4	3.7	2.9
Технология В	3.8	3.4	3.6	2.9	2.8	3.0	3.4	3.0

18. У приборов двух партий, изготовленных с применением различной технологии, измерялось дифференциальное сопротивление канала R_i . Результаты измерений (в микроомах) следующие:

Технология А	0.01	0.02	0.12	0.30	0.29	0.15	0.21
Технология В	0.15	0.07	0.25	0.15	0.22	0.18	0.18

Влияет ли технология изготовления на величину дифференциального сопротивления канала R_i ?

19. В следующей таблице приведено время работы (в сотнях часов) электронных ламп А и В до выхода из строя.

A	32	34	35	37	42	43	47	58	59	62	69	71
B	39	48	54	65	70	76	87	90	111	118	126	127

Проверить гипотезу о различии среднего времени работы ламп этих двух типов.

20. Приведены результаты двух серий измерений, полученных при производстве азотной кислоты путем окисления аммиака кислородом воздуха:

Метод А	95.6	94.9	96.2	95.1	95.8	96.3	92.1	95.3	94.0
Метод В	93.3	92.1	94.7	90.1	95.6	90.0	94.7	95.2	93.7

Проверить гипотезу о принадлежности наблюдений к общей генеральной совокупности.

21. Данные следующей таблицы основаны на наблюдениях девяти пациентов, принимавших транквилизатор, и представляют степень депрессии, измеренной по специальной шкале. Значения x относятся к первому визиту пациента к врачу, значения y к моменту окончания лечения. Приводит ли прием транквилизатора к улучшению состояния пациентов?

x_i	1.83	0.50	1.62	2.48	1.68	1.88	1.55	3.06	1.30
y_i	0.88	0.65	0.60	2.05	1.06	1.29	1.06	3.14	1.29

22. Приведено содержание хрома (в весовых процентах) в образцах нержавеющей стали: 17.4, 17.9, 17.6, 18.1, 17.6, 18.9, 16.9, 17.5, 17.8, 17.4, 24.6, 26.0. Проверить гипотезу о том, что медиана процента хрома в стали равна 18% против альтернативы, что она не равна 18%.

23. Приведено содержание окислителя (z_i) в воде для орошения, измеряемое в миллионных долях озона: 0.32, 0.21, 0.28, 0.15, 0.08, 0.22, 0.17, 0.35, 0.20, 0.31, 0.17, 0.11. Проверить гипотезу о том, что медиана содержания окислителя равна 0.25, против альтернативы, что она меньше 0.25.

24. В следующей таблице представлены данные, относящиеся к методу прямого определения железистой сыворотки, полученные двумя способами (микрограмм/100 мл):

1-й способ	111	107	100	99	102	106	109	108	104	99
2-й способ	107	108	106	98	105	103	110	105	104	100
1-й способ	101	96	97	102	107	113	116	113	110	98
2-й способ	96	108	103	104	114	114	113	108	106	99

Проверить нулевую гипотезу о том, что обе выборки извлечены из одной генеральной совокупности.

25. На двух аналитических весах, в одном и том же порядке, взвешены десять проб химического вещества и получены следующие результаты взвешивания (в мг):

1-е весы	25	30	28	50	20	40	32	36	42	38
2-е весы	28	31	26	52	24	36	33	35	45	40

Проверить значимо или незначимо различаются результаты взвешиваний на аналитических весах.

26. Две лаборатории одним и тем же методом, в одном и том же порядке, определяли содержание углерода в тринадцати пробах нелегированной стали. Получены следующие результаты анализа (в %):

1-я лаборатория	0.18	0.12	0.12	0.08	0.08	0.12	0.19	0.32	0.27	0.22	0.34	0.14	0.46
2-я лаборатория	0.16	0.09	0.08	0.05	0.13	0.10	0.14	0.30	0.31	0.24	0.28	0.11	0.42

Различаются ли средние результаты анализа у обеих лабораторий?

27. Химическая лаборатория произвела анализ восьми проб двумя методами. Получены следующие результаты (в условных единицах):

1-й метод	15	20	16	22	24	14	18	20
2-й метод	15	22	14	25	29	16	20	24

Установить, значимо или незначимо различаются средние результаты анализа этими двумя методами.

28. Физическая подготовка девяти спортсменов была проверена при поступлении в спортивную школу, а затем после недели тренировки. Итоги проверки в баллах оказались следующими:

При поступлении	76	71	57	49	70	69	26	65	59
После недельной тренировки	81	85	52	52	70	63	33	83	62

Улучшилась или нет физическая подготовка спортсменов после недельной тренировки?

29. Измерительным прибором, практически не имеющим систематической ошибки, было сделано восемь независимых измерений некоторой величины. Результаты измерений таковы: 2504, 2486, 2525, 2495, 2515, 2528, 2492, 2494. Проверить гипотезу о том, что медиана результатов измерений равна 2500, против альтернативы, что она больше 2500.

