ГЛАВА V

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

§ 16. Основные выборочные характеристики

16.1. Основные понятия. *Математическая статистика* — наука о математических методах, позволяющих по статистическим данным, например по реализациям случайной величины (СВ), построить теоретико-вероятностную модель исследуемого явления. Задачи математической статистики являются, в некотором смысле, обратными к задачам теории вероятностей. Центральным понятием математической статистики является выборка.

Определение 16.1. *Однородной выборкой* (*выборкой*) *объема* n при $n \geqslant 1$ называется случайный вектор $Z_n \stackrel{\Delta}{=} \operatorname{col}(X_1, \ldots, X_n)$, компоненты которого X_i, $i = \overline{1,n}$, называемые *элементами выборки*, являются независимыми СВ с одной и той же функцией распределения $F(x)$. Будем говорить, что выборка Z_n *соответствует* функции распределения $F(x)$.

Определение 16.2. *Реализацией выборки* называется неслучайный вектор $z_n \stackrel{\Delta}{=} \operatorname{col}(x_1, \ldots, x_n)$, компонентами которого являются реализации соответствующих элементов выборки X_i, $i = \overline{1,n}$.

Из определений 16.1 и 16.2 вытекает, что реализацию выборки z_n можно также рассматривать как последовательность $x_1, \ldots, x_n$ из n реализаций одной и той же СВ X, полученных в серии из n независимых одинаковых опытов, проводимых в одинаковых условиях. Поэтому можно говорить, что выборка Z_n *порождена наблюдаемой* СВ X, имеющей распределение $F_X(x) \stackrel{\Delta}{=} F(x)$.

Определение 16.3. Если компоненты вектора Z_n независимы, но их распределения $F_1(x_1), \ldots, F_n(x_n)$ различны, то такую выборку называют *неоднородной*.

Определение 16.4. Множество S всех реализаций выборки Z_n называется *выборочным пространством*.

Выборочное пространство может быть всем n-мерным евклидовым пространством $\mathbb{R}^n$ или его частью, если СВ X непрерывна, а

также может состоять из конечного или счетного числа точек из $\mathbb{R}^n$, если СВ X дискретна.

На практике при исследовании конкретного эксперимента распределения $F_1(x_1), \ldots, F_n(x_n)$ СВ $X_1, \ldots, X_n$ редко бывают известны полностью. Часто априори (до опыта) можно лишь утверждать, что распределение $F_{Z_n}(z_n) = F_1(x_1) \cdot \ldots \cdot F_n(x_n)$ случайного вектора Z_n принадлежит некоторому классу (семейству) $\mathcal{F}$.

Определение 16.5. Пара $(S, \mathcal{F})$ называется *статистической моделью* описания серии опытов, порождающих выборку Z_n.

Определение 16.6. Если распределения $F_{Z_n}(z_n, \theta)$ из класса $\mathcal{F}$ определены с точностью до некоторого векторного параметра $\theta \in \Theta \subset \mathbb{R}^s$, то такая статистическая модель называется *параметрической* и обозначается $(S_\theta, F_{Z_n}(z_n, \theta))$, $\theta \in \Theta \subset \mathbb{R}^s$.

В некоторых случаях выборочное пространство может не зависеть от неизвестного параметра θ распределения $F_{Z_n}(z_n, \theta)$.

В зависимости от вида статистической модели в математической статистике формулируются соответствующие задачи по обработке информации, содержащейся в выборке.

Определение 16.7. СВ $Z \stackrel{\Delta}{=} \varphi(Z_n)$, где $\varphi(z_n)$ — произвольная функция, определенная на выборочном пространстве S и не зависящая от распределения $F_{Z_n}(z_n, \theta)$, называется *статистикой*.

16.2. Вариационный ряд.

Определение 16.8. Упорядочим элементы реализации выборки $x_1, \ldots, x_n$ по возрастанию: $x^{(1)} \leqslant x^{(2)} \leqslant \ldots \leqslant x^{(n)}$, где верхний индекс соответствует номеру элемента в упорядоченной последовательности. Обозначим через $X^{(k)}$, $k = \overline{1, n}$, случайные величины, которые при каждой реализации z_n выборки Z_n принимают k-е (по верхнему номеру) значения $x^{(k)}$. Упорядоченную последовательность СВ $X^{(1)} \leqslant \ldots \leqslant X^{(n)}$ называют *вариационным рядом выборки*.

Определение 16.9. Элементы $X^{(k)}$ вариационного ряда называются *порядковыми статистиками*, а крайние члены вариационного ряда $X^{(1)}$, $X^{(n)}$ — *экстремальными порядковыми статистиками*.

Например, для $k = 1$ функция $\varphi(z_n)$ для статистики $X^{(1)} = \varphi(Z_n)$ определяется следующим образом:

$$\varphi(z_n) = \min\left\{x_k\colon\ k = \overline{1, n}\right\}.$$

Если однородная выборка Z_n соответствует распределению $F(x)$, то k-я порядковая статистика $X^{(k)}$ имеет следующую функцию

распределения:

$$F_{(k)}(x) = \mathbf{P}\left\{X^{(k)} \leqslant x\right\} = \sum_{i=k}^{n} C_n^i [F(x)]^i [1 - F(x)]^{n-i}.$$

В частности, для $k = 1$ и $k = n$ имеем

$$F_{(1)}(x) = 1 - [1 - F(x)]^n, \qquad F_{(n)}(x) = [F(x)]^n.$$

Если функция распределения $F(x)$ имеет плотность $f(x)$, то порядковая статистика $X^{(k)}$ имеет следующую плотность распределения:

$$f_{(k)}(x) = \frac{n!}{(k-1)!\,(n-k)!}\,[F(x)]^{k-1}[1 - F(x)]^{n-k} f(x).$$

Определение 16.10. Порядковая статистика $X^{([np]+1)}$ с номером $[np] + 1$, где $[\,\cdot\,]$ — целая часть числа, называется *выборочной квантилью уровня* p.

Если в некоторой окрестности точки x_p плотность распределения $f(x)$ СВ X непрерывна вместе с её производной и, кроме того, $f(x_p) > 0$, то

$$\left(X^{([np]+1)} - x_p\right)\sqrt{\frac{nf^2(x_p)}{p(1-p)}} \xrightarrow{F} U \qquad \text{при } n \to \infty,$$

где СВ U имеет распределение $\mathbf{N}(0;1)$. Таким образом, при больших n можно считать, что выборочная квантиль $X^{([np]+1)}$ близка к x_p, и, более того, распределение статистики $X^{([np]+1)}$ может быть аппроксимировано нормальным распределением $\mathbf{N}\left(x_p;\, \frac{p(1-p)}{nf^2(x_p)}\right)$.

16.3. Выборочная функция распределения.

Пусть серия из n испытаний проводится по схеме Бернулли, т. е. испытания проводятся независимо друг от друга и некоторое событие A при каждом испытании появляется с одной и той же вероятностью $p \triangleq \mathbf{P}(A)$. Пусть $M_n(A)$ — случайное число появлений события A в этой серии, а $W_n(A) \triangleq M_n(A)/n$ — частота события A в серии из n испытаний.

Рассмотрим выборку Z_n, порожденную СВ X с функцией распределения $F_X(x)$. Определим для каждого $x \in \mathbb{R}^1$ событие $A_x \triangleq \{X \leqslant x\}$, для которого $\mathbf{P}(A_x) = F_X(x)$. Тогда $M_n(A_x)$ — случайное число элементов выборки Z_n, не превосходящих x.

О п р е д е л е н и е 16.11. Частота $W_n(A_x)$ события A_x, как функция $x \in \mathbb{R}^1$, называется *выборочной (эмпирической) функцией распределения* СВ X и обозначается

$$\widehat{F}_n(x) \triangleq W_n(A_x).$$

Для каждого фиксированного $x \in \mathbb{R}^1$ СВ $\widehat{F}_n(x)$ является статистикой, реализациями которой являются числа 0, $1/n$, $2/n$, ... , n/n, и при этом

$$\mathbf{P}\left\{\widehat{F}_n(x) = \frac{k}{n}\right\} = \mathbf{P}\{M_n(A_x) = k\}, \quad k = \overline{1,n}.$$

Любая реализация $\overline{F}_n(x)$ выборочной функции $\widehat{F}_n(x)$ является ступенчатой функцией, характерный вид которой показан на рис. 16.1. В точках $x^{(1)} < \ldots < x^{(n)}$, где $x^{(k)}$ — реализация порядковой статистики $X^{(k)}$, функция $\overline{F}_n(x)$ имеет скачки величиной $1/n$ и является непрерывной справа.

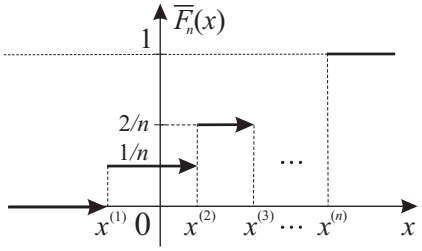

Рис. 16.1

С в о й с т в а $\widehat{F}_n(x)$

1) $\mathbf{M}\left[\widehat{F}_n(x)\right] = F(x)$, для любого $x \in \mathbb{R}^1$ и любого $n \geqslant 1$;

2) $\sup\limits_{x\in\mathbb{R}^1}\left|\widehat{F}_n(x) - F(x)\right| \xrightarrow{\text{п.н.}} 0$ при $n \to \infty$;

3) $\mathbf{M}\left[\left(\widehat{F}_n(x) - F(x)\right)^2\right] = d_n(x) \triangleq \dfrac{F(x)(1-F(x))}{n} \leqslant \dfrac{1}{4n}$;

4) $\left(\widehat{F}_n(x) - F(x)\right) \Big/ \sqrt{d_n(x)} \xrightarrow{F} U$ при $n \to \infty$, где СВ U имеет распределение $\mathbf{N}(0;1)$.

Первые два свойства свидетельствуют о том, что при увеличении числа n испытаний происходит сближение выборочной функции распределения $\widehat{F}_n(x)$ с функцией распределения $F(x)$ СВ X. Последние два свойства позволяют оценить скорость этого сближения в зависимости от объема n выборки Z_n. Утверждение 2) называется *теоремой Гливенко–Кантелли*, а утверждение 4) является следствием *теоремы Муавра–Лапласа*.

16.4. Гистограмма. Рассмотрим процедуру *группировки* выборки. Для этого действительную ось $\mathbb{R}^1 = (-\infty, \infty)$ разделим точками $\alpha_0, \ldots, \alpha_{l+1}$ на $l+1$ непересекающийся полуинтервал (*разряд*) $\Delta_k = [\alpha_k, \alpha_{k+1})$, $k = \overline{0, l}$, таким образом, что $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_l < \alpha_{l+1} = +\infty$, $\alpha_1 \leqslant x^{(1)}$, $\alpha_l \geqslant x^{(n)}$. Обычно длина разрядов Δ_k, $k = \overline{1, l-1}$, выбирается одинаковой, т. е. равной $h_k \stackrel{\Delta}{=} (\alpha_l - \alpha_1)/(l-1)$. Используя реализацию вариационного ряда $x^{(1)} < \ldots < x^{(n)}$, для каждого k-го разряда Δ_k, $k = \overline{1, l-1}$, вычислим частоту попадания элементов реализации выборки в этот разряд. Получаем $\overline{p}_k \stackrel{\Delta}{=} n_k/n$, где n_k — число элементов реализации выборки z_n, попавших в k-й разряд. Если рассмотреть априорную выборку Z_n и случайное число N_k элементов этой выборки, попавших в k-й разряд, то получим набор случайных величин $\widehat{p}_k = N_k/n$.

Таблица 16.1

$[\alpha_1, \alpha_2)$	$\cdots$	$[\alpha_{l-1}, \alpha_l)$
$\overline{p}_1$	$\cdots$	$\overline{p}_{l-1}$

Определение 16.12. Последовательность пар $(\Delta_k, \widehat{p}_k)$, $k = \overline{1, l-1}$, называется *статистическим рядом*, а его реализация $(\Delta_k, \overline{p}_k)$, $k = \overline{1, l-1}$, представляется в виде табл. 16.1.

Изобразим графически статистический ряд.

Определение 16.13. На оси Ox отложим разряды и на них, как на основании, построим прямоугольники с высотой, равной $\overline{p}_k/h_k$, $k = \overline{1, l-1}$. Тогда площадь каждого прямоугольника будет равна $\overline{p}_k$. Полученная фигура называется *столбцовой диаграммой*, а кусочно постоянная функция $\overline{f}_n(x)$, образованная верхними гранями полученных прямоугольников, — *гистограммой* (рис. 16.2).

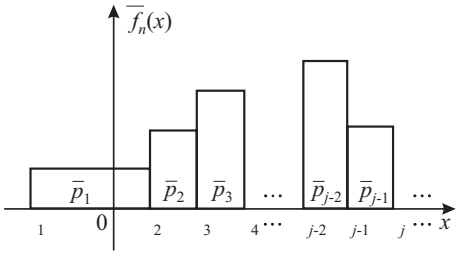

Рис. 16.2

При этом полагают $\overline{f}_n(x) = 0$ для всех $x < \alpha_1$ и $x \geqslant \alpha_l$, так как $n_0 = 0$ и $n_l = 0$.

Пусть плотность распределения $f(x)$ непрерывна и ограничена, а количество разрядов $l_n + 1$ зависит от объема n выборки таким

образом, что $l_n \to \infty$, но при этом $n/l_n \to \infty$ (например, можно выбрать $l_n = c_1 + c_2 \ln n$, где c_1, c_2 — некоторые положительные константы). Тогда выборочная плотность распределения $\widehat{f}_n(x)$, реализациями которой служат гистограммы $\overline{f}_n(x)$, сходится по вероятности к плотности $f(x)$ наблюдаемой СВ, т. е. $\widehat{f}_n(x) \xrightarrow{\mathrm{P}} f(x)$ при $n \to \infty$ для любого $x \in \mathbb{R}^1$. Таким образом, при достаточно «мелком» разбиении отрезка $[\alpha_1, \alpha_l]$ и при большом объеме выборки n высоты построенных прямоугольников можно принимать в качестве приближенных значений плотности $f(x)$ в средних точках соответствующих интервалов. Из этого следует, что гистограмму можно рассматривать как статистический аналог плотности распределения наблюдаемой СВ X. Используя гистограмму, неизвестную плотность можно аппроксимировать кусочно постоянной функцией. Но из математического анализа известно, что если функция является достаточно гладкой, то кусочно линейная аппроксимация оказывается, как правило, лучше кусочно постоянной.

Определение 16.14. Сглаженную гистограмму в виде ломанной, у которой прямые линии последовательно соединяют середины верхних граней прямоугольников, образующих столбцовую диаграмму, называют *полигоном частот*.

16.5. Выборочные моменты. Пусть имеется выборка $Z_n = \mathrm{col}(X_1, \ldots, X_n)$, которая порождена СВ X с функцией распределения $F_X(x)$.

