

ГЛАВА 3

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

3.1. Введение. Основные понятия математической статистики

3.1.1. Выборка, основные задачи математической статистики

Рассмотрим случайный эксперимент, связанный со случайной величиной ξ , принимающей значения в $\mathbb{R}^1$. Проведя n независимых повторений эксперимента, мы получим последовательность из n значений случайной величины ξ , которые обозначим через $X_1, X_2, \dots, X_n$.

Пусть $F(x)$ — функция распределения случайной величины ξ . Совокупность $X_1, X_2, \dots, X_n$ — называется *выборкой объема n из генеральной совокупности с функцией распределения $F(x)$* . Выборку объема n можно рассматривать как n -мерный случайный вектор $X = (X_1, X_2, \dots, X_n)$, где $X_1, X_2, \dots, X_n$ — независимые случайные величины, имеющие одну и ту же функцию распределения $F(x)$. В задачах математической статистики $F(x)$ принято называть *теоретической функцией распределения*.

Основной задачей математической статистики является получение по выборке $X_1, X_2, \dots, X_n$, такой информации, которая позволяет более или менее точно судить о неизвестном распределении $F(x)$.

Пример. Продолжительность горения («время жизни») электрической лампочки является ее качественной характеристикой. Пусть случайная величина ξ — это «время жизни», она имеет показательное распределение с плотностью распределения

$$p(t) = \begin{cases} \lambda \cdot e^{-\lambda t}, & \text{если } t > 0, \\ 0, & \text{иначе} \end{cases}.$$

Пусть у нас имеется некоторая партия лампочек. Предположим, что в результате n наблюдений мы получили следующие значения ξ (выраженные в часах): $X_1, X_2, \dots, X_n$, т.е. получили выборку объема n , извлеченную из всей партии. Теперь, опираясь только на

эти данные, мы должны оценить, скажем, параметр λ . В связи с этим возникают следующие виды задач, которые связаны с этим экспериментом.

1. Как по выборке $X_1, X_2, \dots, X_n$, получить точную оценку параметра λ ? Такая задача называется задачей «**построения точечных оценок**».

2. Указать в какой интервал может попасть значение этого параметра λ с достаточно высокой вероятностью. Такая задача называется задачей «**построения доверительных интервалов**».

3. Можно ли считать всю рассматриваемую партию лампочек бракованной или нет? Поставленный вопрос решается, например, следующим образом.

Рассмотрим два предположения (гипотезы) :

$$H_1 : \frac{1}{n} \cdot \sum_{k=1}^n X_k \geq T \quad \text{и} \quad H_2 : \frac{1}{n} \cdot \sum_{k=1}^n X_k < T,$$

где число T либо зададим, либо определим по выборке. Далее, с помощью некоторой процедуры (критерия) проверяем выполнение H_1 , и H_2 . Если для данных $X_1, X_2, \dots, X_n$, гипотеза H_1 выполняется, то, всю рассматриваемую партию лампочек считаем годной, если же выполняется H_2 , то партию лампочек считаем негодной.

Такая задача называется задачей «**проверки статистических гипотез**».

3.1.2. Эмпирическая функция распределения. Гистограмма выборки

Пусть $X_1, X_2, \dots, X_n$ выборка объема n из генеральной совокупности с функцией распределения $F(x)$. Если расположить выборочные данные в порядке неубывания, то полученный ряд называется *вариационным рядом*: $X_{(1)}, X_{(2)}, \dots, X_{(n)}$.

Пример 1. Если выборка объема 4 следующая: 4, -2, 3, 1, то вариационный ряд выглядит так: -2, 1, 3, 4.

Определение 1. Эмпирической называется функция распределения $F(x)$ дискретной случайной величины ξ , у которой таблица распределения имеет следующий вид:

ξ	$X_{(1)}$	$X_{(2)}$	...	$X_{(n)}$
P	$\frac{1}{n}$	$\frac{1}{n}$		$\frac{1}{n}$

Как показано в 2.2.1 функция распределения дискретной случайной величины

ξ	a_1	a_2	...	a_n
P	p_1	p_2		p_n

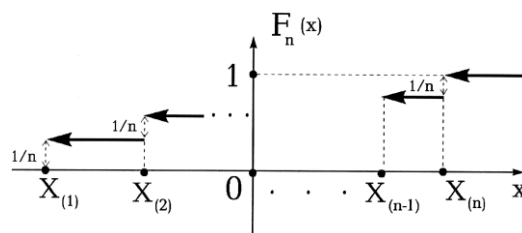
имеет следующий вид:

$$F_{\xi}(x) = \begin{cases} 0, & \text{если } x \leq a_1, \\ \dots \\ p_1 + p_2 + \dots + p_{k-1}, & \text{если } a_{k-1} < x \leq a_k, \quad k = 2, \dots, n, \\ \dots \\ 1, & \text{если } x > a_n. \end{cases}$$

Отсюда

$$F_n(x) = \begin{cases} 0, & \text{если } x \leq X_{(1)}, \\ \dots \\ \frac{k-1}{n}, & \text{если } X_{(k-1)} < x \leq X_{(k)}, \quad k = 2, \dots, n, \\ \dots \\ 1, & \text{если } x > X_{(n)}. \end{cases}$$

Другими словами $F_n(x) = v/n$, где v —число тех выборочных значений X_k , которые меньше x .



Как видно из графика, функция $F_n(x)$ ступенчатая и имеет разрывы в точках $X_{(k)}$ и величина скачка равна $1/n$, если совпадающих друг с другом значений $X_{(k)}$, нет. Если же v значений $X_{(k)}$ совпадают, то величина скачка в этой точке равна v/n .

