

Глава 3. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

§ 1. Первоначальная обработка одномерной выборки. Точечные оценки неизвестных параметров распределения

В математической статистике изучаются методы систематизации и обработки экспериментальных данных с целью извлечения из них информации о случайных величинах, нахождения оценок неизвестных параметров и проверки гипотез.

Пусть при неизменных условиях проводится n независимых опытов, в каждом из которых наблюдается некоторая случайная величина ξ с функцией распределения $F(x)$. Тогда результат этих наблюдений описывается n -мерным случайным вектором (X_1, X_2, \dots, X_n) , где случайные величины X_i независимы и каждая из них распределена так же, как случайная величина ξ . Этот случайный вектор, а также его реализацию, т.е. набор (x_1, x_2, \dots, x_n) значений $X_i = x_i$, $i = 1, 2, \dots, n$, полученных в результате наблюдений, называют *выборкой объема n* .

Закон распределения случайной величины ξ иногда называют распределением генеральной совокупности, а про выборку говорят, что она взята из генеральной совокупности случайной величины ξ .

Элементы выборки, расположенные в порядке возрастания, образуют *вариационный ряд*. Если выборка (x_1, x_2, \dots, x_n) содержит k различных элементов z_1, z_2, \dots, z_k , причем элемент z_i встречается n_i раз ($i = 1, 2, \dots, k$), то число n_i называется *частотой* элемента z_i , отношение $\frac{n_i}{n}$ — *относительной частотой*, а последовательность пар (z_i, n_i) — *статистическим рядом*. Очевидно, что $\sum_{i=1}^k n_i = n$, $\sum_{i=1}^k \frac{n_i}{n} = 1$. Обычно статистический ряд записывается в виде таблицы, первая строка которой содержит элементы z_i , а вторая — их частоты.

При большом объеме n выборки для более компактного представления результатов опытов используется *группировка выборки*. Для этого отрезок, содержащий все элементы выборки, разбивается на k непересекающихся промежутков (обычно одинаковой длины), называемых интервалами группировки. Если n_i , $i = 1, 2, \dots, k$ — количество элементов выборки, попавших в i -й интервал (частоты), то последовательность из интервалов группировки и соответствующих частот называется *группированным статистическим рядом*.

Функция $\hat{F}_n(x) = \frac{m(x)}{n} = \sum_{i: z_i < x} \frac{n_i}{n}$, где n — объем выборки, а $m(x)$ — число

элементов x_i в выборке, меньших x , называется *эмпирической* (или *выборочной*) *функцией распределения*. Функция $\hat{F}_n(x)$ представляет собой неубывающую кусочно-постоянную функцию и является оценкой неизвестной функции распределения $F(x)$ (а именно, $\mathbf{P}(|\hat{F}_n(x) - F(x)| < \varepsilon) \rightarrow 1$ при $n \rightarrow \infty$ $\forall \varepsilon > 0$, $\forall x \in \mathbb{R}$).

Для оценки неизвестной плотности вероятности $p(x)$ непрерывной случайной величины ξ служит *гистограмма относительных частот*. Она представляет собой фигуру из прямоугольников (построенных в прямоугольной сис-

теме координат), основаниями которых являются интервалы группировки длины b_i , а высоты соответственно равны $\frac{n_i}{nb_i}$ (оценка плотности вероятности на i -м интервале). Очевидно, площадь гистограммы равна единице.

Всякую функцию от выборки (X_1, X_2, \dots, X_n) называют *статистикой*. (*Точечной*) *оценкой* неизвестного параметра θ называется статистика $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$, значения которой приближенно равны оцениваемому параметру θ . Оценка $\hat{\theta}_n$ параметра θ называется *несмешенной*, если $\mathbf{M}\hat{\theta}_n = \theta$, т.е. если она не имеет систематической ошибки.

Точечной оценкой для математического ожидания $\mathbf{M}\xi$ служит *выборочное среднее*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ (несмешенная оценка),}$$

для дисперсии $\mathbf{D}\xi$ — *несмешенная оценка дисперсии*

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

для среднего квадратичного отклонения $\sigma\xi$ — оценка $s = \sqrt{s^2}$.