30. При измерении угла теодолитом получены следующие результаты: $20^{\circ}40'20''$, $20^{\circ}40'34''$, $20^{\circ}40'42''$, $20^{\circ}40'28''$, $20^{\circ}40'34''$, $20^{\circ}40'27''$, $20^{\circ}40'25''$, $20^{\circ}40'32''$, $20^{\circ}40'46''$. Проверить гипотезу, что медиана измерений равна $20^{\circ}40'30''$, против альтернативы, что она не равна этому значению.

СЛОВАРЬ ИСПОЛЬЗУЕМЫХ ТЕРМИНОВ

Alternative hypothesis – альтернативная гипотеза.

Analysis of variance (ANOVA) – дисперсионный анализ.

Analysis options – процедуры анализа.

Analysis summary – сводка анализа.

Asymptotically confidence interval – асимптотический доверительный интервал.

Asymptotic distribution – асимптотическое распределение.

Asymptotic efficiency – асимптотическая эффективность.

Average – среднее значение.

Average rank – средний ранг.

Backward selection – уменьшение группы переменных в процедуре множественной регрессии.

Box-and-whisker plot – «ящик с усами». График в виде прямоугольника, построенный от сгиба до сгиба и имеющий поперечную черту на медиане с «усами» до указанных значений.

Central confidence interval – симметричный относительно центра доверительный интервал.

Chi-squared distribution – распределение χ^2 .

Compare – сравнение данных.

Comparison of alternative models – сравнение альтернативных моделей.

Confidence interval – доверительный интервал.

Consistent estimator – состоятельная оценка.

Continuous random variable – непрерывная случайная величина.

Contrast – контраст.

Correlation coefficient – коэффициент корреляции.

Count – число наблюдений на данном уровне фактора.

Covariance – ковариация, второй смешанный момент.

Critical region – критическая область.

Cumulative distribution function – интегральная функция распределения.

Degree of freedom – степени свободы.

Density function – функция плотности вероятности.
Density trace – график функции плотности.
Describe – описание данных.
Discrete random variable – дискретная случайная величина.
Dispersion – дисперсия, рассеяние.
Distribution fitting – подбор распределений.
Distribution-free test – свободный от распределения критерий.
Empirical distribution function – эмпирическая функция распределения.
Estimator – оценка; статистика, используемая в качестве оценки.
Expectation (of a continuous random variable) – математическое ожидание (непрерывной случайной величины).
Factor – фактор, обстоятельство.
F-distribution – F-распределение (распределение Фишера).
Fit – аппроксимация.
Fit the model – подбор модели.
Forecasts – предсказания.
Forward selection – увеличение группы переменных в процедуре множественной регрессии.
Frequency – частота.
Frequency histogram – гистограмма частот.
Greather than (больше чем) – выбор правостороннего критерия значимости.
Goodness-of-fit-test – критерий согласия.
Gross error – грубая ошибка.
Hazard function – функция риска.
Homogeneous groups – однородные группы.
Hypothesis test – критерий для проверки гипотезы.
Independent variable – независимая случайная величина.
Intercept – свободный член (уравнения регрессии).
Inverse CDF – обратная функция распределения.
Kruskal-Wallis tests – ранговый однофакторный критерий Краскела-Уоллиса.
Kurtosis – коэффициент эксцесса.
Lack-of-fit – неадекватность, рассогласованность.
Less then (меньше чем) – выбор левостороннего критерия значимости.
Level – уровень.
Level of factor – уровень фактора.
Linear regression – линейная регрессия.
Lower – нижний.
Mean (of a sample) – выборочное среднее.

Median – медиана.
Midpoint – середина интервала группировки.
Modify arrangement – задание классификации.
Multiple range test – множественные сравнения.
Multiple regression – множественная регрессия.
Multiple variable analysis – анализ многих переменных.
Nonparametric statistical procedure – непараметрический статистический метод.
Normal population – (генеральная) совокупность с нормальным распределением.
Normal probability plot – график на нормальной вероятностной бумаге.
Normal probability plot of residuals – нормальный вероятностный график остатков.
Not equal (не равно) – выбор двустороннего критерия значимости.
Null hypothesis – нулевая гипотеза.
Numeric data – числовые данные.
Observed versus predicted – график предсказанных значений.
One-sided test – односторонний критерий.
One-variable analysis – анализ одной переменной.
One-way ANOVA – однофакторный дисперсионный анализ.
Pane options – панель процедур.
Percentile – процентиль.
Plot of fitted model – график подобранной модели.
Point estimator – точечная оценка.
Probability distribution – распределение вероятностей.
Pure error – полная (чистая) ошибка.
Quantile – квантиль.
Random numbers – случайные числа.
Ratio – отношение.
Rejection region – область отклонения (гипотезы).
Relate – отношения данных.
Relative frequency – относительная частота.
Residual – остаток.
Response – отклик.
Ridge regression – ридж-регрессия или гребневая регрессия.
Sample standard deviation – выборочное среднее квадратическое отклонение.
Sample variance – выборочная дисперсия.
Scatterplot – диаграмма рассеивания.
Signed rank – знаковый ранг.
Significance level – уровень значимости.