Представляет интерес предельное поведение $F_n(x)$ при $n \rightarrow \infty$.

Теорема 1. Пусть $X_1, X_2, \dots, X_n$ выборка объема n из генеральной совокупности с функцией распределения $F(x)$. Тогда при $n \rightarrow \infty$ со для любого $x \in \mathfrak{R}^1$ справедливо

$$F_n(x) \xrightarrow{P} F(x), \text{ или, другими словами, для любого } \varepsilon > 0, \lim_{n \rightarrow \infty} \mathbf{P}(|F_n(x) - F(x)| > \varepsilon) = 0.$$

Доказательство. Пусть $\xi_i = \begin{cases} 1, & \text{если } X_i < x, \\ 0, & \text{если } X_i \geq x \end{cases}$ - такие дискретные случайные величины, что $\mathbf{P}(\xi_i = 0) = q$ и $\mathbf{P}(\xi_i = 1) = p, i = 1, 2, \dots, n$.

Легко видеть, что $F(x) = \mathbf{P}(X_i < x) = \mathbf{P}(\xi_i = 1) = p$ и $\mathbf{E}\xi_i = 0 \cdot q + 1 \cdot p = F(x)$. Тогда по закону больших чисел (см. 2.7.2) для эмпирической функции распределения $F_n(x) = \frac{1}{n} \sum_{i=1}^n \xi_i$ при $n \rightarrow \infty$ получим $F_n(x) \xrightarrow{P} F(x)$.

Прежде чем сформулировать еще одну теорему, приведем следующее определение.

Определение 2. Последовательность случайных величин $\xi_1, \xi_2, \dots, \xi_n, \dots$ сходится к ξ с вероятностью 1 (единица) (или почти наверное), если выполняется следующее равенство

$$\mathbf{P}\left(\lim_{n \rightarrow \infty} \xi_n = \xi\right) = 1.$$

Теперь сформулируем (без доказательства, его можно найти в [2]) следующую теорему.

Теорема 2 (Гливленко - Кантелли). В условиях предыдущей теоремы справедливо

$$\mathbf{P}\left(\lim_{n \rightarrow \infty} \max_{-\infty < x < \infty} |F_n(x) - F(x)| = 0\right) = 1.$$

Эти результаты показывают, что при больших n эмпирическая функция распределения дает хорошее приближение для теоретической функции распределения $F(x)$.

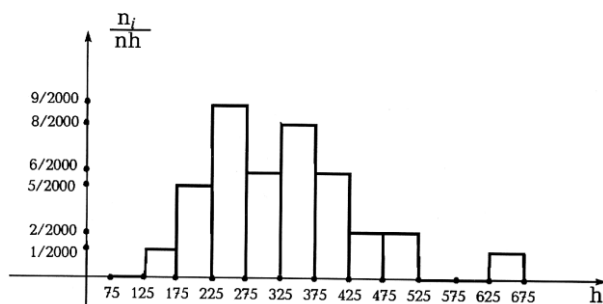
Выборки объема n из генеральной совокупности с непрерывным распределением $F(x)$ на практике часто подвергаются группировке. В этом случае указываются не выборочные значения, а число выборочных значений, попавших в интервалы некоторого определенного разбиения генеральной совокупности (разбиения множества возможных значений случайной величины, имеющей функцию распределения $F(x)$). Как правило, интервалы берутся одинаковой длины, скажем h . Если обозначить через n_i число выборочных значений, попавших в i -интервал, то этот интервал принимается за

основание прямоугольника высоты $n_i/(n \cdot h)$. Получающаяся при этом фигура называется *гистограммой выборки*. Площадь каждого прямоугольника гистограммы равна частоте n_i/n соответствующей группы. При больших n эта площадь будет приблизительно равна вероятности попасть в соответствующий интервал, т.е. будет приблизительно равна интегралу от плотности распределения $p(t)$, вычисленному по данному интервалу. Таким образом, верхняя часть контура гистограммы дает хорошее приближение для плотности распределения.

Пример 2. Испытывалась чувствительность одного из каналов $n = 40$ телевизоров. Данные испытаний указаны в следующей таблице, где в первой строке даны интервалы чувствительности в микровольтах, во второй - число телевизоров, чувствительность которых оказалась данным интервале:

интервал	75-125	125-175	175-225	225-275	275-325	325-375	375-425
n_i	0	1	5	9	6	8	6
интервал	425-475	475-525	525-575	575-625	625-675	675-725	
n_i	2	2	0	0	1	0	

Здесь длина интервала $h = 50$. Построим гистограмму.



3.1.3. Выборочные характеристики случайной величины

Пусть $X_1, X_2, \dots, X_n$ выборка объема n из генеральной совокупности с функцией распределения $F(x)$.

Определение 1. Выборочным начальным моментом порядка k (случайной величины ξ с функцией распределения $F(x)$) называется величина

$$a_k = \frac{1}{n} \cdot \sum_{i=1}^n X_i^k, \quad k \geq 1.$$

Выборочным центральным моментом порядка k называется число

$$m_k = \frac{1}{n} \cdot \sum_{i=1}^n (X_i - a_1)^k, \quad k \geq 2.$$

Замечание 1. Число $a_1 = \frac{1}{n} \cdot \sum_{i=1}^n X_i$ называется *выборочным средним* и обозначается через $\bar{X}$, а число $m_2 = \frac{1}{n} \cdot \sum_{i=1}^n (X_i - a_1)^2$ называется *выборочной дисперсией* и обозначается через s^2 , а s называется *выборочным стандартным отклонением*.

В заключение этого пункта приведем теорему, которая легко получается из закона больших чисел (проверьте самостоятельно!).