Заметим, что при вычислении оценки по конкретной выборке вместо случайных величин (X_1, X_2, \dots, X_n) подставляются их конкретные числовые значения $X_i = x_i$, $i = 1, 2, \dots, n$, поэтому

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i z_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^k n_i (z_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^k n_i z_i^2 - n\bar{x}^2 \right),$$

где z_i , как и выше, означают различные элементы среди элементов выборки, n_i — частота элемента z_i . В случае группированной выборки в качестве значений z_i берутся середины интервалов группировки.

Пример 1. Записать в виде вариационного и статистического рядов выборку 4, 5, 3, 7, 2, 10, 4, 5, 5, 3, 10, 4, 7, 4, 4. Найти оценки математического ожидания, дисперсии и среднего квадратичного отклонения.

□ Объем выборки $n = 15$. Упорядочив элементы выборки по величине, получим вариационный ряд

$$2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 7, 7, 10, 10.$$

Различными элементами в данной выборке являются элементы $z_1 = 2$, $z_2 = 3$, $z_3 = 4$, $z_4 = 5$, $z_5 = 7$, $z_6 = 10$; их частоты соответственно равны $n_1 = 1$, $n_2 = 2$, $n_3 = 5$, $n_4 = 3$, $n_5 = 2$, $n_6 = 2$ (для контроля правильности находим $\sum n_i = 15$).

Следовательно, статистический ряд исходной выборки имеет вид:

z_i	2	3	4	5	7	10
n_i	1	2	5	3	2	2

Оценка математического ожидания (выборочное среднее)

$$\bar{x} = \frac{1}{15}(1 \cdot 2 + 2 \cdot 3 + 5 \cdot 4 + 3 \cdot 5 + 2 \cdot 7 + 2 \cdot 10) = \frac{77}{15} \approx 5,133;$$

дисперсии (несмешенная оценка дисперсии)

$$s^2 = \frac{1}{14} \left(\sum_{i=1}^6 n_i z_i^2 - 15 \cdot \bar{x}^2 \right) = \frac{1}{14} [1 \cdot 2^2 + 2 \cdot 3^2 + 5 \cdot 4^2 + 3 \cdot 5^2 + 2 \cdot 7^2 + 2 \cdot 10^2 - 15 \cdot (\frac{77}{15})^2] \approx 5,695;$$

среднего квадратичного отклонения $s = \sqrt{s^2} \approx 2,386$. ■

Пример 2. Найти эмпирическую функцию распределения для выборки, представленной в виде статистического ряда:

z_i	1	4	6
n_i	10	15	25

□ Объем выборки $n = 50$. При любом $x \leq 1$ элементы выборки, меньшие, чем x , отсутствуют, поэтому $\hat{F}_n(x) = 0$. При $1 < x \leq 4$ число элементов исходной выборки, меньших x , равно 10 (частота элемента $z_1 = 1 < x$), поэтому $\hat{F}_n(x) = \frac{10}{50} = 0,2$. При $4 < x \leq 6$ число элементов выборки, меньших x , равно $10 + 15 = 25$ (сумма частот элементов $z_1 = 1 < x$ и $z_2 = 4$, меньших x), откуда $\hat{F}_n(x) = \frac{25}{50} = 0,5$. Наконец, при любом $x > 6$ все 50 элементов выборки меньше, чем x , поэтому $\hat{F}_n(x) = \frac{50}{50} = 1$.

Таким образом,

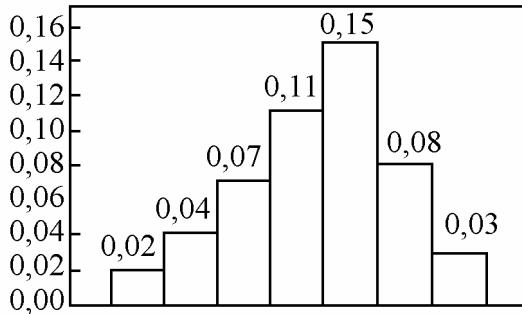
$$\hat{F}_n(x) = \begin{cases} 0 & \text{если } x \leq 1, \\ 0,2 & \text{если } 1 < x \leq 4, \\ 0,5 & \text{если } 4 < x \leq 6, \\ 1 & \text{если } x > 6. \end{cases}$$

Пример 3. Построить гистограмму выборки, представленной группированным статистическим рядом

Номер интервала i	Границы интервала	Частота n_i
1	10–12	2
2	12–14	4
3	14–16	7
4	16–18	11
5	18–20	15
6	20–22	8
7	22–24	3

и найти выборочное среднее и несмешенную оценку дисперсии.