Определение 16.15. Для выборки Z_n объема n *выборочными начальными* и *центральными моментами порядка* r СВ X называются следующие СВ:

$$\widehat{\nu}_r(n) \triangleq \frac{1}{n}\sum_{k=1}^{n}(X_k)^r, \quad r = 1, 2, \ldots;$$

$$\widehat{\mu}_r(n) \triangleq \frac{1}{n}\sum_{k=1}^{n}(X_k - \widehat{\nu}_1(n))^r, \quad r = 2, 3, \ldots$$

Определение 16.16. *Выборочным средним* и *выборочной дисперсией* СВ X называются соответственно

$$\widehat{m}_X(n) \triangleq \widehat{\nu}_1(n) \triangleq \frac{1}{n}\sum_{k=1}^{n}X_k,$$

$$\widehat{d}_X(n) \triangleq \widehat{\mu}_2(n) \triangleq \frac{1}{n}\sum_{k=1}^{n}(X_k - \widehat{m}_X)^2.$$

В дальнейшем мы будем использовать сокращенные обозначения $\widehat{m}_X \triangleq \widehat{m}_X(n)$, $\widehat{d}_X \triangleq \widehat{d}_X(n)$, если это не будет приводить к путанице. Пусть имеется также выборка $V_n \triangleq \operatorname{col}(Y_1, \ldots, Y_n)$, порожденная СВ Y с функцией распределения $F_Y(y)$.

Определение 16.17. *Выборочным коэффициентом корреляции* СВ X и Y называют

$$\widehat{r}_{XY} \triangleq \frac{\sum\limits_{k=1}^{n}(X_k - \widehat{m}_X)(Y_k - \widehat{m}_Y)}{n\sqrt{\widehat{d}_X \widehat{d}_Y}}.$$

Пусть существуют исследуемые моменты ν_r, μ_r. Тогда справедливы следующие свойства.

Свойства выборочных моментов $\widehat{m}_X$

1) $\mathbf{M}[\widehat{\nu}_r(n)] = \nu_r$ для любого $n \geqslant 1$ и для всех $r = 1, 2, \ldots$;

2) $\widehat{\nu}_r(n) \xrightarrow{\text{п.н.}} \nu_r$ при $n \to \infty$ для всех $r = 1, 2, \ldots$;

3) $\widehat{\mu}_r(n) \xrightarrow{\text{п.н.}} \mu_r$ при $n \to \infty$ для всех $r = 2, 3, \ldots$;

4) $\mathbf{D}[\widehat{m}_X] = d_X/n$, где $d_X \triangleq \mathbf{D}[X]$;

5) $\mathbf{M}\left[\widehat{d}_X\right] = \dfrac{n-1}{n}\, d_X$;

6) $(\widehat{m}_X - m_X) \Big/ \sqrt{d_X/n} \xrightarrow{F} U$ при $n \to \infty$, где СВ U имеет распределение $\mathbf{N}(0;1)$;

7) $\left(\widehat{d}_X - d_X\right) \Big/ \sqrt{(\nu_4 - \nu_2^2)/n} \xrightarrow{F} U$ при $n \to \infty$, где СВ U имеет распределение $\mathbf{N}(0;1)$.

Второе и третье свойства указывают, что с увеличением объема выборки выборочные моменты будут сколь угодно близки к соответствующим теоретическим моментам. Пример нахождения объема выборки, гарантирующего в определенном смысле близость выборочной характеристики к ее истинному значению, приведен в § 23.

Из первого свойства вытекает, что математические ожидания (МО) выборочных начальных моментов совпадают с соответствующими значениями начальных моментов СВ X, т. е. в этом смысле обладают свойством «несмещенности». А МО выборочной дисперсии $\widehat{d}_X$ не совпадает с дисперсией d_X СВ X, т. е. в этом смысле СВ $\widehat{d}_X$ является «смещенной» выборочной характеристикой d_X. Поэтому часто вместо $\widehat{d}_X$ используют «*исправленную*» выборочную дисперсию $\widehat{s}_X \triangleq \dfrac{n}{n-1}\, \widehat{d}_X$, для которой $\mathbf{M}[\widehat{s}_X] = d_X$.

16.6. Типовые задачи.

З а д а ч а 16.1. В метеорологии принято характеризовать температуру месяца ее средним значением (среднее значение температуры месяца равно сумме температур всех дней данного месяца, деленной на число дней в этом месяце). В табл. 16.2 приведены значения средней температуры января в г. Саратове и г. Алатыре.

Т а б л и ц а 16.2

Год	1891	1892	1893	1894	1895	1896	1897
Саратов	−19,2	−14,8	−19,6	−11,1	−9,4	−16,9	−13,7
Алатырь	−21,8	−15,4	−20,8	−11,3	−11,6	−19,2	−13,0

Год	1899	1911	1912	1913	1914	1915	
Саратов	−4,9	−13,9	−9,4	−8,3	−7,9	−5,3	
Алатырь	−7,4	−15,1	−14,4	−11,1	−10,5	−7,2	

Требуется по данным реализациям найти: *а*) выборочное среднее и выборочную дисперсию средней температуры января в г. Саратове и г. Алатыре; *б*) выборочный коэффициент корреляции средней температуры января в г. Саратове и средней температуры января в г. Алатыре.

Р е ш е н и е. Пусть СВ X — средняя температура января в г. Саратове, а СВ Y — средняя температура января в г. Алатыре. В таблице приведена реализация $x_1, \ldots, x_{13}$ выборки $X_1, \ldots, X_{13}$, порожденной СВ X, и реализация $y_1, \ldots, y_{13}$ выборки $Y_1, \ldots, Y_{13}$, порожденной СВ Y. Выборочное среднее $\widehat{m}_X$ СВ X для данной реализации $x_1, \ldots, x_{13}$ равно

$$\widehat{m}_X = \frac{1}{13}\sum_{i=1}^{13} x_i \approx -11{,}87,$$

а выборочное среднее $\widehat{m}_Y$ СВ Y равно

$$\widehat{m}_Y = \frac{1}{13}\sum_{i=1}^{13} y_i \approx -13{,}75.$$

Выборочная дисперсия $\widehat{d}_X$ СВ X для данной реализации $(x_1, \ldots, x_{13})$ равна

$$\widehat{d}_X = \frac{1}{13}\sum_{i=1}^{13} \left(x_i - (-11{,}87)\right)^2 \approx 22{,}14,$$

а выборочная дисперсия $\widehat{d}_Y$ СВ Y равна

$$\widehat{d}_Y = \frac{1}{13}\sum_{i=1}^{13} \left(y_i - (-13{,}75)\right)^2 \approx 20{,}09.$$

Выборочный коэффициент корреляции СВ X и Y:

$$\widehat{r}_{XY} = \frac{\sum\limits_{i=1}^{13} (x_i - (-11{,}87))\,(y_i - (-13{,}75))}{13\sqrt{22{,}14}\sqrt{20{,}09}} \approx 0{,}95.$$

Ответ. $\widehat{m}_X \approx -11{,}87$, $\widehat{m}_Y \approx -13{,}75$, $\widehat{d}_X \approx 22{,}14$, $\widehat{d}_Y \approx 20{,}09$, $\widehat{r}_{XY} \approx \approx 0{,}95$.

Задача 16.2. В 1889–1890 гг. был измерен рост 1000 взрослых мужчин (рабочих московских фабрик). Результаты измерений представлены в табл. 16.3 (данные взяты из [10]). По имеющимся наблюдениям требуется построить гисторамму.

Таблица 16.3

рост [см]	143–146	146–149	149–152	152–155	155–158
число мужчин	1	2	8	26	65
рост [см]	158–161	161–164	164–167	167–170	170–173
число мужчин	120	180	201	170	120
рост [см]	173–176	176–179	179–182	182–185	185–188
число мужчин	64	28	10	3	1

Решение. Пусть СВ X — рост взрослого мужчины. Тогда приведенная таблица содержит реализацию выборки $X_1, \ldots, X_{1000}$, порожденной СВ X.

В данной задаче группировка выборки уже проведена: действительная ось $\mathbb{R}^1$ разделена на 17 полуинтервалов Δ_k, $k = \overline{0,16}$, где $\Delta_0 = (-\infty, 143)$, $\Delta_{16} = [188, +\infty)$, а остальные 15 полуинтервалов имеют одинаковую длину $h = 3$. Во второй строке таблицы приведены числа n_k, $k = \overline{1,15}$, равные количеству элементов выборки, попавших в k-й разряд. Вычислим частоту попадания в k-й полуинтервал, $k = \overline{1,15}$: $\overline{p}_k = n_k/1000$, и построим реализацию статистического ряда (см. табл. 16.4).

Таблица 16.4

Δ_k	143–146	146–149	149–152	152–155	155–158
$\overline{p}_k$	0,0003	0,0006	0,002	0,008	0,022
Δ_k	158–161	161–164	164–167	167–170	170–173
$\overline{p}_k$	0,04	0,06	0,067	0,057	0,04
Δ_k	173–176	176–179	179–182	182–185	185–188
$\overline{p}_k$	0,021	0,009	0,003	0,001	0,0003

Теперь на оси OX отложим разряды Δ_k, $k = \overline{1,15}$, и на них, как на основании, построим прямоугольники высотой $\overline{p}_k$ (рис. 16.3).

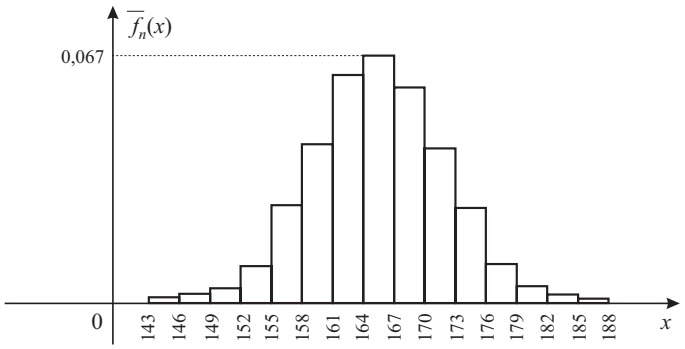

Рис. 16.3

§ 17. Основные распределения в статистике

17.1. Распределение хи-квадрат.

Определение 17.1. Пусть U_k, $k = \overline{1,n}$, — набор из n независимых нормально распределенных СВ, $U_k \sim \mathbf{N}(0;1)$. Тогда СВ

$$X_n \triangleq \sum_{k=1}^{n} U_k^2$$

имеет *распределение хи-квадрат* (χ^2-*распределение*) с n *степенями свободы*, что обозначается как $X_n \sim \chi^2(n)$.

Свойства распределения хи-квадрат $\chi^2(n)$

1) СВ X_n имеет следующую плотность распределения:

$$f(x,n) = \begin{cases} \dfrac{1}{2^{(n/2)}\Gamma(n/2)}\, x^{(n/2)-1} e^{-x/2}, & x > 0, \\ 0, & x \leqslant 0, \end{cases}$$

где $\Gamma(m) \triangleq \int\limits_0^{+\infty} y^{m-1} e^{-y}\, dy$ — гамма-функция. Графики функций $f(x,n)$ (рис. 17.1), называемые *кривыми Пирсона*, асимметричны и начиная с $n > 2$ имеют один максимум в точке $x = n-2$.

2) Характеристическая функция СВ X_n имеет вид

$$g(t,n) = \int\limits_{-\infty}^{+\infty} e^{itx} f(x,n)\, dx = (1-2ti)^{-n/2}.$$

3) СВ $X_n \sim \chi^2(n)$ имеет следующие моменты:

$$\mathbf{M}[X_n] = n, \qquad \mathbf{D}[X_n] = 2n.$$

4) Сумма любого числа m независимых СВ X_k, $k = \overline{1,m}$, имеющих распределение хи-квадрат с n_k степенями свободы, имеет распределение хи-квадрат с $n \triangleq \sum_{k=1}^{m} n_k$ степенями свободы.

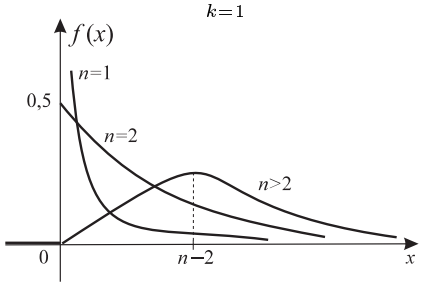

Рис. 17.1

5) Распределение хи-квадрат обладает свойством *асимптотической нормальности*:

$$\frac{X_n - n}{\sqrt{2n}} \xrightarrow{F} U \qquad \text{при } n \to \infty,$$

где СВ U имеет распределение $\mathbf{N}(0;1)$. Это означает, что при достаточно большом объеме n выборки можно приближенно считать $X_n \sim \mathbf{N}(n;2n)$. Фактически эта аппроксимация имеет место уже при $n \geqslant 30$.

Пример 17.1. Приведем пример, в котором возникает распределение хи-квадрат. Пусть выборка Z_n соответствует нормальному распределению $\mathbf{N}(m;\sigma^2)$. Рассмотрим выборочную дисперсию

$$\widehat{d}_X = \frac{1}{n}\sum_{k=1}^{n}(X_k - \widehat{m}_X)^2,$$

где $\widehat{m}_X$ — выборочное среднее. Тогда СВ $Y_n \triangleq n\widehat{d}_X \Big/ \sigma^2$ имеет распределение $\chi^2(n-1)$ и не зависит от $\widehat{m}_X$.

17.2. Распределение Стьюдента.

Определение 17.2. Пусть U и X_n — независимые СВ, $U \sim \mathbf{N}(0;1)$, $X_n \sim \chi^2(n)$. Тогда СВ $T_n \triangleq U \Big/ \sqrt{X_n/n}$ имеет *распределение Стьюдента с n степенями свободы*, что обозначают как $T_n \sim \mathbf{S}(n)$.

Свойства распределения Стьюдента $\mathbf{S}(n)$

1) СВ T_n имеет плотность распределения

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

Графики плотностей $f(t,n)$ (рис. 17.2), называемые *кривыми Стьюдента*, симметричны при всех $n=1,2,\ldots$ относительно оси ординат.

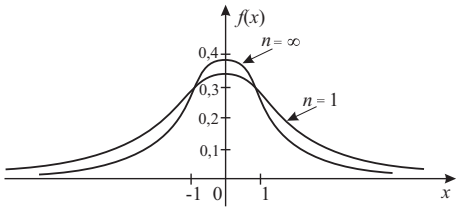

Рис. 17.2

2) СВ T_n имеет МО, равное $\mathbf{M}[T_n]=0$ для всех $n\geqslant 2$, и дисперсию $\mathbf{D}[T_n]=n/(n-2)$ при $n>2$. При $n=2$ дисперсия $\mathbf{D}[T_n]=+\infty$.

3) При $n=1$ распределение Стьюдента совпадает с *распределением Коши*, плотность которого равна

$$f(t,1)=\frac{1}{\pi}\frac{1}{1+t^2}.$$

Но, как известно, математическое ожидание и дисперсия СВ T_1, имеющей распределение Коши, не существуют, так как бесконечен предел

$$\lim_{a\to\infty} I(a)=+\infty,$$

где $I(a)\triangleq\dfrac{1}{\pi}\displaystyle\int_0^a\frac{t}{t^2+1}\,dt.$

4) Можно показать, что при $n\to\infty$ распределение $\mathbf{S}(n)$ асимптотически нормально, т. е. $T_n\xrightarrow{F}U$, где СВ U имеет распределение $\mathbf{N}(0;1)$. При $n\geqslant 30$ распределение Стьюдента $\mathbf{S}(n)$ практически не отличается от $\mathbf{N}(0;1)$.

Пример 17.2. Приведем пример, в котором встречается распределение Стьюдента. Пусть выборка Z_n соответствует нормальному распределению $\mathbf{N}(m;\sigma^2)$. Пусть $\widehat{m}_X$ — выборочное среднее, а $\widehat{d}_X$ —

выборочная дисперсия. Тогда СВ

$$T_n = \sqrt{n-1}\,\frac{\widehat{m}_X - m}{\sqrt{\widehat{d}_X}}$$

имеет распределение Стьюдента $\mathbf{S}(n-1)$.

17.3. Распределение Фишера.