Теорема 2. Справедливы следующие утверждения:

- 1) $a_k \rightarrow^p E \xi^k$ при $n \rightarrow \infty$;
- 2) $m_k \rightarrow^p E(\xi - E\xi)^k$ при $n \rightarrow \infty$.

Замечание 2. Из теоремы 2 следует, что при больших n выборочные моменты могут служить приближением для неизвестных теоретических моментов.

Рассмотрим *другие выборочные* характеристики.

Разность $X_{(n)} - X_{(1)}$ называется *размахом вариации* и обозначается через R .

Медиана (Me) – это то значение выборки, которое находится в середине вариационного ряда. Если вариационный ряд дискретный, то в случае, когда n – нечетное число, медиана – это то значение выборки, которое находится ровно в середине вариационного ряда. Если n – четное число, то медиана определяется как среднее арифметическое двух выборочных значений, которые находятся в середине вариационного ряда. Если в выборке разброс крайних значений существенен, то в качестве выборочной характеристики среднего лучше использовать медиану (а не $\bar{X}$)

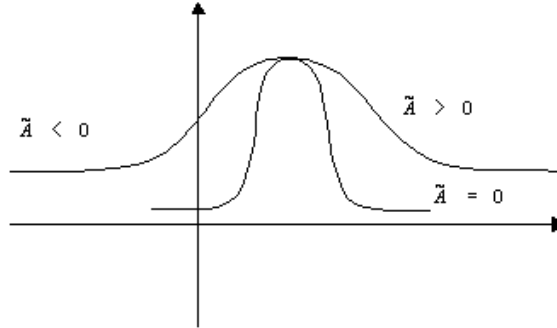
Мода (Mo) – это то значение выборки, которое имеет наибольшую частоту (т.е. это часто встречающееся значение выборки).

Коэффициент вариации (CV). Это безразмерная величина, которая указывает величину вариации (среднеквадратического отклонения) на единицу среднего значения, т.е.

$$CV = \frac{s}{\bar{X}} \cdot 100\%$$

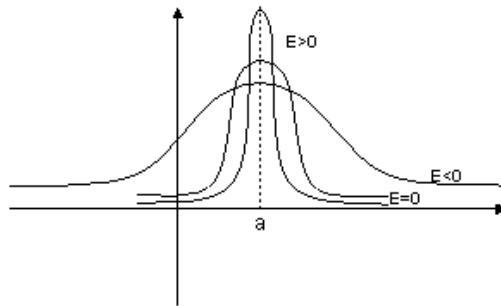
Коэффициент асимметрии ($\tilde{A}$). Эта характеристика определяется так $\tilde{A} = \frac{m_3}{S^3}$.

Коэффициент асимметрии для нормального распределения равен нулю. Графически представим:



Коэффициент эксцесса (E). Эта характеристика определяется так $E = \frac{m_4}{S^4} - 3$.

Коэффициент эксцесса для нормального распределения равен нулю. Графически представим:



Замечание. Введенные нами основные понятия математической статистики естественно распространяются на многомерный случай. Например, предположим, что мы имеем выборку n пар значений двумерной случайной величины ξ : $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. Если приписать каждому из этих значений вероятность $1/n$, то получим распределение выборки. Выборочные характеристики для этого распределения вычисляются по общим правилам для двумерных распределений (см. 2.4.2 и 2.6).

Задачи к 3.1

1. Дано распределение признака X (случайной величины X), полученное по n наблюдениям. В данной задаче X – число сделок на фондовой бирже за квартал; $n=400$ (инвесторов).

X_i	0	1	2	3	4	5	6	7	8	9	10
n_i	146	97	73	34	23	10	6	3	4	2	2

Необходимо¹: 1) построить гистограмму и эмпирическую функцию распределения X ; 2) найти: а) среднюю арифметическую $\bar{X}$; б) медиану M_e и моду M_0 ; в) дисперсию s^2 , среднеквадратическое отклонение s и коэффициент вариации CV ; г) начальные a_k и центральные m_k моменты k -го порядка ($k=1,2,3,4$); д) коэффициенты асимметрии $\tilde{A}$ и эксцесса E . (¹При наличии открытых интервалов значений X типа «менее X_1 » или «свыше X_n » для проведения расчётов их условно заменяют интервалами той же ширины k , т.е. (X_1-k, X_1) или (X_n, X_n+k)).

2. Дано распределение признака X (случайной величины X), полученное по n наблюдениям. В данной задаче X – месячный доход жителя региона (в тыс.руб.); $n=1200$ (жителей).

X_i	Менее 3	3-7	7-11	11-15	15-19	19-23	23-27	27-31	31-35	Свыше 35
n_i	60	96	109	156	197	223	171	78	63	47

Необходимо¹: 1) построить гистограмму и эмпирическую функцию распределения X ; 2) найти: а) среднюю арифметическую $\bar{X}$; б) медиану M_e и моду M_0 ; в) дисперсию s^2 , среднеквадратическое отклонение s и коэффициент вариации CV ; г) начальные a_k и центральные m_k моменты k -го порядка ($k=1,2,3,4$); д) коэффициенты асимметрии $\tilde{A}$ и эксцесса E . (¹При наличии открытых интервалов значений X типа «менее X_1 » или «свыше X_n » для проведения расчётов их условно заменяют интервалами той же ширины k , т.е. (X_1-k, X_1) или (X_n, X_n+k)).