□ Объем выборки $n = 50$. Длина интервалов группировки $b_i = 2$. Строим прямоугольники с длинами оснований $b_i = 2$ и высотами $\frac{n_i}{50 \cdot 2} = \frac{n_i}{100}$, где $n_1 = 2$, $n_2 = 4, \dots, n_7 = 3$:



Оценки \bar{x} и s^2 находим, принимая в качестве различных значений z_i выборки середины интервалов группировки (11, 13, 15, 17, 19, 21, 23):

$$\bar{x} = \frac{1}{50}(2 \cdot 11 + 4 \cdot 13 + 7 \cdot 15 + 11 \cdot 17 + 15 \cdot 19 + 8 \cdot 21 + 3 \cdot 23) = 17,76;$$

$$s^2 = \frac{1}{49} \sum_{i=1}^7 n_i (z_i - 17,76)^2 \approx 8,798. \blacksquare$$

Для каждой из приведенных ниже выборок построить вариационный и статистические ряды.

246. 11, 15, 12, 16, 19, 6, 11, 12, 13, 16, 8, 9, 13, 13, 11.

247. 17, 18, 16, 16, 17, 18, 19, 17, 15, 17, 19, 18, 16, 16, 18, 18.

Для каждой из приведенных ниже выборок, представленных статистическими рядами, найти эмпирическую функцию распределения и построить ее график.

248.

z_i	4	7	8
n_i	5	2	3

249.

z_i	1	4	5	7
n_i	20	10	14	6

250. Построить гистограмму выборки, представленной группированным статистическим рядом:

Границы интервалов	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Частоты n_i	1	2	7	18	12	8	2

251. В результате измерения времени решения контрольной задачи учениками 4-го класса получены следующие результаты (в секундах):

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

Построить для данной выборки группированный статистический ряд и гистограмму, разбив отрезок [14, 77] на 7 интервалов группировки одинаковой длины (элемент, совпадающий с общей границей двух соседних интервалов, относить к правому интервалу).

252. Найти выборочное среднее и несмешенную оценку дисперсии по выборкам:

- а) 1, 2, 3, 4, 5, 5, 9;
- б) 1, 2, 3, 4, 5, 5, 12.

Сравнить полученные результаты.

253. По выборке 3, 1, 2, 2, 3, 2, 3, 5, 2, 1 найти оценки математического ожидания, дисперсии и среднего квадратичного отклонения.

254. По группированной выборке

Границы интервалов	1–3	3–5	5–7	7–9	9–11	11–13
Частоты n_i	1	2	4	2	1	1

найти выборочное среднее и несмешенную оценку дисперсии.

255. В результате 5 независимых измерений некоторой величины измерительным прибором, не имеющим систематической погрешности, получены следующие значения: 2781, 2836, 2807, 2763, 2858. Найти оценку измеряемой величины и оценку среднего квадратичного отклонения погрешности измерения.

256. В партии из 40 деталей измерялись отклонения x_i (в мкм) от номинального размера, после чего были найдены $\sum_i x_i = 689$ и

$\sum_i x_i^2 = 12635$. Найти оценки среднего значения и дисперсии отклонения.

257. Группированная выборка для предела прочности образцов сварного шва (\bar{x} / \bar{x}^2) имеет вид:

Границы интервалов	28–30	30–32	32–34	34–36	36–38	38–40	40–42	42–44
Частоты n_i	8	15	15	12	15	20	10	5

Вычислить выборочное среднее и оценку дисперсии.