Определение 17.3. Пусть независимые СВ X_n и X_m имеют распределения хи-квадрат соответственно с n и m степенями свободы. Тогда СВ $V_{n,m} \stackrel{\Delta}{=} \dfrac{X_n/n}{X_m/m}$ имеет *распределение Фишера с n и m степенями свободы*, что записывают как $V_{n,m} \sim \mathbf{F}(n;m)$.

Свойства распределения Фишера $\mathbf{F}(n;m)$

1) СВ $V_{n,m}$ имеет плотность $f(v,n,m) = 0$ при $v \leqslant 0$ и

$$f(v,n,m) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}\, n^{\frac{n}{2}} m^{\frac{m}{2}} \frac{v^{\frac{n}{2}-1}}{(m+nv)^{\frac{n+m}{2}}} \quad \text{при } v > 0.$$

Графики функции $f(v,n,m)$, называемые *кривыми Фишера*, асимметричны и при $n > 2$ достигают максимальных значений в точках $v = \dfrac{(n-2)m}{(m+2)n}$, близких к единице при больших значениях m и n. Типовой вид кривой Фишера приведен на рис. 17.3.

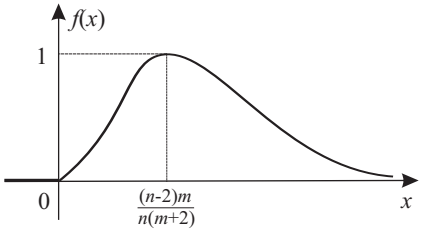

Рис. 17.3

2) СВ $V_{n,m}$ имеет следующие моменты:

$$\mathbf{M}[V_{n,m}] = \frac{m}{m-2} \quad \text{при } m > 2,$$

$$\mathbf{D}[V_{n,m}] = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)} \quad \text{при } m > 4.$$

Пример 17.3. Пусть $Z_n = \mathrm{col}(X_1, \ldots, X_n)$ — выборка объема n, порожденная СВ X с нормальным распределением $\mathbf{N}(m_X;\sigma^2)$, а

$W_n = \mathrm{col}(Y_1, \ldots, Y_m)$ — выборка объема m, порожденная СВ Y с нормальным распределением $\mathbf{N}(m_Y; \sigma^2)$, и СВ Z_n и W_n независимы. Тогда СВ, образованная отношением исправленных выборочных дисперсий СВ X и Y, т. е.

$$V_{n,m} \triangleq \frac{\frac{1}{n-1}\sum\limits_{k=1}^{n}(X_k - \widehat{m}_X)^2}{\frac{1}{m-1}\sum\limits_{j=1}^{m}(Y_j - \widehat{m}_Y)^2},$$

имеет распределение $\mathbf{F}(n-1; m-1)$.

§ 18. Точечные оценки

18.1. Основные понятия.

Определение 18.1. *Параметром распределения* $\theta \in \Theta \subset \mathbb{R}^1$ СВ X называется любая числовая характеристика этой СВ (математическое ожидание, дисперсия и т. п.) или любая константа, явно входящая в выражение для функции распределения.

В общем случае будем предполагать, что параметр распределения θ может быть векторным, т. е. $\theta \in \Theta \subset \mathbb{R}^s$.

В случае параметрической статистической модели $(S_\theta, F_{Z_n}(z_n, \theta))$ таким параметром распределения может служить неизвестный вектор $\theta \in \Theta \subset \mathbb{R}^s$, характеризующий распределение $F_{Z_n}(z_n, \theta)$.

Пусть имеется выборка $Z_n = \mathrm{col}(X_1, \ldots, X_n)$ с реализацией $z_n = \mathrm{col}(x_1, \ldots, x_n)$.

Определение 18.2. *Точечной* (*выборочной*) *оценкой* неизвестного параметра распределения $\theta \in \Theta \subset \mathbb{R}^s$ называется произвольная статистика $\widehat{\theta}(Z_n)$, построенная по выборке Z_n и принимающая значения в множестве Θ.

Замечание 18.1. Реализацию $\widehat{\theta}(z_n)$ оценки $\widehat{\theta}(Z_n)$ принимают, как правило, за приближенное значение неизвестного параметра θ.

Ясно, что существует много разных способов построения точечной оценки, которые учитывают тип статистической модели. Для параметрической и непараметрической моделей эти способы могут быть различны. Рассмотрим некоторые свойства, которые характеризуют качество введенной оценки.

Определение 18.3. Оценка $\widehat{\theta}(Z_n)$ параметра θ называется *несмещенной*, если ее МО равно θ, т. е. $\mathbf{M}\left[\widehat{\theta}(Z_n)\right] = \theta$ для любого $\theta \in \Theta$.

Определение 18.4. Оценка $\widehat{\theta}(Z_n)$ параметра θ называется *состоятельной*, если она сходится по вероятности к θ, т. е. $\widehat{\theta}(Z_n) \xrightarrow{\mathrm{P}} \theta$ при $n \to \infty$ для любого $\theta \in \Theta$.

Определение 18.5. Оценка $\widehat{\theta}(Z_n)$ параметра θ называется *сильно состоятельной*, если она сходится почти наверное к θ, т. е. $\widehat{\theta}(Z_n) \xrightarrow{\text{п.н.}} \theta$ при $n \to \infty$ для любого $\theta \in \Theta$.

Очевидно, что если оценка сильно состоятельная, то она является также состоятельной.

Пример 18.1. Оценка $\widehat{\theta}_1(Z_n) \stackrel{\Delta}{=} \widehat{m}_X$ неизвестного МО $\theta_1 \stackrel{\Delta}{=} m_X$ СВ X является несмещенной (по свойству 1) $\widehat{m}_X$), а оценка $\widehat{\theta}_2(Z_n) \stackrel{\Delta}{=} \widehat{d}_X$ неизвестной дисперсии $\theta_2 \stackrel{\Delta}{=} d_X$ — смещенной, так как $\mathbf{M}\left[\widehat{d}_X\right] = \dfrac{n-1}{n}\, d_X$ (по свойству 5) $\widehat{m}_X$). Оценка $\widehat{\theta}_2(Z_n) \stackrel{\Delta}{=} \widehat{s}_X$ является несмещенной оценкой d_X по определению $\widehat{s}_X$. Если существуют моменты m_X, d_X, то сильная состоятельность оценок $\widehat{m}_X$, $\widehat{d}_X$ гарантируется свойствами 2), 3) $\widehat{m}_X$.

Свойствами состоятельности и несмещенности могут обладать сразу несколько оценок неизвестного параметра θ.

Определение 18.6. Несмещенная оценка $\widehat{\theta}^*(Z_n)$ скалярного параметра θ называется *эффективной*, если $\mathbf{D}\left[\widehat{\theta}^*(Z_n)\right] \leqslant \mathbf{D}\left[\widehat{\theta}(Z_n)\right]$ для всех несмещенных оценок $\widehat{\theta}(Z_n)$ параметра θ, т. е. ее дисперсия минимальна по сравнению с дисперсиями других несмещенных оценок при одном и том же объеме n выборки Z_n.

Пример 18.2. Пусть СВ X имеет нормальное распределение $\mathbf{N}(m_X; \sigma_X^2)$ с неизвестными параметрами $\theta_1 \stackrel{\Delta}{=} m_X$, $\theta_2 \stackrel{\Delta}{=} \sigma_X^2$. В этом случае выборочное среднее $\widehat{m}_X$ является эффективной оценкой m_X.

В классе параметрических моделей $(S_\theta, F_{Z_n}(z_n, \theta))$, $\theta \in \Theta \subset \subset \mathbb{R}^1$, рассмотрим подкласс статистических моделей, удовлетворяющих некоторым естественным условиям регулярности. С этой целью введем следующие понятия.

Определение 18.7. *Функцией правдоподобия* для неизвестного параметра $\theta \in \Theta \subset \mathbb{R}^s$ называется: в случае непрерывной наблюдаемой СВ X — плотность распределения

$$L(z_n, \theta_1, \ldots, \theta_s) \stackrel{\Delta}{=} f_{Z_n}(z_n, \theta_1, \ldots, \theta_s) = \prod_{k=1}^{n} f_X(x_k, \theta_1, \ldots, \theta_s),$$

где $f_X(x, \theta_1, \ldots, \theta_s)$ — плотность распределения СВ X, а в случае

дискретной наблюдаемой СВ X — произведение вероятностей

$$L(z_n, \theta_1, \dots, \theta_s) \triangleq \prod_{k=1}^{n} p_X(x_k, \theta_1, \dots, \theta_s),$$

где $p_X(x_k, \theta_1, \dots, \theta_s)$ — вероятность события $\{X = x_k\}$.

Аналогично определяется функция правдоподобия $L(z_n, \theta_1, \dots, \theta_s)$ при неоднородной выборке $Z_n \triangleq \operatorname{col}(X_1, \dots, X_n)$, когда СВ X_k, $k = \overline{1,n}$, по-прежнему независимы, но имеют различные плотности распределения $f_{X_k}(x_k, \theta_1, \dots, \theta_s)$, зависящие от одного и того же набора неизвестных параметров $\theta_1, \dots, \theta_s$.

Определение 18.8. *Логарифмической функцией правдоподобия* для неизвестного параметра $\theta \in \Theta \subset \mathbb{R}^s$ называется функция $\ln L(z_n, \theta_1, \dots, \theta_s)$.

Определение 18.9. Параметрическая статистическая модель $(S_\theta, F_{Z_n}(z_n, \theta))$, $\theta \in \Theta \subset \mathbb{R}^1$, называется *регулярной*, если выполняются следующие условия:

1) функция правдоподобия $L(z_n, \theta) > 0$ для всех $\theta \in \Theta$ и $z_n \in S_\theta$ дифференцируема по параметру $\theta \in \Theta$;

2) для любого измеримого множества $A \subset S$ выполняется условие

$$\frac{\partial}{\partial \theta} \int\limits_A L(z_n, \theta)\, dz_n = \int\limits_A \frac{\partial}{\partial \theta} L(z_n, \theta)\, dz_n.$$

Из условия 2) вытекает, в частности, что в случае регулярной модели выборочное пространство $S_\theta = S$, т. е. не зависит от неизвестного параметра θ.

Пример 18.3. Пусть выборка Z_n соответствует равномерному распределению $\mathbf{R}(0; b)$ с неизвестным параметром $\theta \triangleq b$. В этом случае параметрическая статистическая модель $(S_\theta, F_{Z_n}(z_n, \theta))$ не является регулярной, так как выборочное пространство S_θ определяется на отрезках $[0, b]$, а следовательно, зависит от параметра b.

Определение 18.10. В случае регулярной статистической модели $(S, F_{Z_n}(z_n, \theta))$, $\theta \in \Theta \subset \mathbb{R}^1$, величина

$$I_n(\theta) \triangleq \mathbf{M}\left[\left(\frac{\partial \ln L(Z_n, \theta)}{\partial \theta}\right)^2\right]$$

называется *информацией Фишера* о параметре $\theta \in \Theta$, содержащейся в выборке Z_n.

В случае регулярной модели $(S, F_{Z_n}(z_n, \theta))$ для любой несмещенной оценки $\widehat{\theta}(Z_n)$ параметра $\theta \in \Theta \subset \mathbb{R}^1$ справедливо *неравенство Рао–Крамера*

$$\mathbf{D}\left[\widehat{\theta}(Z_n)\right] \geqslant \frac{1}{I_n(\theta)}.$$

Это неравенство дает нижнюю границу для дисперсии несмещенной оценки.

Определение 18.11. Несмещенная оценка $\widehat{\theta}(Z_n)$ параметра $\theta \in \Theta \subset \mathbb{R}^1$ называется *R-эффективной оценкой*, если для этой оценки в неравенстве Рао–Крамера достигается равенство, т. е. $\mathbf{D}\left[\widehat{\theta}(Z_n)\right] = 1/I_n(\theta)$.

Если R-эффективная оценка существует, то она является также эффективной в смысле минимума дисперсии (см. определение 18.6).

Способ построения R-эффективных оценок вытекает из *критерия эффективности*, состоящего в следующем. Оценка $\widehat{\theta}(Z_n)$ параметра θ является R-эффективной тогда и только тогда, когда

$$\widehat{\theta}(Z_n) - \theta = \frac{1}{I_n(\theta)} \sum_{k=1}^{n} \frac{\partial \ln f(X_k, \theta)}{\partial \theta}.$$

Таблица 18.1

Модель	$I_n(\theta)$	$\widehat{\theta}(Z_n)$	$\mathbf{D}[\widehat{\theta}(Z_n)]$
$\mathbf{N}(\theta; \sigma_X^2)$	$\frac{n}{\sigma_X^2}$	$\widehat{m}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\frac{\sigma_X^2}{n}$
$\mathbf{N}(m_X; \theta)$	$\frac{n}{2\theta^2}$	$\widehat{d}_X = \frac{1}{n}\sum_{i=1}^{n}(X_i - m_X)^2$	$\frac{2\theta^2}{n}$
$\mathbf{Bi}(k; \theta)$	$\frac{kn}{\theta(1-\theta)}$	$\frac{\widehat{m}_X}{k} = \frac{1}{nk}\sum_{i=1}^{n} X_i$	$\frac{\theta(1-\theta)}{kn}$
$\Pi(\theta)$	$\frac{n}{\theta}$	$\widehat{m}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\frac{\theta}{n}$

Пример 18.4. Приведем примеры R-эффективных (а следовательно, и эффективных) оценок $\widehat{\theta}(Z_n)$ неизвестных параметров θ некоторых распространенных распределений, их дисперсии $\mathbf{D}\left[\widehat{\theta}(Z_n)\right]$, а также значения информациии Фишера $I_n(\theta)$ (см. табл. 18.1).

Пример 18.5. Приведем пример, когда эффективная оценка неизвестного параметра существует, а R-эффективная оценка не существует. Пусть выборка Z_n соответствует распределению $\mathbf{N}(m;\sigma^2)$ с неизвестными параметрами $\theta_1 \triangleq m$, $\theta_2 \triangleq \sigma^2$. В данном случае параметрическая модель является регулярной. Из п. 16.5 следует, что статистика

$$\widehat{s}_X(Z_n) = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \widehat{m}_X)^2$$

является несмещенной оценкой неизвестной дисперсии $d_X = \sigma^2$. Как указано в примере 17.1 из п. 17.1, СВ

$$Y_n \triangleq \frac{(n-1)\widehat{s}_X(Z_n)}{\sigma^2}$$

имеет распределение хи-квадрат $\chi^2(n-1)$, а следовательно, её дисперсия равна $\mathbf{D}[Y_n] = 2(n-1)$. Поэтому

$$\mathbf{D}\left[\widehat{s}_X(Z_n)\right] = \frac{2\sigma^4}{n-1}.$$

Согласно неравенству Рао-Крамера нижняя граница дисперсии любой несмещенной оценки в этой статистической модели равна $2\sigma^4/n$. Таким образом, $\widehat{s}_X(Z_n)$ не является R-эффективной оценкой. Но можно показать, что эта оценка является эффективной, т. е. $\widehat{s}_X(Z_n)$ имеет минимальную дисперсию. Таким образом, нижняя граница в неравенстве Рао-Крамера не достигается, а следовательно, R-эффективной оценки не существует.