3.2. Примеры параметрических семейств распределений

Пусть $X_1, X_2, \dots, X_n$ выборка объема n из генеральной совокупности с функцией распределения $F(x)$. Предположим, что распределение генеральной совокупности зависит от параметров. Чаще всего, в задачах математической статистики параметры распределений неизвестны и могут принимать те или иные значения. В связи с этим вводится некоторое множество Θ — множество возможных значений данного параметра θ . Понятно, что Θ может быть множеством любой природы, но чаще всего $\Theta \subseteq \mathbb{R}^1$ или $\Theta \subseteq \mathbb{R}^2$. Поэтому выражение «дана выборка $X_1, X_2, \dots, X_n$ объема n из генеральной совокупности с функцией распределения $F(x)$ » часто заменяется на следующее «дана выборка $X_1, X_2, \dots, X_n$ из семейства $\{p_\theta(t), \theta \in \Theta\}$ », где для фиксированного θ функция $p_\theta(t)$ определяется однозначно:

$$p_\theta(t) = \begin{cases} \text{плотность распределения, если } \xi - \text{непрерывная сл. величина,} \\ P(\xi = t), \text{ если } \xi - \text{дискретная сл. величина.} \end{cases}$$

Теперь перечислим некоторые важные семейства параметрических распределений, которые часто встречаются в задачах математической статистики. На известных нам из Главы 1 распределениях подробно останавливаться не будем.

1. Нормальное распределение с параметрами μ, σ^2 . (Функцию нормального распределения будем обозначать через Φ_{μ, σ^2}).

2. Гамма распределение с параметрами α и λ , $\alpha, \lambda > 0$ ($\Gamma_{\lambda, \alpha}$)

Плотность распределения имеет следующий вид:

$$p(t) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot t^{\alpha-1} e^{-\lambda t}, & \text{если } t > 0, \\ 0, & \text{иначе,} \end{cases}$$

где $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \cdot e^{-x} dx$ – гамма функция, которая обладает свойствами:

$$\Gamma(1) = 1, \quad \Gamma(\alpha+1) = \alpha \cdot \Gamma(\alpha)$$

Приведем (без вычисления) некоторые числовые характеристики данного распределения: $E\xi = \frac{\alpha}{\lambda}; D\xi = \frac{\alpha}{\lambda^2}$.

3. Показательное распределение с параметром $\lambda, \lambda > 0$ ($\Gamma_{\lambda,1}$).

Это частный случай гамма распределения, когда $\alpha = 1$.

4. Равномерное распределение с параметрами a и b ($U_{a,b}$).

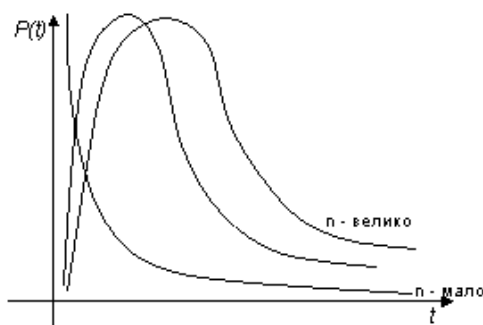
5. Биномиальное распределение с параметрами n и p (B_p^n).

Отметим, что при $n = 1$ получаем распределение Бернулли (B_p^1).

6. Распределение Пуассона с параметром $\lambda, \lambda > 0$ (Π_λ).

7. Распределение хи - квадрат (H_n).

Это тоже частный случай гамма распределения, когда $\alpha = n/2, \lambda = 1/2$. Иногда его обозначают через χ_n^2 . Можно показать, что (см. [2]) случайная величина $\eta = \xi_1^2 + \xi_2^2 + \dots + \xi_n^2$ имеет распределение H_n , где $\xi_i, i = 1, \dots, n$ являются независимыми случайными величинами, имеющие распределение $\Phi_{0,1}$ (стандартное нормальное распределение, которого в предыдущих главах мы обозначали через Φ). Здесь n называют числом степеней свободы распределения H_n . График плотности распределения χ^2 выглядит так



Отметим, то распределение χ^2 приближенно совпадает с нормальным распределением при $n \rightarrow \infty$. Приведем (без вычисления) некоторые числовые характеристики данного распределения $E\eta = n; D\eta = 2n$;

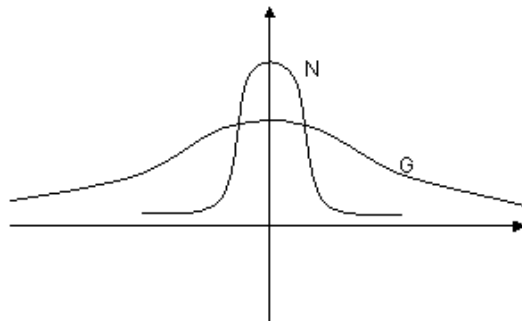
8. Распределение Стьюдента (G_n).

Это распределение следующей случайной величины t :

$$t = \frac{\xi_0}{\sqrt{\frac{1}{n}(\xi_1^2 + \xi_2^2 + \dots + \xi_n^2)}},$$

где $\xi_i, i = 0, 1, \dots, n$ — независимые случайные величины, имеющие распределение $\Phi_{0,1}$. Здесь n также называют числом степеней свободы.

Следует отметить, что при $n \rightarrow \infty$ распределение Стьюдента приближенно совпадает с нормальным распределением. График плотности распределения Стьюдента выглядит так (по сравнению с графиком плотности стандартного нормального закона)



Приведем (без вычисления) некоторые числовые характеристики данного распределения $Et = 0, Dt = n/(n-2)$.

3.3. Оценивание неизвестных параметров

3.3.1. Оценка. Определение

Пусть $X_1, X_2, \dots, X_n$ выборка объема n из семейства $\{p_\theta(t), \theta \in \Theta\}$.