Significance test – критерий значимости.
Simple regression – простая регрессия.
Size – объем, размер.
Skewness – коэффициент асимметрии.
Slope – угловой коэффициент (наклон).
Source – источник.
Summary statistics – описание данных.
Survivor function – функция выживаемости.
Tail areas – площади хвостов (распределений).
Tail areas probabilities – вероятности хвостов (распределений).
t-distribution – распределение Стьюдента.
Test for normality – критерий на принадлежность выборки к нормальному распределению.
Test statistic – статистика, лежащая в основе критерия.
Type I error – ошибка I рода.
Type II error – ошибка II рода.
Upper – верхний.
Unusual residuals – необычные остатки.
Variance check – тесты дисперсий.

Библиографический список

1. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. М.: Наука, 1983.
2. *Браунли К.А.* Статистическая теория и методология в науке и технике. М.: Наука, 1977.
3. *Вадзинский Р.Н.* Справочник по вероятностным распределениям. СПб.: Наука, 2001.
4. *Гаек Я., Шидак З.* Теория ранговых критериев. М.: Наука, 1971.
5. *Губарев В.В.* Алгоритмы статистических измерений. М.: Энергоатомиздат, 1985.
6. *Джонсон Н., Лион Ф.* Статистика и планирование эксперимента в технике и науке. М.: Мир. Т.1, 1980. Т.2, 1981.
7. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. М.: Финансы и статистика, 1987.
8. *Дэйвид Г.* Порядковые статистики. М.: Наука, 1979.
9. *Дюк В.* Обработка данных на ПК в примерах. СПб.: Питер, 1997.
10. *Калинина В.Н., Панкин В.Ф.* Математическая статистика. М.: Высшая школа, 1998.
11. *Кнут Д.Е.* Искусство программирования. Т. 2. Получисленные алгоритмы. М.: Мир, 1977.
12. *Мэйндональд Дж.* Вычислительные алгоритмы в прикладной статистике. М.: Финансы и статистика, 1988.

13. *Песаран М., Слейтер Л.* Динамическая регрессия: теория и алгоритмы. М.: Финансы и статистика, 1984.
14. *Плескунин В.И., Воронина Е.Д.* Теоретические основы организации и анализа выборочных данных в эксперименте. Л.: Из-во Лен. гос. ун-та, 1979.
15. *Пугачев В.С.* Теория вероятностей и математическая статистика. М.: Наука, 1979.
16. *Сборник задач по математике. Специальные курсы* / Под ред. А.В. Ефимова М.: Наука, 1984.
17. *Себер Дж.* Линейный регрессионный анализ. М.: Мир, 1980.
18. *Смирнов Н.В., Дунин-Барковский И.В.* Краткий курс математической статистики для технических приложений. М.: Физматгиз, 1959.
19. *Тьюки Дж.* Анализ результатов наблюдений. Разведочный анализ. М.: Мир, 1981.
20. *Тюрин Ю.Н., Макаров А.А.* Анализ данных на компьютере. М.: Финансы и статистика, 1995.
21. *Факторный, дискриминантный и кластерный анализ* / Под ред. И.С. Енюкова М.: Финансы и статистика, 1989.
22. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений. М.: Мир, 1980.
23. *Хастингс Н., Пикок Дж.* Справочник по статистическим распределениям. М.: Статистика, 1980.
24. *Холлендер М., Вульф Д.* Непараметрические методы статистики. М.: Финансы и статистика, 1983.
25. *Хьюбер П.* Робастность в статистике. М.: Мир, 1984.
26. *Шапорев С.Д.* Методы вычислительной математики и их приложения / Балт. гос. техн. ун-т. СПб., 2002.
27. *Iman R. L.* An approximation to the exact distribution of the Wilcoxon-Mann-Whitney rank sum test statistic. Communication in Statistic, A5(Theory and Method), 1976. p. 587-598.
28. *Wilcoxon F., Katti S.K., Wilcox Roberta A.* Critical values an probability levels for the Wilcoxon rank test.- In: Selected Tables in Mathematical Statistics, vol.1/2-d ed. H.L. Harter, D.B. Owen, eds.- Providence, R. I. Am. Math. Soc., 1973, p. 171-235.