18.2. Метод максимального правдоподобия. Как уже отмечалось выше, на практике часто удается предсказать вид распределения наблюдаемой СВ с точностью до неизвестных параметров $\theta \triangleq \mathrm{col}(\theta_1,\ldots,\theta_s)$, т. е. для непрерывной СВ X оказывается известной плотность $f(x,\theta)$, а для дискретной СВ X — вероятности $p(x_i,\theta) \triangleq \mathbf{P}_\theta\{X = x_i\}$, $i = \overline{1,m}$, где x_i, $i = \overline{1,m}$, — возможные значения СВ X. Например, может быть $\theta_1 = m_X$, $\theta_2 = d_X$ при $s = 2$. Эти неизвестные параметры требуется оценить по имеющейся выборке Z_n.

Рассмотрим параметрическую статистическую модель $(S_\theta, F_{Z_n}(z_n,\theta))$, $\theta \in \Theta \subset \mathbb{R}^s$, для которой известна функция правдоподобия $L(z_n,\theta_1,\ldots,\theta_s)$.

Определение 18.12. *Оценкой максимального правдоподобия* (МП-*оценкой*) параметра $\theta \in \Theta$ называется статистика $\widehat{\theta}(Z_n)$, максимизирующая для каждой реализации z_n функцию правдоподобия,

т. е.

$$\widehat{\theta}(z_n) = \arg\max_{\theta\in\Theta} L(z_n, \theta).$$

Способ построения МП-оценки называется *методом максимального правдоподобия.*

Поскольку функция правдоподобия $L(z_n, \theta)$ и её логарифм $\ln L(z_n, \theta)$ достигают максимума при одних и тех же значениях θ, то часто вместо $L(z_n, \theta)$ рассматривают логарифмическую функцию правдоподобия $\ln L(z_n, \theta)$.

В случае дифференцируемости функции $\ln L(z_n, \theta)$ по θ МП-оценку можно найти, решая относительно $\theta_1, \ldots, \theta_s$ систему *уравнений правдоподобия*

$$\frac{\partial \ln L(z_n, \theta_1, \ldots, \theta_s)}{\partial \theta_1} = 0, \quad \ldots, \quad \frac{\partial \ln L(z_n, \theta_1, \ldots, \theta_s)}{\partial \theta_s} = 0.$$

Отметим, что в тех случаях, когда существует R-эффективная оценка параметра θ, она совпадает с МП-оценкой.

Пример 18.6. Если, например, СВ X имеет нормальное распределение $\mathbf{N}(m_X; \sigma_X^2)$ с неизвестным математическим ожиданием $\theta \triangleq m_X$, то легко установить, что оценкой максимального правдоподобия параметра $\theta = m_X$ при любых σ_X является выборочное среднее $\widehat{m}_X$. Действительно, в этом случае

$$L(z_n, m_X) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left\{-\frac{(x_k - m_X)^2}{2\sigma_X^2}\right\} =$$

$$= \frac{1}{(2\pi)^{n/2}\sigma_X^n} \exp\left\{-\sum_{k=1}^{n} \frac{(x_k - m_X)^2}{2\sigma_X^2}\right\}.$$

Дифференцируя функцию $\ln L(z_n, m_X)$ по m_X и приравнивая нулю получаемое выражение, находим уравнение, решение которого

$$\widehat{\theta}(Z_n) = \frac{1}{n}\sum_{k=1}^{n} X_k \triangleq \widehat{m}_X.$$

Пример 18.7. Пусть СВ X имеет равномерное распределение $\mathbf{R}(a; b)$ с неизвестными параметрами $\theta_1 \triangleq a$, $\theta_2 \triangleq b$. В данном случае плотность распределения $f(x, a, b) = 1/(b - a)$, если $x \in [a, b]$, и $f(x, a, b) = 0$, если $x \notin [a, b]$. Оценим параметры a и b методом максимального правдоподобия. Функция правдоподобия в данном случае имеет вид

$$L(x_1, \ldots, x_n, a, b) = \prod_{k=1}^{n} f(x_k, a, b) = \left(\frac{1}{b - a}\right)^n,$$

если $a \leqslant x_k \leqslant b$ для всех $k = \overline{1,n}$, и $L(x_1, \ldots, x_n, a, b) = 0$ в остальных случаях. Таким образом, функция правдоподобия отлична от нуля, если неизвестные параметры a и b удовлетворяют неравенствам

$$b \geqslant x^{(n)} \triangleq \max_{k=\overline{1,n}} x_k, \qquad a \leqslant x^{(1)} \triangleq \min_{k=\overline{1,n}} x_k.$$

При этом функция $L(x_1, \ldots, x_n, a, b)$ достигает максимума по a и b, когда разность $b - a$ оказывается минимально возможной, не нарушающей полученные неравенства, т. е. в случае достижения в них равенств. Таким образом получаем МП-оценки неизвестных параметров a и b:

$$\widehat{a}(Z_n) = X^{(1)}, \qquad \widehat{b}(Z_n) = X^{(n)},$$

где $X^{(n)}$ и $X^{(1)}$ — крайние члены вариационного ряда.

Как отмечалось выше, данная параметрическая модель не является регулярной, поэтому в данной модели R-эффективные оценки не существуют.

Пример 18.8. Пусть случайная величина X имеет распределение Пуассона $\mathbf{\Pi}(a)$ с неизвестным параметром $\theta = a$. Построим МП-оценку параметра a. Функция правдоподобия в этом случае равна

$$L(z_n, a) = \prod_{k=1}^{n} \frac{a^{x_k} e^{-a}}{x_k!},$$

поэтому логарифмическая функция правдоподобия равна

$$\ln L(z_n, a) = \sum_{k=1}^{n} x_k \ln a - na - \ln(x_1! \cdot \ldots \cdot x_n!).$$

Решая соответствующее уравнение правдоподобия

$$\frac{\partial}{\partial a} \ln L(z_n, a) = 0,$$

находим МП-оценку неизвестного параметра a:

$$\widehat{a}(Z_n) = \frac{1}{n} \sum_{k=1}^{n} X_k \triangleq \widehat{m}_X.$$

Пример 18.9. Найдем МП-оценку для вероятности p «успеха» в схеме испытаний Бернулли. В этом случае имеем распределение Бернулли $\mathbf{Bi}(1; p)$ с неизвестным параметром $\theta = p$. Поэтому

$$L(z_n, a) = \prod_{k=1}^{n} p^{x_k} (1-p)^{1-x_k},$$

где $x_k = 1$, если в k-м испытании был «успех», и $x_k = 0$ — в противном случае. Решая уравнение правдоподобия

$$\frac{\partial}{\partial p} \ln L(z_n, p) = \sum_{k=1}^{n} \left(\frac{x_k}{p} - \frac{1 - x_k}{1 - p} \right) = 0$$

относительно параметра p, находим МП-оценку

$$\widehat{p}(Z_n) = \frac{1}{n} \sum_{k=1}^{n} X_k \triangleq \widehat{m}_X.$$

18.3. Метод моментов. Исторически первым для оценивания неизвестных параметров был предложен следующий метод. Пусть имеется параметрическая статистическая модель $(S_\theta, F_{Z_n}(z_n, \theta))$, $\theta \in \Theta \subset \mathbb{R}^s$. Предположим, что у наблюдаемой СВ X, порождающей выборку Z_n, существуют начальные моменты $\nu_i = \mathbf{M}[X^i]$, $i = \overline{1, s}$. Тогда в общем случае от неизвестных параметров будут зависеть и начальные моменты, т. е. $\nu_i = \nu_i(\theta)$.

Пусть $\widehat{\nu}_i$, $i = \overline{1, s}$, — выборочные начальные моменты. Рассмотрим систему уравнений

$$\nu_i(\theta) = \widehat{\nu}_i, \quad i = \overline{1, s}$$

и предположим, что её можно разрешить относительно параметров $\theta_1, \ldots, \theta_s$, т. е. найти функции $\widehat{\theta}_i = \varphi_i(\widehat{\nu}_1, \ldots, \widehat{\nu}_s)$, $i = \overline{1, s}$.

Определение 18.13. Решение полученной системы уравнений $\widehat{\theta}_i = \varphi_i(\widehat{\nu}_1, \ldots, \widehat{\nu}_s)$, $i = \overline{1, s}$, называется *оценкой* параметра θ, найденной по *методу моментов*, или ММ-*оценкой*.

Если функции $\varphi_1(\cdot), \ldots, \varphi_s(\cdot)$ непрерывны, то ММ-оценки являются состоятельными.

Замечание 18.2. Уравнения метода моментов часто оказываются более простыми по сравнению с уравнениями правдоподобия, и их решение не связано с большими вычислительными трудностями.

Пример 18.10. Пусть $Z_n \triangleq \mathrm{col}(X_1, \ldots, X_n)$ — выборка, соответствующая нормальному распределению $\mathbf{N}(m; \sigma^2)$ с неизвестными параметрами $\theta_1 \triangleq m$ и $\theta_2 \triangleq \sigma^2$. Оценим параметры m и σ^2 с помощью метода моментов. В данном случае $\nu_1 = m$, $\nu_2 = m^2 + \sigma^2$ и система уравнений для метода моментов принимает вид

$$\begin{cases} m = \widehat{\nu}_1 \triangleq \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} X_k, \\ m^2 + \sigma^2 = \widehat{\nu}_2 \triangleq \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} X_k^2. \end{cases}$$

Решая эту систему, находим ММ-оценки

$$\widehat{\theta}_1(Z_n) = \widehat{\nu}_1, \qquad \widehat{\theta}_2(Z_n) = \widehat{\nu}_2 - \widehat{\nu}_1^2.$$

Пример 18.11. Пусть $Z_n = \mathrm{col}(X_1, \ldots, X_n)$ — выборка, соответствующая равномерному распределению $\mathbf{R}(a; b)$ с неизвестными параметрами $\theta_1 \triangleq a$ и $\theta_2 \triangleq b$. Оценим параметры a и b с помощью метода моментов. Поскольку для данного распределения

$$\nu_1 = \frac{a+b}{2}, \qquad \nu_2 = \frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2,$$

то система уравнений метода моментов принимает вид

$$\begin{cases} \dfrac{a+b}{2} = \widehat{\nu}_1 \triangleq \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k, \\ \dfrac{(b-a)^2}{12} + \left(\dfrac{a+b}{2}\right)^2 = \widehat{\nu}_2 \triangleq \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k^2. \end{cases}$$

Решая эту систему, получаем ММ-оценки неизвестных параметров:

$$\widehat{a}(Z_n) = \widehat{\nu}_1 - \sqrt{3\widehat{d}_X}, \qquad \widehat{b}(Z_n) = \widehat{\nu}_1 + \sqrt{3\widehat{d}_X},$$

где $\widehat{d}_X \triangleq \dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(X_k - \widehat{m}_X)^2$.

Оценки, полученные в примере 18.10 с помощью метода моментов, совпадают с МП-оценками, найденными в примере 18.6 из п. 18.2, а ММ-оценки в примере 18.11 не совпадают с МП-оценками, построенными в примере 18.7 из того же пункта.

§ 19. Интервальные оценки

19.1. Основные понятия. Пусть имеется параметрическая статистическая модель $(S_\theta, F_{Z_n}(z_n, \theta))$, $\theta \in \Theta \subset \mathbb{R}^1$, и по выборке $Z_n = \mathrm{col}(X_1, \ldots, X_n)$, соответствующей распределению $F(x, \theta)$ наблюдаемой СВ X, требуется определить неизвестный параметр θ. Вместо точечных оценок, рассмотренных ранее, рассмотрим другой тип оценок неизвестного параметра $\theta \in \Theta \subset \mathbb{R}^1$.

Определение 19.1. Интервал $[\theta_1(Z_n), \theta_2(Z_n)]$ со случайными концами, «накрывающий» с вероятностью $1 - \alpha$, $0 < \alpha < 1$, неизвестный параметр θ, т. е.

$$\mathbf{P}\{\theta_1(Z_n) \leqslant \theta \leqslant \theta_2(Z_n)\} = 1 - \alpha,$$

называется *доверительным интервалом* (или *интервальной оценкой*) уровня надежности $1-\alpha$ параметра θ.

Аналогично определяется доверительный интервал для произвольной функции от параметра θ.

Определение 19.2. Число $\delta \triangleq 1-\alpha$ называется *доверительной вероятностью* или *уровнем доверия*.

Определение 19.3. Доверительный интервал $[\theta_1(Z_n), \theta_2(Z_n)]$ называется *центральным*, если выполняются следующие условия:

$$\mathbf{P}\{\theta \geqslant \theta_2(Z_n)\} = \frac{\alpha}{2}, \qquad \mathbf{P}\{\theta_1(Z_n) \geqslant \theta\} = \frac{\alpha}{2}.$$

Часто вместо двусторонних доверительных интервалов рассматривают односторонние доверительные интервалы, полагая $\theta_1(Z_n) = -\infty$ или $\theta_2(Z_n) = +\infty$.

Определение 19.4. Интервал, границы которого удовлетворяют условию:

$$\mathbf{P}\{\theta \geqslant \theta_2(Z_n)\} = \alpha \qquad (\text{или} \quad \mathbf{P}\{\theta_1(Z_n) \geqslant \theta\} = \alpha),$$

называется соответственно *правосторонним* (или *левосторонним*) *доверительным интервалом*.

Рассмотрим два различных способа построения доверительных интервалов.

19.2. Использование центральной статистики.

Определение 19.5. Функция $Y \triangleq G(Z_n, \theta)$ случайной выборки Z_n, такая, что её распределение $F_Y(y)$ не зависит от параметра θ и при любом значении z_n функция $G(z_n, \theta)$ является непрерывной и монотонной по θ, называется *центральной статистикой* для параметра θ.

Зная распределение $F_Y(y)$ центральной статистики $Y \triangleq \triangleq G(Z_n, \theta)$, можно найти числа g_1 и g_2, удовлетворяющие условию

$$\mathbf{P}\{g_1 \leqslant G(Z_n, \theta) \leqslant g_2\} = 1-\alpha.$$

Тогда границы доверительного интервала $[\theta_1(Z_n), \theta_2(Z_n)]$ для параметра θ могут быть найдены, если разрешить, учитывая свойства функции $G(z_n, \theta)$, следующие неравенства:

$$g_1 \leqslant G(Z_n, \theta) \leqslant g_2.$$

В частности, если $G(z_n, \theta)$ — монотонно возрастающая по θ функция, то

$$\theta_1(Z_n) = G^{-1}(Z_n, g_1), \qquad \theta_2(Z_n) = G^{-1}(Z_n, g_2),$$

где $G^{-1}(z_n, g_1)$ — функция, обратная по отношению к $G(z_n, \theta)$. Если $G(z_n, \theta)$ — монотонно убывающая по θ функция, то

$$\theta_1(Z_n) = G^{-1}(Z_n, g_2), \qquad \theta_2(Z_n) = G^{-1}(Z_n, g_1).$$

Применим данный метод для построения доверительных интервалов неизвестных параметров нормального распределения $\mathbf{N}(m_X; \sigma_X^2)$. С этой целью сформулируем утверждение, с помощью которого можно определить центральные статистики для неизвестных параметров m_X, σ_X^2. Но перед его формулировкой напомним следующее понятие из линейной алгебры.

Матрица C размерности $n \times n$ с элементами c_{ij}, $i = \overline{1,n}$, $j = \overline{1,n}$, называется *ортогональной*, если $C^T C = I$, где I — единичная матрица размерности $n \times n$, а C^T — *транспонированная матрица* с элементами $c_{ij}^T = c_{ji}$, $i = \overline{1,n}$, $j = \overline{1,n}$.

Рассмотрим ортогональную матрицу C специального вида, у которой первая строка состоит из элементов $c_{1i} = 1/\sqrt{n}$, $i = \overline{1,n}$, а остальные строки — произвольные, ненарушающие ортогональность матрицы C. Такая матрица всегда существует. Рассмотрим свойства введенной матрицы.