Определение. *Оценкой θ^* неизвестного параметра θ называется всякая функция $\theta^* = \theta^*(X_1, X_2, \dots, X_n)$ от выборки, которая предназначена для использования вместо неизвестного параметра в качестве его приближения.*

Например, функция $\theta^* = \theta^*(X_1, X_2, \dots, X_n) = \frac{1}{n} \cdot \sum_{i=1}^n X_i$ — является оценкой. Отметим, что любая функция от выборки называется в общем случае *статистикой*, и оценка — это та статистика, которая не зависит от параметра.

3.3.2. Свойства оценок

Если рассматривается выборка $X_1, X_2, \dots, X_n$ из семейства распределений $\{p_\theta(t), \theta \in \Theta\}$, то в обозначениях числовых характеристик оценки θ^* (как случайной величины)

будет указан индекс θ , например, $E_{\theta}\theta^*$ — означает математическое ожидание оценки $\theta^* = \theta^*(X_1, \dots, X_n)$, по распределению $\{p_{\theta}(t)\}$, зависящему от неизвестного параметра θ .

Заметим, что произвольная оценка θ^* не может считаться «хорошей» для оценивания параметра θ до тех пор, пока не установлены некоторые свойства, связывающие θ^* и θ . Рассмотрим эти **свойства**.

1. Несмещенность.

Определение 1. Оценка θ^* неизвестного параметра θ называется несмещенной, если $E_{\theta}\theta^* = \theta$ при всех $\theta \in \Theta$.

Если это не так, то $h(\theta) = E_{\theta}\theta^* - \theta$ называется *смещением* оценки θ^* .

Пример. Пусть $X_1, X_2, \dots, X_n$ выборка из распределения $U_{a,b}$, где b — неизвестный параметр. В этом случае, плотность распределения имеет вид:

$$p(t) = \begin{cases} \frac{1}{b}, & \text{если } t \in [0, b], \\ 0, & \text{иначе.} \end{cases}$$

Рассмотрим следующие оценки

$$b_1^* = \frac{X_1^2 + X_2^2 + \dots + X_n^2}{n}, \quad b_2^* = \max\{X_1, X_2, \dots, X_n\}, \quad b_3^* = \frac{2(X_1 + X_2 + \dots + X_n)}{n}$$

Вычислим математические ожидания этих оценок.

Сначала для оценки b_1^* :

$$E_{\theta}b_1^* = E_{\theta} \left(\frac{X_1^2 + X_2^2 + \dots + X_n^2}{n} \right) = \frac{1}{n} \cdot \sum_{i=1}^n E_{\theta}X_i^2 = \frac{1}{n} \cdot n \cdot E_{\theta}X_1^2 = E_{\theta}X_1^2.$$

Далее,

$$E_{\theta}b_1^* = E_{\theta}X_1^2 = \int_0^b t^2 \cdot \frac{1}{b} dt = \frac{b^2}{3},$$

отсюда следует, что $E_b b_1^* \neq b$, что означает — оценка b_1^* является смещенной.

Вычислим теперь $E_b X_2^*$.

$$\mathbf{P}(b_2^* < t) = \mathbf{P}(X_1 < t, X_2 < t, \dots, X_n < t) = (\mathbf{P}(X_1 < t))^n = (U_{0,b}(t))^n.$$

Плотность распределения оценки b_2^* имеет вид:

$$\mathbf{p}(t) = \left((U_{0,b}(t))^n \right)' = n \cdot \left((U_{0,b}(t))^{n-1} \right) \cdot (U_{0,b}(t))' = n \cdot \left(\frac{t}{b} \right)^{n-1} \cdot \frac{t}{b}, \text{ где } t \in [0, b].$$

Отсюда получаем, что

$$\mathbf{E}_\theta b_2^* = \int_0^b n \cdot \left(\frac{t}{b} \right)^{n-1} \cdot \frac{t}{b} dt = \frac{n}{n+1} b,$$

что позволяет нам заключить, что b_2^* смещенная оценка.

И наконец, проверим оценку b_3^* на несмещенность.

$$\mathbf{E}_\theta b_1^* = 2 \cdot \frac{1}{n} \cdot \sum_{i=1}^n \mathbf{E}_\theta X_i = 2 \cdot \frac{1}{n} \cdot n \cdot \mathbf{E}_\theta X_1 = 2 \cdot \frac{b}{2} = b.$$

так как $\mathbf{E}_b X_1 = b/2$, следовательно, b_3^* — несмещенная оценка.

2. Асимптотическая несмещенность.

Определение 2. Оценка θ^* неизвестного параметра θ называется асимптотически несмещенной, если $\mathbf{E}_\theta \theta^* \rightarrow \theta$ при всех $\theta \in \Theta$.

Оценка из предыдущего примера b_1^* не является асимптотически несмещенной, а оценки b_2^* и b_3^* являются асимптотически несмещенными.

3. Состоятельность

Определение 3. Оценка θ^* неизвестного параметра θ называется состоятельной, если

$$\theta^* \xrightarrow{P_\theta} \theta \text{ при } n \rightarrow \infty.$$

Проверим состоятельность вышеприведенных оценок.

Начнем с b_1^* . По закону больших чисел мы получаем, что

$$b_1^* = \frac{X_1^2 + X_2^2 + \dots + X_n^2}{n} \xrightarrow{P_\theta} \mathbf{E}_\theta X_1^2 \text{ при } n \rightarrow \infty.$$

Так как $\mathbf{E}_\theta X_1^2 = \frac{b^2}{3}$, отсюда следует, что условие $b_1^* \xrightarrow{P_\theta} b$ при $n \rightarrow \infty$ не выполняется,

т.е. b_1^* не является состоятельной.