§ 2. Доверительные интервалы для параметров нормального распределения

Интервальной оценкой (доверительным интервалом) для неизвестного параметра θ называется интервал (θ_1, θ_2) , который с заданной вероятностью p содержит оцениваемый параметр: $P(\theta_1 < \theta < \theta_2) = p$. При этом вероятность p называется доверительной, а число $\alpha = 1 - p$ — уровнем значимости. Обычно используются значения p , равные 0,90; 0,95; 0,99 (соответственно $\alpha = 0,1; 0,05; 0,01$).

Квантилью порядка p ($0 < p < 1$) непрерывной случайной величины ξ с возрастающей на $(-\infty, +\infty)$ функцией распределения $F(x)$ называется число x_p , определяемое равенством $F(x_p) = P(\xi < x_p) = p$ (т.е. $x_p = F^{-1}(p)$, где F^{-1} — функция, обратная к F).

1 Пусть случайная величина ξ имеет нормальное распределение. Для оценки ее математического ожидания a по выборке объема n в случае, если известна дисперсия $D\xi = \sigma^2$, служит доверительный интервал $(\bar{x} - \delta, \bar{x} + \delta)$, где \bar{x} — выборочное среднее, $\delta = u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$ — точность оценки, $u_{1-\alpha/2}$ — квантиль порядка $1 - \alpha/2$ стандартного нормального распределения $N(0,1)$, α — заданный уровень значимости. Для приведенных выше значений доверительной вероятности соответствующие квантили равны:

$$u_{0,95} \approx 1,645; \quad u_{0,975} \approx 1,960; \quad u_{0,995} \approx 2,576.$$

2 Более естественной является ситуация, когда оба параметра a и σ неизвестны. В этом случае точность δ интервальной оценки $(\bar{x} - \delta, \bar{x} + \delta)$ математического ожидания находится по формуле:

$$\delta = t_{1-\alpha/2;n-1} \frac{s}{\sqrt{n}},$$

где $s = \sqrt{s^2}$ — оценка среднего квадратичного отклонения, а $t_{1-\alpha/2;n-1}$ — квантиль порядка $1 - \alpha/2$ так наз. распределения Стьюдента с $n-1$ степенями свободы. Таблица квантилей $t_{p,k}$ распределения Стьюдента с k степенями свободы приводится на стр. 70 (таблица 2).

Для оценки неизвестной дисперсии $D\xi$ нормального распределения служат доверительные интервалы

$$\left(\frac{ns_0^2}{\chi^2_{1-\alpha/2;n}}, \frac{ns_0^2}{\chi^2_{\alpha/2;n}} \right)$$

в случае, когда математическое ожидание a известно, и

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2;n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2;n-1}} \right)$$

при неизвестном a ; здесь $s_0^2 = \frac{1}{n} \sum_{i=1}^k n_i(z_i - a)^2$ и s^2 — несмешенные оценки дисперсии.

персии в том и другом случае, а $\chi^2_{p,k}$ — квантиль порядка p так наз. *распределения χ^2 с k степенями свободы* (таблица 3, стр. 70).

Пример. По выборке из нормального распределения, представленной в виде статистического ряда

z_i	-2	1	2	3	4	5
n_i	2	1	2	2	2	1

оценить с помощью доверительных интервалов математическое ожидание a и дисперсию σ^2 с доверительными вероятностями 0,90 и 0,99.

□ Объем выборки $n = 10$. Находим выборочное среднее $\bar{x} = 2$ и несмещенную оценку дисперсии $s^2 = \frac{52}{9}$. При доверительной вероятности 0,90 уровень значимости $\alpha = 0,1$, откуда $\alpha/2 = 0,05$ и $1 - \alpha/2 = 0,95$. По таблице 2 находим квантиль порядка 0,95 распределения Стьюдента с $n - 1 = 9$ степенями свободы: $t_{0,95;9} \approx 1,833$. Вычисляем точность интервальной оценки математического ожидания:

$$\delta \approx 1,833 \cdot \frac{\sqrt{\frac{52}{9}}}{\sqrt{10}} \approx 1,39.$$