Свойства матрицы C

1) $\sum\limits_{j=1}^{n} c_{kj} c_{ij} = 0$ для всех $k \neq i$, так как $C^T C = I$ и у единичной матрицы все элементы, кроме диагональных, равны нулю.

2) $\sum\limits_{j=1}^{n} c_{kj}^2 = 1$ для всех $k = \overline{1,n}$, так как $C^T C = I$ и диагональные элементы матрицы I равны единице.

3) Если $y = Cx$, где $y = \mathrm{col}(y_1, \ldots, y_n)$ и $x = \mathrm{col}(x_1, ..., x_n)$, то $\sum\limits_{k=1}^{n} y_k^2 = \sum\limits_{k=1}^{n} x_k^2$. Действительно,

$$\sum_{k=1}^{n} y_k^2 = y^T y = (Cx)^T Cx = x^T C^T Cx = x^T x = \sum_{k=1}^{n} x_k^2.$$

4) $\sum\limits_{j=1}^{n} c_{ij} = 0$ для всех $i = \overline{2,n}$. Из свойства 1) C для первой строки $(k = 1)$ при $i > 1$ имеем

$$\sum_{j=1}^{n} c_{ij} = \sqrt{n} \sum_{j=1}^{n} c_{1j} c_{ij} = 0.$$

Теорема 19.1. (*Теорема Фишера, усиленный вариант*). *Пусть* $Z_n \stackrel{\Delta}{=} \operatorname{col}(X_1, \dots, X_n)$ — *выборка, порожденная СВ* $X \sim \mathbf{N}(m_X; \sigma_X^2)$, *а* $\widehat{m}_X$ *и* $\widehat{d}_X$ — *выборочные среднее и дисперсия. Тогда*

1) *СВ* $\overset{*}{M}_X \stackrel{\Delta}{=} (\widehat{m}_X - m_X)\sqrt{n} \Big/ \sigma_X$ *имеет распределение* $\mathbf{N}(0;1)$;

2) *СВ* $\overset{*}{D}_X \stackrel{\Delta}{=} n\widehat{d}_X \,/\sigma_X^2$ *имеет распределение* $\chi^2(n-1)$;

3) $\overset{\circ}{M}_X \stackrel{\Delta}{=} (\widehat{m}_X - m_X)\sqrt{(n-1) \Big/ \widehat{d}_X}$ *имеет распределение* $\mathbf{S}(n-1)$;

4) *СВ* $\widehat{m}_X$ *и* $\widehat{d}_X$ *независимы.*

Доказательство. 1) Докажем вначале первое утверждение. Так как сумма нормальных СВ является нормальной, то $\widehat{m}_X = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$ является нормальной СВ. Кроме того, из свойства 1) $\widehat{m}_X$ вытекает $\mathbf{M}[X] = m_X$. Далее, так как СВ X_i, $i = \overline{1,n}$, независимы и одинаково распределены, то в силу свойства 4)М$[X]$ имеем $\mathbf{D}[\widehat{m}_X] = \sigma_X^2/n$. Таким образом, $(\widehat{m}_X - m_X)\sqrt{n}/\sigma_X \sim \mathbf{N}(0;1)$.

2) Докажем теперь второе утверждение. С этой целью введем новые случайные величины $Y_1, \dots, Y_n$:

$$Y = \frac{1}{\sigma_X} C Z_n,$$

где $Y \stackrel{\Delta}{=} \operatorname{col}(Y_1, \dots, Y_n)$, $Z_n \stackrel{\Delta}{=} \operatorname{col}(X_1, \dots, X_n)$, C — ортогональная матрица, построенная выше, т. е. удовлетворяющая свойствам 1) C–4) C. Тогда имеем

$$Y_k = \sum_{j=1}^{n} \frac{c_{kj}}{\sigma_X} X_j, Y_1 = \sum_{j=1}^{n} \frac{c_{1j}}{\sigma_X} X_j = \frac{1}{\sigma_X \sqrt{n}} \sum_{j=1}^{n} X_j = \frac{\sqrt{n}}{\sigma_X} \widehat{m}_X.$$

Найдем моменты новых СВ. Учитывая свойства 1) $\widehat{m}_X$ и 4) C, имеем

$$\mathbf{M}[Y_1] = \frac{\sqrt{n}}{\sigma_X} m_X,$$

$$\mathbf{M}[Y_k] = \mathbf{M}\Big[\sum_{j=1}^{n} \frac{c_{kj}}{\sigma_X} X_j\Big] = \frac{m_X}{\sigma_X} \sum_{j=1}^{n} c_{kj} = 0, \quad k = \overline{2,n}.$$

Учитывая свойства 4) $\widehat{m}_X$ и 2) C, получаем

$$\mathbf{D}[Y_k] = \sum_{j=1}^{n} \frac{c_{kj}^2}{\sigma_X^2} \mathbf{D}[X_j] = \sum_{j=1}^{n} c_{kj}^2 = 1, \quad k = \overline{1,n}.$$

Пусть $k \neq i$, $k = \overline{2,n}$, $i = \overline{2,n}$. Тогда с учетом независимости СВ X_j, $j = \overline{1,n}$, и их одинаковой распределенности имеем

$$\sigma_X^2 k_{Y_k Y_i} = \sigma_X^2 \mathbf{M}[Y_k Y_i] = \mathbf{M}\Big[\sum_{j=1}^n c_{kj} X_j \sum_{l=1}^n c_{il} X_l\Big] =$$

$$= \sum_{j \neq l}^n c_{kj} c_{il} \mathbf{M}[X_j]\mathbf{M}[X_l] + \sum_{j=1}^n c_{kj} c_{ij} \mathbf{M}[X_j^2] =$$

$$= m_X^2 \sum_{j=1}^n c_{kj} \sum_{l=1}^n c_{il} - m_X^2 \sum_{j=1}^n c_{kj} c_{ij} + \mathbf{M}[X^2] \sum_{j=1}^n c_{kj} c_{ij} = 0,$$

так как каждая из этих сумм равна нулю в силу свойств $1)C$, $4)C$.

Аналогично устанавливается, что $k_{Y_1 Y_i} = k_{Y_k Y_1} = 0$ для всех $i = \overline{1,n}$, $k = \overline{1,n}$. Таким образом, мы установили, что все СВ Y_k, $k = \overline{1,n}$, некоррелированы, но они имеют нормальное распределение, так как образованы суммой нормальных СВ X_j, $j = \overline{1,n}$. Поэтому, СВ Y_k, $k = \overline{1,n}$, независимы. Причем, $Y_k \sim \mathbf{N}(0;1)$, $k = \overline{2,n}$.

Рассмотрим выборочную дисперсию $\widehat{d}_X$, введенную согласно определению 16.16. Тогда

$$\frac{n\widehat{d}_X}{\sigma_X^2} = \frac{1}{\sigma_X^2} \sum_{j=1}^n (X_j - \widehat{m}_X)^2 = \frac{1}{\sigma_X^2} \sum_{j=1}^n (X_j^2 - 2X_j \widehat{m}_X + \widehat{m}_X^2) =$$

$$= \frac{1}{\sigma_X^2} \Big(\sum_{j=1}^n X_j^2 - 2n\widehat{m}_X^2 + n\widehat{m}_X^2\Big) = \frac{1}{\sigma_X^2} \sum_{j=1}^n X_j^2 - \frac{n}{\sigma_X^2} \widehat{m}_X^2 =$$

$$= \sum_{k=1}^n Y_k^2 - Y_1^2 = \sum_{k=2}^n Y_k^2,$$

где $Y_k \sim \mathbf{N}(0;1)$, $k = \overline{2,n}$. Таким образом, мы доказали, что, согласно определению 17.1, $\overset{*}{D}_X \overset{\Delta}{=} n\widehat{d}_X/\sigma_X^2 \sim \chi^2(n-1)$. Учитывая теперь, что

$$\widehat{d}_X = \frac{\sigma_X^2}{n} \sum_{k=2}^n Y_k^2, \qquad \widehat{m}_X = \frac{\sigma_X}{\sqrt{n}} Y_1,$$

а также независимость СВ $Y_1, Y_2, \ldots, Y_n$, приходим к выводу, что независимы также и СВ $\widehat{d}_X$, $\widehat{m}_X$, т. е. верно последнее утверждение теоремы.

3) Наконец, докажем третье утверждение. Рассмотрим СВ

$$\overline{Y}_1 \overset{\Delta}{=} Y_1 - \frac{\sqrt{n}}{\sigma_X} m_X = \frac{\sqrt{n}}{\sigma_X} (\widehat{m}_X - m_X),$$

которая имеет распределение $\mathbf{N}(0;1)$, так как установлено выше, что $\mathbf{M}[Y_1] = \sqrt{n} m_X/\sigma_X$, $\mathbf{D}[Y_1] = 1$, $\mathbf{M}[\widehat{m}_X] = m_X$ и СВ Y_1 имеет нормальное распределение. Очевидно также, что СВ $\overline{Y}_1$ и $\widehat{d}_X$ независимы. Рассмотрим СВ T, имеющую распределение Стьюдента $\mathbf{S}(n-1)$. Согласно определению 17.2, используя полученные представления для $\widehat{m}_X$ и $\widehat{d}_X$, находим

$$T \stackrel{\Delta}{=} \overline{Y}_1 \sqrt{\frac{n-1}{Y_2^2 + \ldots + Y_n^2}} = (\widehat{m}_X - m_X)\,\frac{\sqrt{n}}{\sigma_X} \sqrt{\frac{(n-1)\sigma_X^2}{n\widehat{d}_X}} \stackrel{\Delta}{=} \overset{\circ}{M}_X .$$

Таким образом, $\overset{\circ}{M}_X \sim \mathbf{S}(n-1)$.

Пример 19.1. По выборке Z_n из нормального распределения требуется построить доверительный интервал для неизвестного МО m_X при известной дисперсии σ_X^2. Из приведенного выше утверждения следует, что СВ $\overset{*}{M}_X$ имеет нормальное распределение $\mathbf{N}(0;1)$, которое не зависит от m_X, и, кроме того, функция $G(Z_n, m_X) \stackrel{\Delta}{=} \overset{*}{M}_X = (\widehat{m}_X - m_X)\sqrt{n}/\,\sigma_X$ является непрерывной и убывающей по m_X. Это значит, что СВ $\overset{*}{M}_X$ является центральной статистикой. Поэтому доверительный интервал для неизвестного m_X можно найти, если разрешить относительно m_X двойное неравенство

$$g_1 \leqslant \frac{(\widehat{m}_X - m_X)\sqrt{n}}{\sigma_X} \leqslant g_2,$$

где величины g_1 и g_2 подобраны таким образом, что это неравенство выполняется с вероятностью $1-\alpha$. Заметим, что данное условие неоднозначно определяет g_1, g_2. Выберем доверительный интервал минимальной длины. Учитывая симметрию относительно оси OY плотности стандартного нормального распределения, можно показать, что такой интервал будет иметь минимальную длину, если положить $g_1 = -g_2$, и при этом он оказывается центральным. Таким образом, получаем следующий доверительный интервал:

$$\widehat{m}_X - \frac{\sigma_X}{\sqrt{n}}\, u_\gamma \leqslant m_X \leqslant \widehat{m}_X + \frac{\sigma_X}{\sqrt{n}}\, u_\gamma,$$

где u_γ — квантиль уровня $\gamma \stackrel{\Delta}{=} 1 - \alpha/2$ стандартного нормального распределения $\mathbf{N}(0;1)$. В данном случае длина доверительного интервала равна $\Delta = 2u_\gamma \sigma_X/\sqrt{n}$ и не случайна. Поэтому, задавшись значениями любых двух из трех величин Δ, α, n, можно определить значение третьей величины.

Пример 19.2. Используя утверждение 3) теоремы 19.1, можно построить аналогичный центральный доверительный интервал

для неизвестного m_X СВ X, имеющей нормальное распределение $\mathbf{N}(m_X; \sigma_X^2)$, и в случае, когда величина σ_X неизвестна:

$$\widehat{m}_X - \sqrt{\frac{\widehat{d}_X}{n-1}}\, t_\gamma(n-1) \leqslant m_X \leqslant \widehat{m}_X + \sqrt{\frac{\widehat{d}_X}{n-1}}\, t_\gamma(n-1),$$

где $t_\gamma(n-1)$ — квантиль уровня $\gamma \stackrel{\Delta}{=} 1 - \alpha/2$ распределения Стьюдента $\mathbf{S}(n-1)$.

В отличие от предыдущего примера длина доверительного интервала случайна и зависит от СВ $\widehat{d}_X$. Но при $n \geqslant 30$ интервалы из примеров 19.1 и 19.2 практически совпадают, так как при $n \geqslant$ $\geqslant 30$ распределение Стьюдента близко к стандартному нормальному распределению.

Пример 19.3. Центральный доверительный интервал для неизвестного параметра σ_X^2 СВ $X \sim \mathbf{N}(m_X; \sigma_X^2)$ при неизвестном m_X можно получить, используя утверждение 2) теоремы 19.1,

$$\frac{n}{x_\gamma(n-1)}\, \widehat{d}_X \leqslant \sigma_X^2 \leqslant \frac{n}{x_{1-\gamma}(n-1)}\, \widehat{d}_X,$$

где $x_\gamma(n-1)$ и $x_{1-\gamma}(n-1)$ — квантили уровней $\gamma \stackrel{\Delta}{=} 1 - \alpha/2$ и $1 -$ $- \gamma = \alpha/2$ для распределения хи-квадрат с $n-1$ степенью свободы.

Пример 19.4. Построим доверительный интервал для неизвестного параметра b равномерного распределения $\mathbf{R}(0; b)$. Можно показать, что СВ $G(Z_n, b) \stackrel{\Delta}{=} \left(X^{(n)} \big/ b\right)^n$ имеет распределение $\mathbf{R}(0; 1)$ для любого $b > 0$. Кроме того, функция $G(z_n, b)$ — убывающая по b. Следовательно, $G(Z_n, b)$ является центральной статистикой. Тогда получаем условие

$$\mathbf{P}\left\{g_1 \leqslant \left(\frac{X^{(n)}}{b}\right)^n \leqslant g_2\right\} = 1 - \alpha$$

для некоторых чисел g_1, g_2. Разрешая двойное неравенство в этом вероятностном условии относительно b, получаем следующий доверительный интервал:

$$\frac{X^{(n)}}{g_2^{1/n}} \leqslant b \leqslant \frac{X^{(n)}}{g_1^{1/n}}.$$

Отметим, что этот интервал будет иметь наименьшую длину, если $g_2 = 1$, а g_1 является квантилью уровня α распределения $\mathbf{R}(0; 1)$, т. е. $g_1 = \alpha$.

19.3. Использование точечной оценки. Пусть $T \stackrel{\Delta}{=} \widehat{\theta}(Z_n)$ — точечная оценка неизвестного параметра θ с функцией распределения $F_T(t,\theta)$, которая монотонна по θ. Пусть $t_n \stackrel{\Delta}{=} \widehat{\theta}(z_n)$ — реализация оценки $T \stackrel{\Delta}{=} \widehat{\theta}(Z_n)$. Пусть $\theta_1(z_n)$ и $\theta_2(z_n)$ — такие два числа, что

$$\theta_1(z_n) \stackrel{\Delta}{=} \max\left\{\theta : F_T(t_n,\theta) = \frac{\alpha}{2}\right\},$$
$$\theta_2(z_n) \stackrel{\Delta}{=} \min\left\{\theta : F_T(t_n-0,\theta) = 1-\frac{\alpha}{2}\right\},$$

если $F_T(t,\theta)$ — возрастающая по θ, и

$$\theta_1(z_n) \stackrel{\Delta}{=} \max\left\{\theta : F_T(t_n-0,\theta) = 1-\frac{\alpha}{2}\right\},$$
$$\theta_2(z_n) \stackrel{\Delta}{=} \min\left\{\theta : F_T(t_n,\theta) = \frac{\alpha}{2}\right\},$$

если $F_T(t,\theta)$ — убывающая по θ.