Теперь возьмем оценку b_2^* :

$$\begin{aligned} \mathbf{P}(|b_2^* - b| > \varepsilon) &= \mathbf{P}(b_2^* > b + \varepsilon) + \mathbf{P}(b_2^* < b - \varepsilon) = \left(\mathbf{P}(X_1 < b - \varepsilon) \right)^n = \\ &= \left(\frac{b - \varepsilon}{b} \right)^n = \left(1 - \frac{\varepsilon}{b} \right)^n \rightarrow 0 \text{ при } n \rightarrow \infty, \end{aligned}$$

так как $1 - \frac{\varepsilon}{b} < 1$. Следовательно, оценка b_2^* — состоятельная оценка.

Далее, поскольку по закону больших чисел:

$$b_3^* = 2 \cdot \frac{X_1 + X_2 + \dots + X_n}{n} \xrightarrow{P_\theta} 2 \cdot \mathbf{E}_\theta X_1 = b \text{ при } n \rightarrow \infty,$$

то, получаем, что b_3^* — состоятельная оценка.

4. Асимптотическая нормальность

Определение 4. Оценка, θ^* неизвестного параметра θ называется асимптотически нормальной с коэффициентом σ^2 , если имеет место

$$\mathbf{P}(\sqrt{n} \cdot (\theta^* - \theta) < x) \rightarrow \Phi_{0, \sigma^2}(x).$$

Отметим, что b_1^* и b_2^* — не являются асимптотически нормальными (доказательство приводить не будем).

Рассмотрим оценку b_3^* . По центральной предельной теореме :

$$\begin{aligned} \mathbf{P}(\sqrt{n} \cdot (b_3^* - b) < x) &= \mathbf{P}\left(\sqrt{n} \cdot \frac{2 \cdot \sum_{i=1}^n X_i - 2n \cdot \mathbf{E}_\theta X_1}{n} < x\right) = \\ &= \mathbf{P}\left(2 \cdot \frac{\sum_{i=1}^n (X_i - \mathbf{E}_\theta X_1)}{\sqrt{n}} < x\right) \rightarrow \Phi_{0,1}\left(\frac{x\sqrt{3}}{b}\right). \end{aligned}$$

Поскольку

$$\begin{aligned}\Phi_{0,1}\left(\frac{x\sqrt{3}}{b}\right) &= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x\sqrt{3}/b} e^{-t^2/2} dt = \left| t = \frac{\sqrt{3}}{b} \cdot y \right| = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^x e^{-3y^2/2b^2} \frac{\sqrt{3}}{b} dy \\ &= \frac{1}{(b/\sqrt{3}) \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^x e^{-3y^2/2b^2} dy = \Phi_{0,1(b/\sqrt{3})}(x).\end{aligned}$$

Итак, мы получили, что b_3^* — асимптотически нормальная оценка с коэффициентом $b/\sqrt{3}$.

Задачи к 3.3

1. Имеются следующие результаты выборок, извлечённых из нормально распределённой генеральной совокупности:

$$n = 9; \sum X_i = 36; \sum (X_i - \bar{X})^2 = 288; \quad n = 16; \sum X_i = 64; \sum (X_i - \bar{X})^2 = 180;$$

$$n = 25; \sum X_i = 500; \sum X_i^2 = 11400.$$

Что является лучшей оценкой средней арифметической, дисперсии, среднеквадратического отклонения выборочных данных?

2. Найти несмещённую оценку дисперсии случайной величины ξ на основании данного распределения выборки:

X_i	1	5	6	8
n_i	6	4	7	3

3.4. Методы получения оценок

3.4.1. Метод моментов

Метод моментов был предложен известным английским статистиком К.Пирсоном и отличается простотой получения оценок. Суть этого метода заключается в приравнивании друг другу теоретических и выборочных моментов.

Пусть $X_1, X_2, \dots, X_n$ выборка объема n из семейства распределений $\{p_\theta(t), \theta \in \Theta\}$. Пусть $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ — неизвестный параметр, и предположим, что существуют первые k моментов у случайной величины ξ с распределением $p_\theta(t)$: $\alpha_i = E\xi^i, i = 1, 2, \dots, k$. С

помощью выборки $X_1, X_2, \dots, X_n$, мы можем вычислить первые k выборочных моментов:

$$a_i = \frac{1}{n} \cdot \sum_{j=1}^n X_j^i, \quad i = 1, 2, \dots, k.$$

Составим систему уравнений, приравняв, например, соответствующие теоретические и выборочные начальные моменты:

$$\begin{cases} \alpha_1(\theta_1, \theta_2, \dots, \theta_k) = a_1, \\ \alpha_2(\theta_1, \theta_2, \dots, \theta_k) = a_2, \\ \dots \\ \alpha_k(\theta_1, \theta_2, \dots, \theta_k) = a_k. \end{cases}$$

Решение этой системы $\theta^* = (\theta_1^*, \dots, \theta_k^*)$ и будет оценкой, полученной по методу моментов.

Отметим, что можно составить и систему уравнений, приравняв соответствующие теоретические и выборочные центральные моменты:

$$\begin{cases} \mu_1(\theta_1, \theta_2, \dots, \theta_k) = m_1, \\ \mu_2(\theta_1, \theta_2, \dots, \theta_k) = m_2, \\ \dots \\ \mu_k(\theta_1, \theta_2, \dots, \theta_k) = m_k. \end{cases}$$

и решив эту систему, получим оценку $\theta^* = (\theta_1^*, \dots, \theta_k^*)$ по методу моментов.