Следовательно, с вероятностью 0,90

$$0,61 < a < 3,39.$$

Для нахождения доверительного интервала для дисперсии находим по таблице 3 квантили порядков 0,95 и 0,05 распределения χ^2 с 9 степенями свободы:

$$\chi^2_{0,95;9} \approx 16,9 \text{ и } \chi^2_{0,05;9} \approx 3,33.$$

Вычисляем границы доверительного интервала:

$$\frac{9 \cdot \frac{52}{9}}{16,9} \approx 3,1 \text{ и } \frac{9 \cdot \frac{52}{9}}{3,33} \approx 15,6.$$

Итак, $3,1 < \sigma^2 < 15,6$ с вероятностью 0,90.

Аналогично, для доверительной вероятности 0,99 находим:

$$t_{0,995;9} \approx 3,250, \quad \delta \approx 2,47 \text{ и } -0,47 < a < 4,47;$$

$$\chi^2_{0,995;9} \approx 23,6, \quad \chi^2_{0,005;9} \approx 1,73 \text{ и } 2,2 < \sigma^2 < 30,1. \blacksquare$$

258. В течение продолжительного срока при анализе данного материала на содержание железа установлено среднее квадратичное отклонение $\sigma = 0,12\%$. Найти с доверительной вероятностью 0,95 доверительный интервал для истинного содержания a железа в образце, если по результатам 6 анализов среднее содержание железа составило 32,56%.

259. Средняя продолжительность горения электролампы, определенная по выборке объема $n = 100$ из большой партии ламп, оказалась равной 1000 ч. Найти с доверительной вероятностью 0,99 доверительный интервал для средней продолжительности a горения лампы всей партии, если известно, что среднее квадратичное отклонение продолжительности горения лампы равно 40 ч.

260. По выборке объема $n = 100$ вычислено выборочное среднее диаметра изготовленных валиков. Найти с доверительной вероятностью 0,90 точность, с которой выборочное среднее оценивает математическое ожидание a диаметра изготавляемых валиков, зная, что их среднее квадратичное отклонение равно 200 мкм.

261. Оценка \bar{x} (кОм) сопротивления для большой партии однотипных резисторов определяется по результатам измерений n случайно отобранных экземпляров. Сколько измерений нужно произвести, чтобы с вероятностью 0,95 утверждать, что для всей партии резисторов сопротивление находится в пределах $\bar{x} \pm 0,1$ кОм, если дисперсия сопротивления в партии $\sigma^2 = 1$ кОм²?

262. Для определения углового размера изготовленной детали используют среднее арифметическое нескольких измерений. Среднее квадратичное отклонение измерительного прибора равно 1,5'. Найти количество измерений, которое нужно произвести, чтобы погрешность результата с вероятностью 0,99 не превосходила 1'.

263. По выборке объема $n = 25$ были определены выборочные оценки для содержания углерода в единице продукта: $\bar{x} = 18$ г, $s^2 = 16$ г². Найти 90% и 99% доверительные интервалы для среднего содержания углерода.

264. По выборке объема $n = 16$ из партии конденсаторов определены выборочные оценки для емкости конденсатора: $\bar{x} = 20$ мкФ, $s^2 = 4$ мкФ². Найти с доверительной вероятностью 0,95 доверительный интервал для средней емкости конденсатора в партии.

265. Произведено пять независимых равноточных измерений для определения заряда электрона. Получены следующие результаты (в кулонах): $1,594 \cdot 10^{-19}$; $1,597 \cdot 10^{-19}$; $1,598 \cdot 10^{-19}$; $1,593 \cdot 10^{-19}$; $1,590 \cdot 10^{-19}$. Найти доверительный интервал для заряда электрона с доверительной вероятностью 0,99.

266. По выборке объема $n = 16$ найдена несмещенная оценка дисперсии диаметра изготовленных валиков: $s^2 = 1000$ мкм². Найти с доверительной вероятностью 0,95 доверительные интервалы для дисперсии и среднего квадратичного отклонения диаметра валика. (Указание: $\chi^2_{0,025;15} = 6,26$, $\chi^2_{0,975;15} = 27,5$).