Интервал $[\theta_1(Z_n), \theta_2(Z_n)]$ со случайными концами, определенными выше, является центральным доверительным интервалом, «накрывающим» с вероятностью $1-\alpha$ неизвестный параметр θ.

Во многих случаях такой доверительный интервал совпадает с интервалом, построенным на основе центральной статистики. Например, этот факт имеет место в случае нормального распределения. Действительно, в частности, при неизвестном m_X и известном σ_X в качестве точечной оценки m_X можно принять $T \stackrel{\Delta}{=} \widehat{m}_X$, которая имеет распределение $\mathbf{N}(m_X; \sigma_X^2/n)$. Очевидно, что распределение $F_T(t, m_X)$ является в данном случае убывающей по m_X функцией. Поэтому, если строить центральный доверительный интервал с помощью распределения точечной оценки, то он совпадет с доверительным интервалом минимальной длины, найденным в примере 19.1 предыдущего п. 19.2. Но в ряде случаев вообще не удается построить доверительный интервал на основе центральной статистики, тогда как доверительный интервал, основанный на распределении точечной оценки, существует. Рассмотрим несколько таких примеров. Отметим также, что в примере 19.4 из предыдущего п. 19.2 наблюдается противоположная ситуация.

Пример 19.5. Пусть СВ X имеет распределение Бернулли $\mathbf{Bi}(1;p)$ с неизвестным параметром p и $Z_n = \mathrm{col}(X_1, \ldots, X_n)$ — соответствующая выборка. МП-оценкой для p является выборочное среднее

$$\widehat{\theta}(Z_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

СВ $T \stackrel{\Delta}{=} \widehat{\theta}(Z_n)$ может принимать значения $0, 1/n, 2/n, \ldots, n/n$, а ее функция распределения имеет вид

$$F_T\left(\frac{k}{n}, p\right) = \sum_{i=0}^{k} C_n^i p^i (1-p)^{n-i}.$$

Дифференцируя по p функцию $F_T(k/n, p)$, можно убедиться, что ее производная отрицательна, т. е. $F_T(k/n, p)$ является монотонно убывающей по p при $k < n$. Поэтому можно применить описанную выше методику для построения доверительного интервала для p. Таким образом, границы центрального доверительного интервала $[\theta_1(z_n), \theta_2(z_n)]$ находятся из решения следующих уравнений:

$$1 - F_T\left(\frac{k-1}{n}, \theta_1\right) = \sum_{i=k}^{n} C_n^i \theta_1^i (1-\theta_1)^{n-i} = \frac{\alpha}{2},$$

$$F_T\left(\frac{k}{n}, \theta_2\right) = \sum_{i=0}^{k} C_n^i \theta_2^i (1-\theta_2)^{n-i} = \frac{\alpha}{2}.$$

Отметим, что при построении одностороннего доверительного интервала для p можно положить, например, $\theta_2 = 1$, и тогда получаем следующее уравнение для определения θ_1:

$$1 - F_T\left(\frac{k-1}{n}, \theta_1\right) = \sum_{i=k}^{n} C_n^i \theta_1^i (1-\theta_1)^{n-i} = \alpha.$$

Пример 19.6. Пусть СВ X имеет распределение Пуассона $\mathbf{\Pi}(a)$ с неизвестным параметром a. МП-оценкой параметра a является выборочное среднее

$$\widehat{\theta}(Z_n) = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Оценка $T \stackrel{\Delta}{=} \widehat{\theta}(Z_n)$ принимает значения k/n при $k = 0, 1, 2, \ldots, n$, а ее функция распределения имеет вид

$$F_T\left(\frac{k}{n}, a\right) = \sum_{i=0}^{k} e^{-na} \frac{(na)^i}{i!}.$$

Производная данной функции оказывается отрицательной, следовательно, функция $F_T(k/n, a)$ является монотонно убывающей по a. Поэтому воспользуемся описанным выше способом построения центрального доверительного интервала $[\theta_1(z_n), \theta_2(z_n)]$. В данном случае

для определения θ_1 и θ_2 получаем следующие уравнения:

$$1 - F_T\left(\frac{k-1}{n}, \theta_1\right) = \sum_{i=k}^{\infty} e^{-n\theta_1} \frac{(n\theta_1)^i}{i!} = \frac{\alpha}{2},$$

$$F_T\left(\frac{k}{n}, \theta_2\right) = \sum_{i=0}^{k} e^{-n\theta_2} \frac{(n\theta_2)^i}{i!} = \frac{\alpha}{2}.$$

Правая граница одностороннего доверительного интервала находится из уравнения

$$F_T\left(\frac{k}{n}, \theta_2\right) = \sum_{i=0}^{k} e^{-n\theta_2} \frac{(n\theta_2)^i}{i!} = \alpha.$$

19.4. Типовые задачи.

З а д а ч а 19.1. Используя данные типовой задачи 16.1, построить центральный доверительный интервал уровня доверия 0,95 для неизвестного математического ожидания средней температуры января в г. Саратове (СВ X). Будем предполагать, что СВ X имеет гауссовское распределение.

Р е ш е н и е. Согласно примеру 19.2 центральный доверительный интервал для неизвестного математического ожидания m_X уровня доверия 0,95 имеет вид

$$\widehat{m}_X - \sqrt{\frac{\widehat{d}_X}{n-1}}\, t_{0,975}(n-1) \leqslant m_X \leqslant \widehat{m}_X + \sqrt{\frac{\widehat{d}_X}{n-1}}\, t_{0,975}(n-1),$$

где объем выборки $n = 13$, $\widehat{m}_X = -11{,}87$, $\widehat{d}_X = 22{,}14$ (результат решения задачи 16.1), а квантиль распределения Стьюдента $t_{0,975}(12) = 2{,}18$ находится по таблице.

Сделав необходимые вычисления, получаем доверительный интервал для m_X : $[-14{,}43, -8{,}9]$.

О т в е т. $[-14{,}43, -8{,}9]$.

З а д а ч а 19.2. Используя данные типовой задачи 16.1, построить центральный доверительный интервал уровня доверия 0,95 для неизвестной дисперсии средней температуры января в г. Саратове (СВ X). Будем предполагать, что СВ X имеет гауссовское распределение.

Р е ш е н и е. Согласно примеру 19.3 центральный доверительный интервал для неизвестной дисперсии σ_X^2 уровня доверия 0,95 имеет вид

$$\frac{n}{x_{0,975}(n-1)}\, \widehat{d}_X \leqslant \sigma_X^2 \leqslant \frac{n}{x_{0,025}(n-1)}\, \widehat{d}_X,$$

где объем выборки $n = 13$, $\widehat{d}_X = 22{,}14$ (результат решения задачи 16.1), а $x_{0,975}(12) = 23{,}3$, $x_{0,025}(12) = 4{,}4$ — квантили распределения $\chi^2(12)$, которые находятся по таблице.

Таким образом, получаем доверительный интервал для σ_X^2 : [11,4, 60,38].

Ответ. [11,4, 60,38].

§ 20. Проверка статистических гипотез

20.1. Основные понятия.

Определение 20.1. *Статистической гипотезой* H или просто *гипотезой* называется любое предположение относительно параметров или закона распределения СВ X, проверяемое по выборке Z_n.

Определение 20.2. Проверяемая гипотеза называется *основной* (или *нулевой*) и обозначается H_0. Гипотеза, конкурирующая с H_0, называется *альтернативной* и обозначается H_1.

Определение 20.3. Статистическая гипотеза H_0 называется *простой*, если она однозначно определяет параметр или распределение СВ X. В противном случае гипотеза H_0 называется *сложной*.

Пример 20.1. По выборке Z_n требуется проверить гипотезу H_0 о том, что $m_X = m_0$, где m_0 — некоторое фиксированное число, против гипотезы H_1 о том, что $m_X \neq m_0$. Или проверить гипотезу H_0 против гипотезы H_2 о том, что $m_X > m_0$.

Одна из основных задач математической статистики состоит в проверке соответствия результатов эксперимента предполагаемой гипотезе H_0. С этой целью выбирается некоторая статистика $Z = \varphi(Z_n)$, для которой предполагается известным условное распределение $F(z|H_0)$ относительно проверяемой гипотезы H_0. С помощью этой статистики строится процедура (правило) проверки гипотезы.

Определение 20.4. *Статистическим критерием* (*критерием согласия, критерием значимости* или *решающим правилом*) проверки гипотезы H_0 называется правило, в соответствии с которым по реализации $z \triangleq \varphi(z_n)$ статистики Z гипотеза H_0 принимается или отвергается.

Определение 20.5. *Критической областью* $\overline{G}$ статистического критерия называется область реализаций z статистики Z, при которых гипотеза H_0 отвергается.

Определение 20.6. *Доверительной областью* G статистического критерия называется область значений z статистики Z, при которых гипотеза H_0 принимается.

Например, в качестве статистического критерия можно использовать правило:

1) если значение $z = \varphi(z_n)$ статистики $Z = \varphi(Z_n)$ лежит в критической области $\overline{G}$, то гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1;

2) если реализация $z = \varphi(z_n)$ статистики $Z = \varphi(Z_n)$ лежит в доверительной области G, то гипотеза H_0 принимается.

При реализации этого правила возникают ошибки двух видов.

Определение 20.7. *Ошибкой 1-го рода* называется событие, состоящее в том, что гипотеза H_0 отвергается, когда она верна.

Определение 20.8. *Ошибкой 2-го рода* называется событие, состоящее в том, что принимается гипотеза H_0, когда верна гипотеза H_1.

Определение 20.9. *Уровнем значимости* статистического критерия называется вероятность ошибки 1-го рода $\alpha \stackrel{\Delta}{=} \mathbf{P}\{Z \in \overline{G} | H_0\}$. Вероятность ошибки 1-го рода α может быть вычислена, если известно распределение $F(z|H_0)$.

Вероятность ошибки 2-го рода равна $\beta \stackrel{\Delta}{=} \mathbf{P}\{Z \in G | H_1\}$ и может быть вычислена, если известно условное распределение $F(z|H_1)$ статистики Z при справедливости гипотезы H_1.

Ясно, что с уменьшением вероятности α ошибки 1-го рода возрастает вероятность β ошибки 2-го рода, и наоборот, т. е. при выборе критической и доверительной областей должен достигаться определенный компромисс. Поэтому часто при фиксированной вероятности ошибки 1-го рода критическая область выбирается таким образом, чтобы вероятность ошибки 2-го рода была минимальна.

Проверка статистической гипотезы может быть подразделена на следующие этапы:

1) сформулировать проверяемую гипотезу H_0 и альтернативную к ней гипотезу H_1;

2) выбрать уровень значимости α;

3) выбрать статистику Z для проверки гипотезы H_0;

4) найти распределение $F(z|H_0)$ статистики Z при условии, что гипотеза H_0 верна;

5) построить, в зависимости от формулировки гипотезы H_1, критическую область $\overline{G}$;

6) получить выборку наблюдений $x_1, \ldots, x_n$ и вычислить выборочное значение $z = \varphi(x_1, \ldots, x_n)$ статистики Z критерия;

7) принять статистическое решение на уровне доверия $1 - \alpha$: если $z \in \overline{G}$, то отклонить гипотезу H_0 как не согласующуюся с результатами наблюдений, а если $z \in G$, то принять гипотезу H_0 как не противоречащую результатам наблюдений.

20.2. Проверка гипотезы о значении параметра. Пусть имеется параметрическая статистическая модель $(S_\theta, F_{Z_n}(z_n,\theta))$, $\theta \in \Theta \subset \subset \mathbb{R}^1$, т. е. выборка Z_n соответствует распределению $F(x,\theta)$ с неизвестным параметром θ. Проверим простую гипотезу H_0, состоящую в том, что $\theta = \theta_0$, где θ_0 — некоторое фиксированное число из множества Θ.

Формулировка альтернативной гипотезы H_1 и уровень значимости α определяют размер и положение критической области $\overline{G}$ на множестве значений статистики Z. Например, если альтернативная гипотеза H_1 формулируется как $\theta > \theta_0$ (или $\theta < \theta_0$), то критическая область размещается на правом (или левом) «хвосте» распределения статистики Z, т. е.

$$\overline{G} = \{z > z_{1-\alpha}\} \quad (\text{или } \overline{G} = \{z < z_\alpha\}),$$

где $z_{1-\alpha}$ и z_α — квантили уровней $1-\alpha$ и α, соответственно, распределения $F(z|H_0)$ статистики Z при условии, что верна гипотеза H_0. В этом случае статистический критерий называется *односторонним*. Если альтернативная гипотеза H_1 формулируется как $\theta \neq \theta_0$, то критическая область $\overline{G}$ размещается на обоих «хвостах» распределения статистики Z, т. е. определяется совокупностью неравенств

$$\overline{G} = \left\{z < z_{\alpha/2}\right\} \bigcup \left\{z > z_{1-\alpha/2}\right\},$$

где $z_{\alpha/2}$ и $z_{1-\alpha/2}$ — квантили уровней $\alpha/2$ и $1-\alpha/2$, соответственно, распределения $F(z|H_0)$. В этом случае критерий называется *двусторонним*.

Таблица 20.1

Предположение	Статистика Z критерия	Распределение $F(z\|H_0)$	Доверительная область G гипотезы H_0
σ_X^2 известна	$\dfrac{(\widehat{m}_X - m_0)\sqrt{n}}{\sigma_X}$	$\mathbf{N}(0;1)$	$[-u_\gamma, u_\gamma]$
σ_X^2 неизвестна	$\dfrac{(\widehat{m}_X - m_0)\sqrt{n-1}}{\sqrt{\widehat{d}_X}}$	$\mathbf{S}(n-1)$	$[-t_\gamma(n-1), t_\gamma(n-1)]$

Пример 20.2. Пусть известно, что СВ X имеет нормальное распределение. Требуется, используя реализацию выборки z_n, проверить гипотезу H_0, состоящую в том, что $m_X = m_0$ (m_0 — некоторое фиксированное число), против альтернативной гипотезы H_1 о том, что $m_X \neq m_0$. Возможны два случая: дисперсия σ_X^2 известна или неизвестна. Статистики Z для обоих случаев можно выбрать, используя утверждение из п. 19.2. Представим эти случаи в виде табл. 20.1.

Для каждого случая в соответствии с утверждениями 1) и 3) теоремы 19.1 получаем свою доверительную область, где u_γ, $t_\gamma(n-1)$ — квантили уровня $\gamma \triangleq 1-\alpha/2$ распределений $\mathbf{N}(0;1)$ и $\mathbf{S}(n-1)$ соответственно.