Пример. Пусть $X_1, X_2, \dots, X_n$ выборка из показательного распределения $\Gamma_{\lambda,1}$ с плотностью распределения

$$p(t) = \begin{cases} \lambda \cdot e^{-\lambda t}, & \text{если } t > 0, \\ 0, & \text{иначе.} \end{cases}$$

Известно, что $E_\lambda X_1 = 1/\lambda$. Так как $a_1 = E_\lambda X_1$ — момент порядка 1, то $\lambda = 1/a_1$. По определению, оценкой λ_1^* неизвестного параметра λ по методу моментов будет следующая оценка: $\lambda_1^* = \frac{1}{a_1} = \frac{1}{\bar{X}}$, где $\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$.

3.4.2. Метод максимального правдоподобия

Пусть $X_1, X_2, \dots, X_n$ выборка объема n из семейства распределений $\{p_\theta(t), \theta \in \Theta\}$

Определение 1. Статистика

$$f(\theta) = \prod_{i=1}^n p_{\theta}(X_i) = p_{\theta}(X_1) \cdot p_{\theta}(X_2) \cdot \dots \cdot p_{\theta}(X_n)$$

называется функцией правдоподобия, а $L_n(\theta) = \ln f(\theta)$ — логарифмической функцией правдоподобия.

Определение 2. Оценка θ^* неизвестного параметра θ называется оценкой максимального правдоподобия, если

$$f(\theta^*) = \max_{\theta} f(\theta), \text{ или } L_n(\theta^*) = \max_{\theta} L_n(\theta).$$

Алгоритм получения такой оценки следующий: если $f(\theta)$, и соответственно $L_n(\theta)$, гладкая функция θ (т.е. непрерывно дифференцируемая в окрестности каждой точки), тогда оценка максимального правдоподобия ищется как решение уравнения (системы уравнений, если параметр многомерный):

$$\frac{\partial f(\theta)}{\partial \theta} = 0 \text{ или } \frac{\partial L_n(\theta)}{\partial \theta} = 0,$$

поскольку

$$\frac{\partial L_n(\theta)}{\partial \theta} = \frac{\partial \ln f(\theta)}{\partial \theta} = \frac{1}{f(\theta)} \cdot \frac{\partial f(\theta)}{\partial \theta}.$$

Пример 1. Пусть $X_1, X_2, \dots, X_n$ выборка из распределения Бернулли B_p^1 , $\theta = p$ — неизвестный параметр. Напомним, что $p_{\theta}(X_i) = \begin{cases} p, & \text{если } X_i = 1, \\ 1-p, & \text{если } X_i = 0, \end{cases}$ и v — число успехов (число единиц в выборке).

Функция правдоподобия имеет вид:

$$f(\theta) = p^v (1-p)^{n-v},$$

а логарифмическая функция правдоподобия

$$L_n(\theta) = \ln f(\theta) = \ln(p^v (1-p)^{n-v}) = v \cdot \ln p + (n-v) \cdot \ln(1-p).$$

Тогда

$$\frac{\partial L_n(p)}{\partial p} = \frac{v}{p} + \frac{n-v}{1-p} \cdot (-1),$$

и получаем уравнение

$$\frac{v}{p} - \frac{n-v}{1-p} = 0.$$

Решение этого уравнения и есть оценка максимального правдоподобия

$$p^* = \frac{v}{n}.$$

Пример 2. Пусть $X_1, X_2, \dots, X_n$ выборка из нормального распределения Φ_{a, σ^2} , где $\theta = (a, \sigma^2)$ — неизвестный параметр. Напомним, что

$$p_{\theta}(X_i) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(X_i - a)^2}{2\sigma^2}}.$$

Функция правдоподобия имеет следующий вид:

$$f(\theta) = f(a, \sigma^2) = \left(\frac{1}{\sigma \cdot \sqrt{2\pi}} \right)^n \cdot e^{-\frac{\sum_{i=1}^n (X_i - a)^2}{2\sigma^2}}$$

Здесь удобно воспользоваться логарифмической функцией правдоподобия:

$$L_n(a, \sigma^2) = \ln \left(\frac{1}{\sigma \cdot \sqrt{2\pi}} \right)^n - n \cdot \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - a)^2.$$

Теперь составим систему уравнений правдоподобия:

$$\begin{cases} \frac{\partial L_n(a, \sigma^2)}{\partial a} = \frac{1}{2\sigma^2} \cdot 2 \cdot \sum_{i=1}^n (X_i - a) = 0, \\ \frac{\partial L_n(a, \sigma^2)}{\partial \sigma^2} = \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - a)^2 - \frac{n}{\sigma} = 0. \end{cases}$$

Решив эту систему, получаем оценки максимального правдоподобия для параметров a и σ^2 :

$$a^* = \frac{1}{n} \cdot \sum_{i=1}^n X_i = \bar{X}, \quad (\sigma^2)^* = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = s^2.$$

Теперь рассмотрим пример отыскания оценки максимального правдоподобия в случае, когда $f(\theta)$ не является гладкой функцией.

Пример 3. Пусть $X_1, X_2, \dots, X_n$ выборка из равномерного распределения $U_{0,b}$, $\theta = b$ — неизвестный параметр. Напомним, что

$$p(t) = \begin{cases} \frac{1}{b}, & \text{если } X_i \in [0, b], \\ 0, & \text{иначе.} \end{cases}$$

Функция правдоподобия имеет следующий вид:

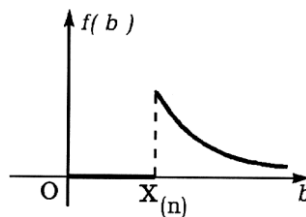
$$f(\theta) = \begin{cases} \frac{1}{b^n}, & \text{если } X_i \in [0, b], \text{ для всех } i = 1, 2, \dots, n, \\ 0, & \text{иначе.} \end{cases}$$

Очевидно, что эта функция не является гладкой. Мы должны найти такое значение b^* , которое удовлетворяет равенству $f(b^*) = \max_b f(b)$. Преобразуем функцию $f(b)$. Легко убедиться в том, что $X_i \in [0, b]$ для любого i эквивалентно $b \in [\max_{1 \leq i \leq n} X_i, \infty)$, поскольку $b \geq \max_{1 \leq i \leq n} X_i$. Напомним, что $\max_{1 \leq i \leq n} X_i = X_{(n)}$ - последний член вариационного ряда.