267. По данным задачи 263 найти 90% и 99% доверительные интервалы для дисперсии. (Указание: $\chi^2_{0,005;24} = 9,89$, $\chi^2_{0,05;24} = 13,8$, $\chi^2_{0,95;24} = 36,4$, $\chi^2_{0,995;24} = 45,6$).

n_i	1	1	1	3	2	3	1	2	1
-------	---	---	---	---	---	---	---	---	---

— статистический ряд.

247. 15, 16, 16, 16, 16, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 19, 19 — вариационный ряд,

z_i	15	16	17	18	19
n_i	1	4	4	5	2

— статистический ряд.

248. $\hat{F}_n(x) = 0$ при $x \leq 4$; 0,5 при $4 < x \leq 7$; 0,7 при $7 < x \leq 8$; 1 при $x > 8$.

249. $\hat{F}_n(x) = 0$ при $x \leq 1$; 0,4 при $1 < x \leq 4$; 0,6 при $4 < x \leq 5$; 0,88 при $5 < x \leq 7$; 1 при $x > 7$. **250.** $n = 50$; $b_i = 10$; высоты прямоугольников: (2, 4, 14, 36, 24, 16, 4)· 10^{-3} .

251. $n = 50$; $b_i = 9$;

Номер интервала i	Границы интервала	Частота n_i	$\frac{n_i}{nb_i}$
1	14–23	2	$\approx 44 \cdot 10^{-4}$
2	23–32	3	$\approx 67 \cdot 10^{-4}$
3	32–41	6	$\approx 133 \cdot 10^{-4}$
4	41–50	17	$\approx 378 \cdot 10^{-4}$
5	50–59	10	$\approx 222 \cdot 10^{-4}$
6	59–68	9	$200 \cdot 10^{-4}$
7	68–77	3	$\approx 67 \cdot 10^{-4}$

252. а) $\bar{x} \approx 4,143$, $s^2 \approx 6,810$; б) $\bar{x} \approx 4,571$, $s^2 \approx 12,952$. **253.** $\bar{x} = 2,4$, $s^2 \approx 1,378$,

$s \approx 1,174$. **254.** $\bar{x} \approx 6,545$, $s^2 \approx 8,073$. **255.** $\bar{x} = 2809$, $s \approx 38,8$. **256.** $\bar{x} \approx 17,2$,

$s^2 \approx 19,67$. **257.** $\bar{x} = 35,72$, $s^2 \approx 16,12$.

§ 2

258. $32,46\% < a < 32,66\%$. **259.** $989,7 < a < 1010,3$. **260.** ≈ 33 мкм. **261.** $n \geq 385$.

262. $n \geq 15$. **263.** (16,63; 19,37); (15,76; 20,24). **264.** (18,9; 21,1). **265.** ($1,588 \cdot 10^{-19}$;

$1,601 \cdot 10^{-19}$). **266.** (545, 2396); (23, 49). **267.** (10,55; 27,83); (8,42; 38,83). **268.1.** (9,84;

22,16). **268.2.** (9,53; 20,37).

§ 3

269. Да ($\chi^2 \approx 0,78$, $\chi^2_{0,95;1} \approx 3,84$). **270.** Нет ($\chi^2 = 24$, $\chi^2_{0,95;1} \approx 3,84$). **271.** Нет ($\chi^2 = 20$,

$\chi^2_{0,99;4} \approx 13,3$). **272.** H_0 принимается: $\lambda = \bar{x} = 0,4$, $\chi^2 \approx 0,69$, $\chi^2_{0,95;2} \approx 5,99$. **273.** H_0 от-

клоняется: $\lambda = \bar{x} = 0,5$, $\chi^2 \approx 5,40$, $\chi^2_{0,90;1} \approx 2,71$. **274.** H_0 принимается: $\chi^2 \approx 3,26$,

$\chi^2_{0,90;5} \approx 9,24$. **275.** H_0 отклоняется: $\chi^2 \approx 6,22$, $\chi^2_{0,90;2} \approx 4,61$.

§ 4