Пример 20.3. Пусть СВ X нормально распределена, а ее дисперсия неизвестна. Требуется на основе реализации z_n выборки Z_n, порожденной СВ X, проверить гипотезу H_0 о том, что $\sigma_X^2 = \sigma_0^2$ (σ_0 — некоторое фиксированное число), против альтернативной гипотезы H_1, состоящей в том, что $\sigma_X^2 \neq \sigma_0^2$. Возможны два случая: m_X — известно или m_X — неизвестно. Представим эти случаи в виде следующей таблицы:

Таблица 20.2

Предположение	Статистика Z критерия	Распределение $F(z\|H_0)$	Доверительная область G гипотезы H_0
m_X известно	$\dfrac{\sum\limits_{k=1}^{n}(X_k - m_X)^2}{\sigma_0^2}$	$\chi^2(n)$	$[x_{1-\gamma}(n), x_\gamma(n)]$
m_X неизвестно	$\dfrac{n\widehat{d}_X}{\sigma_0^2}$	$\chi^2(n-1)$	$[x_{1-\gamma}(n-1), x_\gamma(n-1)]$

Здесь $x_\gamma(k)$, $x_{1-\gamma}(k)$ — квантили уровня $\gamma \triangleq 1-\alpha/2$ и $1-\gamma$ распределения $\chi^2(k)$ с k степенями свободы, где $k = n,\ n-1$.

Замечание 20.1. На практике обычно задают $\alpha \in [0{,}01,\ 0{,}05]$.

20.3. Проверка гипотезы о виде закона распределения. Пусть имеется реализация z_n выборки Z_n, порожденной СВ X с неизвестной функцией распределения $F(x)$. Требуется проверить гипотезу H_0, состоящую в том, что СВ X имеет определенный закон распределения $\overline{F}(x,\theta)$ (например, нормальный, равномерный и т. д.). Истинный закон распределения $F(x)$ неизвестен. Для проверки такой гипотезы можно использовать *статистический критерий хи-квадрат* (*критерий Пирсона*). Правило проверки заключается в следующем.

1) Формулируется гипотеза H_0, состоящая в том, что СВ X имеет распределение определенного вида $\overline{F}(x,\theta_1,\ldots,\theta_s)$ с s неизвестными параметрами $\theta_1,\ldots,\theta_s$ (например, m и σ^2 для нормального распределения, a и b — для равномерного и т. д.).

2) По реализации z_n выборки Z_n методом максимального правдоподобия находятся оценки $\widehat{\theta}_1, \ldots, \widehat{\theta}_s$ неизвестных параметров $\theta_1, \ldots, \theta_s$.

3) Действительная ось $\mathbb{R}^1$ разбивается на $l+1$ непересекающийся полуинтервал (разряд) $\Delta_0, \ldots, \Delta_l$ так, как это сделано в при построении гистограммы в п. 16.4. Подсчитывается число n_k элементов выборки, попавших в каждый k–й разряд Δ_k, $k = \overline{1, l-1}$, за исключением Δ_0 и Δ_l. Полагается $n_0 = n_l = 0$.

4) Вычисляются гипотетические вероятности p_k попадания СВ X в полуинтервалы Δ_k, $k = \overline{0, l}$. Если у распределения $\overline{F}(x, \theta_1, \ldots, \theta_s)$ имеется плотность $\overline{f}(x, \theta_1, \ldots, \theta_s)$, то вероятности p_k могут быть вычислены следующим образом:

$$p_k = \int\limits_{\alpha_k}^{\alpha_{k+1}} \overline{f}\left(x, \widehat{\theta}_1, \ldots, \widehat{\theta}_s\right) dx,$$

где $\alpha_0 = -\infty$, $\alpha_{l+1} = +\infty$, или приближенно по формуле

$$p_k \cong \overline{f}\left(\overline{x}_k, \widehat{\theta}_1, \ldots, \widehat{\theta}_s\right)(\alpha_{k+1} - \alpha_k), \quad k = \overline{1, l-1},$$

где $\overline{x}_k \triangleq (\alpha_{k+1} + \alpha_k)/2$ — середина разряда Δ_k.

5) Вычисляется реализация статистики критерия хи-квадрат по формуле

$$z \triangleq \varphi(z_n) \triangleq np_0 + \sum_{k=1}^{l-1} \frac{(n_k - np_k)^2}{np_k} + np_l.$$

6) Известно, что при соблюдении некоторых естественных условий регулярности и достаточно большом объеме n выборки Z_n распределение $F(z|H_0)$ статистики $Z = \varphi(Z_n)$ хорошо аппроксимируется распределением $\chi^2(l-s)$ с $l-s$ степенями свободы, где s — количество неизвестных параметров предполагаемого закона распределения $\overline{F}(x, \theta_1, \ldots, \theta_s)$, а $l+1$ — количество разрядов, вероятность попадания в которые ненулевая. Тогда критическая область $\overline{G}$ принимает вид: $\overline{G} = (x_{1-\alpha}(l-s), +\infty)$, где $x_{1-\alpha}(l-s)$ — квантиль уровня $1-\alpha$ распределения $\chi^2(l-s)$, α — заданный уровень значимости (обычно $\alpha = 0{,}05$).

7) В соответствии с критерием хи-квадрат гипотеза H_0 принимается (т. е. реализация выборки z_n согласуется с гипотезой H_0) на уровне надежности $1-\alpha$, если $\varphi(z_n) \in G = [0, x_{1-\alpha}(l-s)]$. Если же $\varphi(z_n) \in \overline{G}$, то гипотеза H_0 отвергается.

Замечание 20.2. Если при разбиении на полуинтервалы Δ_k оказалось, что $np_k < 5$ для $k = \overline{1, l-1}$, то рекомендуется объединить соседние полуинтервалы.

Если при обработке наблюдений имеется только реализация статистического ряда, то вычисляя выборочные моменты, считают все выборочные значения, попавшие в k-й интервал, равными середине этого интервала. Это вносит известную ошибку, особенно заметную при малом числе интервалов. Для уменьшения ошибок, вносимых группировкой, применяют *поправки Шеппарда*. Если все интервалы Δ_k имеют длину, равную h, то с учетом поправок Шеппарда первые четыре выборочных момента $\widehat{\nu}'_i$, $i = \overline{1,4}$, соответственно равны:

$$\widehat{\nu}'_1 = \widehat{\nu}_1, \qquad \widehat{\nu}'_2 = \widehat{\nu}_2 - \frac{1}{12}\,h^2,$$
$$\widehat{\nu}'_3 = \widehat{\nu}_3 - \frac{1}{4}\,\widehat{\nu}_1 h^2, \qquad \widehat{\nu}'_4 = \widehat{\nu}_4 - \frac{1}{2}\,\widehat{\nu}_2 h^2 + \frac{7}{240}\,h^4.$$

20.4. Проверка гипотезы о независимости двух СВ. Пусть имеется случайный вектор $V \stackrel{\Delta}{=} \mathrm{col}(X, Y)$ с функцией распределения $F_V(x, y)$. Компонентами V являются СВ X и Y с функциями распределения $F_X(x)$ и $F_Y(y)$ соответственно. Пусть имеется выборка $Z_n \stackrel{\Delta}{=} \mathrm{col}(V_1, \ldots, V_n)$, где $V_k = \mathrm{col}(X_k, Y_k)$. Здесь выборка $X_1, \ldots, X_n$ соответствует распределению $F_X(x)$ СВ X, а выборка $Y_1, \ldots, Y_n$ соответствует распределению $F_Y(y)$ СВ Y. Требуется проверить гипотезу H_0 о независимости СВ X и Y, т. е. что $F_V(x, y) = F_X(x)F_Y(y)$ для всех $x \in \mathbb{R}^1$, $y \in \mathbb{R}^1$. Для проверки этой гипотезы используем критерий хи-квадрат, процедура применения которого состоит в следующих действиях.

1) Множество значений СВ X разбивается на s непересекающихся интервалов (разрядов) $\Delta_{x,1}, \ldots, \Delta_{x,s}$, а множество значений СВ Y — на r непересекающихся интервалов $\Delta_{y,1}, \ldots, \Delta_{y,r}$.

2) Для каждого $i = \overline{1,s}$ и каждого $j = \overline{1,r}$ вычисляется число n_{ij} элементов выборки z_n, принадлежащих прямоугольнику $\Delta_{x,i} \times \Delta_{y,j}$.

3) Вычисляется суммарное число $n_{x,i}$ элементов выборки z_n, первая компонента которой попала в i-й разряд $\Delta_{x,i}$ для СВ X, и аналогично — число $n_{y,j}$ элементов той же выборки z_n, вторая компонента которой попала в j-й разряд $\Delta_{y,j}$ для СВ Y:

$$n_{x,i} = \sum_{j=1}^{r} n_{ij}, \quad n_{y,j} = \sum_{i=1}^{s} n_{ij}, \quad n = \sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij}.$$

4) Вычисляется значение статистики критерия хи-квадрат по

формуле

$$z \triangleq \varphi(z_n) = n\left(\sum_{i=1}^{s}\sum_{j=1}^{r}\frac{n_{ij}^2}{n_{x,i}n_{y,j}} - 1\right).$$

5) При справедливости гипотезы H_0 и достаточно большом n распределение статистики $Z \triangleq \varphi(Z_n)$ хорошо аппроксимируется распределением хи-квадрат с $m = (s-1)(r-1)$ степенями свободы. Поэтому критическая область имеет вид

$$\overline{G} = (x_{1-\alpha}(m), +\infty),$$

где $x_{1-\alpha}(m)$ — квантиль уровня $1-\alpha$ распределения $\chi^2(m)$.

6) Принимается статистическое решение на уровне доверия $1-\alpha$: отклонить гипотезу H_0, если $\varphi(z_n) \in \overline{G}$, и принять гипотезу H_0 — в противном случае.

20.5. Проверка гипотезы об однородности наблюдений. Рассмотрим вспомогательную задачу. Пусть имеется опыт с r возможными исходами $A_1, \ldots, A_r$, имеющими вероятности $p_j \triangleq \mathbf{P}(A_j)$, $j = \overline{1,r}$. Осуществляется s-кратное повторение серий из n_i, $i = \overline{1,s}$, независимых повторений опыта. В данном случае выборку z_n образуют наблюдавшиеся исходы в этих сериях, где $n = n_1 + \ldots + n_s$ — общее число опытов. Требуется проверить гипотезу H_0 о том, что во всех этих сериях наблюдалась одна и та же совокупность вероятностей p_j, $j = \overline{1,r}$. Для решения этой задачи вновь применим критерий хи-квадрат. Последовательность проверки этого критерия состоит в следующем.

1) В каждой i-й серии, $i = \overline{1,s}$, подсчитываются числа n_{ji} появлений событий A_j, $j = \overline{1,r}$.

2) Подсчитывается суммарное число N_j появлений события A_j, $j = \overline{1,r}$, во всех сериях, а также числа n_i, $i = \overline{1,s}$, и n:

$$N_j = \sum_{i=1}^{s} n_{ji}, \quad n_i = \sum_{j=1}^{r} n_{ji}, \quad n = \sum_{i=1}^{s}\sum_{j=1}^{r} n_{ji}.$$

3) Вычисляется значение z статистики критерия хи-квадрат по формуле:

$$z \triangleq \varphi(z_n) = n\left(\sum_{i=1}^{s}\sum_{j=1}^{r}\frac{n_{ji}^2}{n_i N_j} - 1\right).$$

4) Известно, что при больших n распределение статистики $Z \stackrel{\Delta}{=} \varphi(Z_n)$ хорошо аппроксимируется распределением хи-квадрат $\chi^2(m)$ с $m = (s-1)(r-1)$ степенями свободы. Формируется критическая область $\overline{G} = (x_{1-\alpha}(m), +\infty)$, где $x_{1-\alpha}(m)$ — квантиль уровня $1-\alpha$ распределения $\chi^2(m)$.

5) Принимается статистическое решение: отклонить гипотезу H_0, если $\varphi(z_n) \in \overline{G}$ и принять гипотезу H_0, если $\varphi(z_n) \in G$.

Пример 20.4. В частном случае, при $r = 2$, т. е. когда опыт имеет два исхода $A_1 = A$ и $A_2 = \overline{A}$, для которых $p = \mathbf{P}(A)$, $q = \mathbf{P}(\overline{A})$, описанная процедура проверяет гипотезу о том, что во всех сериях наблюдений вероятность p остается неизменной.

Пример 20.5. Описанная процедура может быть применена и во многих других случаях. Например, пусть имеется выборка $X_1, \ldots, X_n$, соответствующая функции распределения $F_X(x)$ СВ X и, кроме того, имеется выборка $Y_1, \ldots, Y_m$, соответствующая функции распределения $F_Y(y)$ СВ Y. Требуется проверить гипотезу H_0 об *однородности выборок* $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$, т. е. что $F_Y(t) = F_X(t)$ для всех $t \in \mathbb{R}^1$. Для того чтобы применить приведенную выше процедуру, достаточно предварительно использовать метод группировки данных, разделяя множества возможных значений СВ X и Y на одни и те же разряды, попадания в которые интерпретируется как события A_j, $j = \overline{1,r}$. В данном случае $s = 2$, и статистика критерия принимает следующий вид:

$$z \stackrel{\Delta}{=} \varphi(z_n) = n \cdot m \sum_{j=1}^{r} \frac{1}{n_{j1} + n_{j2}} \left(\frac{n_{j1}}{n} - \frac{n_{j2}}{m} \right)^2 .$$

20.6. Типовые задачи.

Задача 20.1. Используя данные типовой задачи 16.1, проверить на уровне доверия $1 - \alpha = 0{,}95$ гипотезу H_0, состоящую в том, что математическое ожидание m_X средней температуры января в г. Саратове (СВ X) равно $-13{,}75$, т. е. что $m_X = -13{,}75$, против альтернативной гипотезы H_1, состоящей в том, что $m_X \neq -13{,}75$. Будем предполагать, что СВ X имеет гауссовское распределение.

Решение. Дисперсия σ_X^2 СВ X неизвестна, поэтому выберем для проверки гипотезы H_0 статистику

$$Z = \frac{(\widehat{m}_X - m_0)\sqrt{n-1}}{\sqrt{\widehat{d}_X}} .$$

В данной задаче $n = 13$, $m_0 = -13{,}75$ а $\widehat{m}_X = -11{,}87$ и $\widehat{d}_X = 22{,}14$ (результат решения задачи 16.1). Согласно примеру 20.2

распределением $F(z|H_0)$ статистики Z при справедливости H_0 является распределение Стьюдента $\mathbf{S}(n-1)$.

Таким образом, доверительная область G имеет вид $G = [-t_{0,975}(12),\, t_{0,975}(12)] = [-2{,}18,\, 2{,}18]$. Вычисляя значение z статистики Z для данной реализации выборки, имеем

$$z = \frac{(-11{,}87 + 13{,}75)\sqrt{12}}{\sqrt{22{,}14}} \approx 1{,}38.$$

Таким образом, значение z статистики Z попадает в доверительную область G, и, следовательно, на уровне доверия $1 - \alpha = 0{,}95$ можно считать, что результаты наблюдений не противоречат гипотезе H_0, состоящей в том, что $m_X = -13{,}75$.

Ответ. Гипотеза H_0 принимается.

Задача 20.2. Используя данные типовой задачи 16.1, проверить на уровне доверия $1 - \alpha = 0{,}95$ гипотезу H_0, состоящую в том, что дисперсия σ_X^2 средней температуры января в г. Саратове (СВ X) равна 20, т. е. что $\sigma_X^2 = 20$, против альтернативной гипотезы H_1, состоящей в том, что $\sigma_X^2 \neq 20$. Будем предполагать, что СВ X имеет гауссовское распределение.

Решение. Математическое ожидание m_X СВ X неизвестно, поэтому выберем для проверки гипотезы H_0 статистику

$$Z = \frac{n\widehat{d}_X}{\sigma_0^2}\,.$$

В данной задаче $n = 13$, $\sigma_0^2 = 20$, а $\widehat{d}_X = 22{,}14$ (результат решения задачи 16.1). Согласно примеру 20.3 распределением $F(z|H_0)$ статистики Z при справедливости H_0 является распределение $\chi^2(12)$.