О Г Л А В Л Е Н И Е

1. Случайные величины и их законы распределения	3
1.1. Законы распределения дискретных случайных величин	3
1.2. Числовые характеристики дискретных случайных величин, их свойства	6
1.3. Законы распределения непрерывных случайных величин	9
1.4. Числовые характеристики непрерывных случайных величин	10
1.5. Выборочные аналоги интегральной и дифференциальной функций распределения	13
1.6. Лабораторная работа № 1. Методы описательной статистики в пакете STATGRAPHICS	18
1.7. Нормальное распределение и его числовые характеристики	28
2. Распределения, связанные с нормальным распределением	31
2.1. χ^2 -распределение	31
2.2. t -распределение Стьюдента	37
2.3. F -распределение (распределение Фишера) или распределение дисперсионного отношения	40
2.4. Распределение Колмогорова	44
2.5. Гамма-распределение	47
2.6. Распределение Вейбулла (Вейбулла – Гнеденко)	48
2.7. Лабораторная работа № 2. Семейства вероятностных распределений в математических пакетах STATGRAPHICS и MATHCAD	50
3. Метод статистических испытаний (метод Монте-Карло)	61
3.1. Общие принципы метода статистических испытаний	61
3.2. Датчики базовой случайной величины (БСВ)	63
3.3. Моделирование на ЭВМ стандартной равномерно распределенной случайной величины (базовой случайной величины)	64
3.4. Моделирование дискретной случайной величины при помощи случайных событий	66
3.5. Моделирование непрерывных случайных величин	68
3.6. Лабораторная работа № 3. Моделирование некоторых распределений с помощью базовых случайных величин в пакете MATHCAD	71
4. Точечные и интервальные оценки параметров распределений и их свойства	81
4.1. Статистические характеристики вариационных рядов и показатели их качества	81
4.2. Типовые принципы, используемые при построении оценок [5]	82
4.3. Точечные оценки вероятности по частоте, математического ожидания и дисперсии	85
4.4. Неравенство Крамера - Рао	89
4.5. Методы получения точечных оценок	92
4.6. Сущность интервального оценивания	96
4.7. Приближенные и точные доверительные интервалы для параметров распределений	96
4.8. Лабораторная работа № 4. Оценивание параметров вероятностных распределений в пакетах STATGRAPHICS и MATHCAD	101
5. Проверка статистических гипотез. Критерий согласия	107
5.1. Понятие статистической гипотезы. Основные этапы проверки гипотез	107
5.2. Критерий Неймана – Пирсона	113
5.3. Проверка гипотез о числовых значениях параметров нормального распределения	115

5.4. Проверка гипотез о параметрах двух нормальных распределений	118
5.5. Лабораторная работа № 5. Проверка статистических гипотез о числовых значениях нормальных распределений в математических пакетах STATGRAPHICS и MATHCAD	123
5.6. Критерии согласия	131
5.7. Лабораторная работа № 6. Критерии согласия в статистическом пакете STATGRAPHICS	142
5.8. Лабораторная работа №7. Критерии согласия в математическом пакете MATHCAD	151
6. Однофакторный дисперсионный анализ	158
6.1. Постановка задачи	158
6.2. Дисперсионный анализ	159
6.3. Ранговый однофакторный анализ	168
6.4. Критерий Краскела - Уоллиса (H-критерий)	170
6.5. Лабораторная работа № 8. Однофакторный ранговый и дисперсионный анализ в статистическом пакете STATGRAPHICS	173
7. Регрессионный анализ	189
7.1. Модели регрессии	189
7.2. Оценка параметров линейной регрессии методом наименьших квадратов	192
7.3. Интервальные оценки параметров линейной регрессии и кривой регрессии	197
7.4. Проверка адекватности линейной регрессии	203
7.5. Выбор наилучшей регрессии	206
7.6. Лабораторная работа № 9. Регрессионный анализ в пакетах STATGRAPHICS и MATHCAD	207
8. Непараметрические методы статистики	222
8.1. Основные понятия и область применимости непараметрических методов	222
8.2. Критерий знаков	223
8.3. Критерий знаков для одномерной выборки	227
8.4. Ранговый критерий (одновыборочный критерий Вилкоксона)	230
8.5. Двухвыборочный ранговый критерий Вилкоксона	236
8.6. Лабораторная работа № 10. Критерии знаков и рангов в пакете MATHCAD	239
Словарь используемых терминов	248
<i>Библиографический список</i>	251

Шапоров Сергей Дмитриевич

Прикладная статистика

Редактор *Г.В. Никитина*
Корректор *А.А. Баутдинова*

Подписано в печать 04.07.2003. Формат 60×84/16. Бумага документная.
Печать трафаретная. Усл. печ. л. 15,875. Уч. - изд. л. 18,5. Тираж 150 экз. Заказ № 73.

Балтийский государственный технический университет
Типография БГТУ
190005, С.-Петербург, 1-я Красноармейская ул., д.1