Таким образом, доверительная область G имеет вид $G = [x_{0,025}(12),\, x_{0,975}(12)] = [4{,}4,\, 23{,}3]$. Вычисляя значение z статистики Z для данной реализации выборки, имеем

$$z = \frac{13 \cdot 22{,}14}{20} = 14{,}39.$$

Таким образом, значение z статистики Z попадает в доверительную область G, и, следовательно, на уровне доверия $1 - \alpha = 0{,}95$ можно считать, что результаты наблюдений не противоречат гипотезе H_0, состоящей в том, что $\sigma_X^2 = 20$.

Ответ. Гипотеза H_0 принимается.

Задача 20.3. В течение Второй мировой войны на южную часть Лондона упало 535 снарядов. Территория южного Лондона была разделена на 576 участков площадью 0,25 км^2. В следующей таблице

приведены числа участков n_k, на каждый из которых упало по k снарядов:

Таблица 20.3

k	0	1	2	3	4	5
n_k	299	211	93	35	7	1

Требуется с помощью критерия хи-квадрат проверить на уровне доверия $1-\alpha=0{,}95$ гипотезу H_0, состоящую в том, что СВ X (число снарядов, упавших на один участок) распределена по закону Пуассона.

Решение. Распределение Пуассона $\mathbf{\Pi}(\theta)$ имеет один параметр, МП–оценка этого параметра $\widehat{\theta}=\widehat{m}_X\approx 0{,}93$ (см. пример 18.8).

В данной задаче действительная ось естественным образом разбивается на 8 непересекающихся полуинтервалов Δ_k: $(-\infty,0)$, $[0,1)$, $[1,2)$, $\ldots$, $[5,6)$, $[6,+\infty)$.

Гипотетические вероятности p_k попадания пуассоновской СВ X в k-й полуинтервал вычисляются по следующим формулам:

$$p_k=\frac{e^{-\widehat{\theta}}\widehat{\theta}^{k-1}}{(k-1)!}\qquad \text{для } k=\overline{1,6},$$

$$p_7=1-\sum_{k=1}^{6}p_k,\qquad p_0=\mathbf{P}\{X<0\}=0.$$

Вычисленные значения вероятностей указаны в табл. 20.4.

Таблица 20.4

k	1	2	3	4	5	6	7
Δ_k	$[0,1)$	$[1,2)$	$[2,3)$	$[3,4)$	$[4,5)$	$[5,6)$	$[6,\infty)$
p_k	0,394	0,366	0,171	0,053	0,012	0,002	0,002

Вычисляя значения z статистики критерия хи-квадрат, получим

$$z=\sum_{k=1}^{6}\frac{(n_k-np_k)^2}{np_k}+np_7\approx 2{,}2.$$

При справедливости гипотезы H_0 статистика Z имеет распределение $\chi^2(5)$. Тогда критическая область $\overline{G}$ имеет вид $\overline{G}=(x_{0,95}(5),+\infty)=$ $=(11{,}1,+\infty)$, а доверительная область — $G=[0,11{,}1]$.

Так как вычисленное по выборке значение статистики попадает в доверительную область G, то с вероятностью 0,95 можно утверждать, что опытные данные согласуются с гипотезой H_0.

Ответ. Гипотеза H_0 принимается.

З а д а ч а 20.4. Используя данные задачи 16.2, проверить на уровне доверия $1-\alpha = 0{,}95$ гипотезу H_0, состоящую в том, что рост взрослого мужчины (СВ X) имеет нормальное распределение.

Р е ш е н и е. Нормальное распределение $\mathbf{N}(\theta_1;\theta_2^2)$ имеет два неизвестных параметра: $\theta_1 = m_X$ и $\theta_2^2 = \sigma_X^2$. МП–оценкой параметра θ_1 является выборочное среднее $\widehat{m}_X$, а параметра θ_2^2 — выборочная дисперсия $\widehat{d}_X$. С учетом поправок Шеппарда получим

$$\widehat{m}_X = 165{,}77, \qquad \widehat{d}_X = 34{,}24.$$

Вычислим гипотетические вероятности p_k, $k = \overline{1,15}$, попадания гауссовской СВ X в полуинтервалы Δ_k (которые определены в задаче 16.2) по приближенной формуле

$$p_k = h \cdot \frac{1}{\sqrt{2\pi}\sqrt{\widehat{d}_X}} \exp\left\{-\frac{(\overline{x}_k - \widehat{m}_X)^2}{2\widehat{d}_X}\right\},$$

где $\overline{x}_k$ — середина интервала Δ_k, а h — длина интервала Δ_k, которая в данной задаче равна 3 для всех $k = \overline{1,15}$. Результаты вычислений для $k = \overline{1,15}$ приведены в табл. 20.5.

Т а б л и ц а 20.5

Δ_k	143–146	146–149	149–152	152–155	155–158
p_k	0,0003	0,002	0,007	0,023	0,058
Δ_k	158–161	161–164	164–167	167–170	170–173
p_k	0,115	0,175	0,204	0,184	0,127
Δ_k	173–176	176–179	179–182	182–185	185–188
p_k	0,067	0,027	0,009	0,002	0,0004

Вероятность попадания в интервал $\Delta_0 = (-\infty, 143)$ равна $p_0 = 4 \times \times 10^{-5}$, а в интервал $\Delta_{16} = [188, +\infty)$ — $p_{16} = 4 \cdot 10^{-5}$. Вычисляя реализацию z статистики Z, получим

$$z = np_0 + \sum_{k=1}^{15} \frac{(n_k - np_k)^2}{np_k} + np_{16} \approx 7{,}177.$$

При справедливости гипотезы H_0 статистика Z имеет распределение $\chi^2(14)$. Тогда критическая область $\overline{G}$ имеет вид $\overline{G} = (x_{0,95}(14), +\infty) = (23{,}7, +\infty)$, а доверительная область — $G = [0,\ 23{,}7]$.

Так как вычисленное по выборке значение статистики попадает в доверительную область G, то с вероятностью 0,95 можно утверждать, что опытные данные согласуются с гипотезой H_0.

О т в е т. Гипотеза H_0 принимается.

З а д а ч а 20.5. По переписи населения Швеции 1936 г. из совокупности всех супружеских пар была получена выборка в 25 263 пары, вступивших в брак в течение 1931–1936 гг. В следующей таблице приведено распределение годовых доходов (в тыс. крон) и количество детей у супружеских пар в этой выборке (данные взяты из [19]):

Т а б л и ц а 20.6

число детей \ доходы	0–1	1–2	2–3	> 3	Сумма
0	2 161	3 577	2 184	1 636	9 558
1	2 755	5 081	2 222	1 052	11 110
2	936	1 753	640	306	3 635
3	225	419	96	38	778
$\geqslant 4$	39	98	31	14	182
Сумма	6 116	10 928	5 173	3 016	25 263

Требуется установить с доверительной вероятностью $1-\alpha = 0{,}95$ являются ли зависимыми количество детей в семье (СВ X) и уровень годового дохода этой семьи (СВ Y).

Р е ш е н и е. Для проверки гипотезы H_0 о независимости СВ X и СВ Y воспользуемся критерием хи-квадрат. В данной задаче множество значений СВ X (количество детей в семье) разбивается на 5 разрядов $\Delta_{x,i}$, $i = \overline{1,5}$, естественным образом: 0, 1, 2, 3 и не менее 4-х детей (см. первый столбец табл. 20.6). Множество значений СВ Y (годовой доход семьи) разбито на 4 разряда $\Delta_{y,j}$, $j = \overline{1,4}$, которые указаны в первой строке табл. 20.6). Вычисляя значение z (см. п. 20.4) по формуле

$$z = n\left(\sum_{i=1}^{s}\sum_{j=1}^{r}\frac{n_{ij}^2}{n_{x,i}n_{y,j}} - 1\right),$$

где $n = 25\,263$, $s = 5$, $r = 4$, числа $n_{x,i}$, $i = \overline{1,5}$, приведены в последнем столбце таблицы, а числа $n_{y,j}$, $j = \overline{1,4}$, приведены в последней строке таблицы, получим $z = 568{,}5$.

При справедливости гипотезы H_0 статистика Z имеет распределение хи-квадрат с числом степеней свободы $m = (s-1)(r-1) = 4 \times \times 3 = 12$. Тогда критическая область $\overline{G}$ имеет вид $\overline{G} = (x_{0,95}(12), +\infty) = (21, +\infty)$, а доверительная область — $G = [0, 21]$.

Так как вычисленное по выборке значение статистики попадает в критическую область $\overline{G}$, то с вероятностью 0,95 гипотеза H_0 о независимости СВ X (количество детей в семье) и СВ Y (уровень годового дохода семьи) отвергается.

О т в е т. Гипотеза H_0 о независимости СВ X (количество детей в семье) и СВ Y (уровень годового дохода семьи) отвергается.

Задача 20.6. В табл. 20.7 приведены данные о распределении доходов (в тыс. крон) всех промышленных рабочих и служащих Швеции в 1930 г. для возрастных групп 40–50 лет и 50–60 лет (данные взяты из [19]). Требуется проверить на уровне доверия $1-\alpha=0{,}95$ гипотезу H_0 о том, что доходы рабочих и служащих возрастной группы 40–50 лет (СВ X) и доходы рабочих и служащих возрастной группы 50–60 лет (СВ Y) распределены одинаково.

Таблица 20.7

доходы \ возраст	40–50 лет	50–60 лет
0–1	7 831	7 558
1–2	26 740	20 685
2–3	35 572	24 186
3–4	20 009	12 280
4–6	11 527	6 776
>6	6 919	4 222
Сумма	108 598	75 707

Решение. Для проверки гипотезы H_0 об одинаковой распределенности (однородности) СВ X и СВ Y применим критерий хи-квадрат. В данной задаче количество серий испытаний $s=2$. Тогда, согласно примеру 20.5 (см. п. 20.5), значение статистики критерия имеет вид

$$z=n\cdot m\sum_{j=1}^{r}\frac{1}{n_{j1}+n_{j2}}\left(\frac{n_{j1}}{n}-\frac{n_{j2}}{m}\right)^2,$$

где $n=108\,598$, $m=75\,707$, количество разрядов $r=6$, числа n_{j1}, $j=\overline{1,6}$, приведены во втором столбце таблицы, а числа n_{j2}, $j=\overline{1,6}$, приведены в третьем столбце таблицы.

Проведя необходимые вычисления, получим $z=840{,}62$.

При справедливости гипотезы H_0 статистика Z имеет распределение хи-квадрат с числом степеней свободы $m=(s-1)(r-1)=1\times\times5=5$.

Тогда критическая область $\overline{G}$ имеет вид $\overline{G}=(x_{0,95}(5),+\infty)=(11{,}1,+\infty)$, а доверительная область — $G=[0,\ 11{,}1]$.

Так как вычисленное по выборке значение статистики попадает в критическую область $\overline{G}$, то с вероятностью 0,95 гипотеза H_0 об однородности СВ X (доходы рабочих и служащих возрастной группы 40–50 лет) и СВ Y (доходы рабочих и служащих возрастной группы 50–60 лет) отвергается.

Ответ. Гипотеза H_0 об однородности СВ X (доходы рабочих и служащих возрастной группы 40–50 лет) и СВ Y (доходы рабочих и служащих возрастной группы 50–60 лет) отвергается.

§ 21. Задачи для самостоятельного решения

1. По данным типовой задачи 20.3 построить гистограмму числа снарядов, упавших на участок площадью 0,25 км^2.

2. Предполагая, что число снарядов, упавших на участок площадью 0,25 км^2 (см. задачу 20.3), имеет распределение Пуассона с неизвестным параметром θ, построить МП–оценку этого параметра.

3. По данным типовой задачи 16.2 найти, учитывая поправки Шеппарда, выборочное среднее и выборочную дисперсию роста взрослого мужчины (СВ X).

4. По данным типовой задачи 16.1 построить центральный доверительный интервал уровня доверия $1-\alpha = 0{,}99$ для математического ожидания средней температуры января в г. Алатыре (СВ Y). Предполагается, что СВ Y имеет нормальное распределение.

5. По данным типовой задачи 16.1 построить центральный доверительный интервал уровня доверия $1-\alpha = 0{,}95$ для дисперсии средней температуры января в г. Алатыре (СВ Y). Предполагается, что СВ Y имеет нормальное распределение.

6. По данным типовой задачи 16.1 проверить на уровне доверия $1-\alpha = 0{,}99$ гипотезу H_0, состоящую в том, что математическое ожидание средней температуры января в г. Алатыре (СВ Y) равно $-11{,}87$, т. е. что $m_Y = -11{,}87$, против альтернативной гипотезы H_1 о том, что $m_Y \neq -11{,}87$, предполагая, что СВ Y имеет нормальное распределение.

7. По данным типовой задачи 16.1 проверить на уровне доверия $1-\alpha = 0{,}95$ гипотезу H_0, состоящую в том, что дисперсия средней температуры января в г. Алатыре (СВ Y) равна 20, т. е. $\sigma_Y^2 = 20$, против альтернативной гипотезы H_1: $\sigma_Y^2 \neq 20$, предполагая, что СВ Y имеет нормальное распределение.

8. В табл. 21.1 приведены результаты испытаний 200 ламп на продолжительность работы T [в часах].

Таблица 21.1

T [час]	0–300	300–600	600–900	900–1200
число ламп	53	41	30	22
T [час]	1200–1500	1500–1800	1800–2100	2100–2400
число ламп	16	12	9	7
T [час]	2400–2700	2700–3000	3000–3300	> 3300
число ламп	5	3	2	0

Пусть СВ X — продолжительность работы лампы. Используя критерий хи-квадрат, проверить гипотезу H_0 о том, что СВ X (реализация статистического ряда которой приведена в таблице) имеет экспоненциальный закон распределения с плотностью распределения вероятности $f(x) = \theta e^{-\theta x}$ при $x \geqslant 0$. Уровень доверия принять равным $1 - \alpha = 0{,}95$.

9. Утверждается, что результат действия лекарства зависит от способа его применения. Проверьте это утверждение по данным, представленным в табл. 21.2. Уровень доверия принять равным 0,95.

Таблица 21.2

результат \ способ применения	A	B	C
неблагоприятный	11	17	16
благоприятный	20	23	19

10. В табл. 21.3 приведены данные о распределении доходов (в тыс. крон) заводских мастеров Швеции в 1930 г. для возрастных групп 40–50 лет и 50–60 лет (данные взяты из [19]).

Таблица 21.3

доходы \ возраст	40–50 лет	50–60 лет
0–1	71	54
1–2	430	324
2–3	1 072	894
3–4	1 609	1 202
4–6	1 178	903
> 6	158	112

Требуется проверить на уровне доверия 0,95 гипотезу о том, что доходы заводских мастеров возрастной группы 40–50 лет (СВ X) и заводских мастеров возрастной группы 50–60 лет (СВ Y) распределены одинаково. Уровень доверия принять равным 0,95. Сравните полученный результат с ответом задачи 20.6.

11. В результате проведенного исследования было установлено, что у 782 светлоглазых отцов сыновья тоже имеют светлые глаза, а у 89 светлоглазых отцов сыновья — темноглазые. У 50 темноглазых отцов сыновья также темноглазые, а у 79 темноглазых отцов сыновья — светлоглазые. Имеется ли зависимость между цветом глаз отцов (СВ X) и цветом глаз их сыновей (СВ Y)? Уровень доверия принять равным 0,99.