С учетом вышесказанного получаем

$$f(\theta) = \begin{cases} \frac{1}{b^n}, & \text{если } b \in [X_{(n)}, \infty), \\ 0, & \text{если } b \in [0, X_{(n)}]. \end{cases}$$

Построим график этой функции:



Из графика видно, что $\max_b f(b)$ достигается в точке $b^* = X_{(n)}$.

Задачи 3.4

1. Для определения средней заработной платы работников определённой отрасли было обследовано 100 человек. Результаты представлены в следующей таблице (данные условные):

Зарплата долларах	в	Число человек	Зарплата долларах	в	Число человек
190-210		1	290-310		19
210-230		5	310-330		11
230-250		9	330-350		4
250-270		22	350-370		1
270-290		28			

Построить гистограмму и график эмпирической функции распределения, найти оценки математического ожидания и дисперсии зарплаты наугад взятого работника.

2. При условии равномерного распределения случайной величины X

$$p(x) = \begin{cases} \frac{1}{b-a}, & \text{если } x \in (a, b), \\ 0, & \text{если } x \notin (a, b) \end{cases}$$

произведена выборка

X_i	2	3	4	5	6
n_i	4	6	5	12	8

Найти оценку параметров a и b .

3. Случайная величина подчиняется нормальному закону распределения с плотностью

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/(2\sigma^2)}.$$

Произведена выборка

X_i	3	5	7	9	11	13	15
n_i	6	9	16	25	20	16	8

Найти оценку параметра a и несмещённую оценку параметра σ .

4. Стекланные однородные изделия отправлены для реализации из Москвы в Новосибирск в 1000 контейнерах. После поступления товара было выявлено количество разбитых изделий в каждом контейнере. Результаты представлены в таблице:

X_i	0	1	2	3	4
n_i	785	163	32	16	4

Считая, что число разбитых изделий описывается законом Пуассона, найти точечную оценку параметра λ .

3.5. Сравнение оценок. Неравенство Рао - Крамера

С помощью различных методов мы получаем множество оценок и нам нужно определить лучшие из них. Для сравнения оценок рассмотрим два подхода.

3.5.1. Среднеквадратический подход

Пусть $X_1, X_2, \dots, X_n$ выборка объема n из семейства распределений $\{p_\theta(t), \theta \in \mathfrak{R}^1\}$ и θ_1^* , θ_2^* — две оценки неизвестного параметра θ .

Определение 1. Оценка θ_1^* лучше, чем θ_2^* , если выполняется

$$E_\theta(\theta_1^* - \theta)^2 < E_\theta(\theta_2^* - \theta)^2.$$

Отметим, что если в определении 1 знак $<$ заменить на $\leq, >, \geq$, то получим понятия: θ_1^* «не хуже», «хуже», «не лучше», чем θ_2^* , соответственно.

Если $h(\theta) = E_\theta \theta^* - \theta$ смещение оценки θ^* , то

$$\begin{aligned} E_\theta(\theta^* - \theta)^2 &= E_\theta(\theta^* + h(\theta) - E_\theta \theta^*)^2 \\ &= E_\theta(\theta^* - E_\theta \theta^*)^2 + E_\theta h^2 = D_\theta \theta^* + (h(\theta))^2. \end{aligned}$$

Так как для несмещенных оценок θ_1^* и θ_2^* смещение $h(\theta) = 0$, то в определении 1 величины $E_\theta(\theta_i^* - \theta)^2$ заменяются на $D_\theta \theta_i^*$, $i = 1, 2$. Тем самым, сравнение несмещенных оценок сводится к сравнению дисперсии этих оценок.

Определение 2. В классе несмещенных оценок неизвестного параметра θ оценка θ^* с минимальной дисперсией называется эффективной оценкой.

3.5.2. Асимптотический подход

Пусть θ_1^* и θ_2^* две асимптотически нормальные оценки, т.е.

$$\frac{\sqrt{n} \cdot (\theta_1^* - \theta)}{\sigma_1} \rightarrow \Phi_{0,1} \text{ и } \frac{\sqrt{n} \cdot (\theta_2^* - \theta)}{\sigma_2} \rightarrow \Phi_{0,1} \text{ при } n \rightarrow \infty,$$

или

$$\sqrt{n} \cdot (\theta_1^* - \theta) \rightarrow \Phi_{0, \sigma_1^2} \text{ и } \sqrt{n} \cdot (\theta_2^* - \theta) \rightarrow \Phi_{0, \sigma_2^2} \text{ при } n \rightarrow \infty.$$

Определение 1. Оценка θ_1^* лучше, чем θ_2^* , если $\sigma_1^2 < \sigma_2^2$.

Естественно асимптотический подход менее предпочтителен, поскольку может быть применен в случае выборки большого объема и только в классе асимптотически нормальных оценок.

Определение 2. В классе асимптотически нормальных оценок параметра θ оценка θ^* с минимальным коэффициентом σ^2 называется асимптотически эффективной оценкой.

3.5.3. Неравенство Рао - Крамера

Попытаемся указать нижнюю границу дисперсии $\mathbf{D}_\theta \theta^*$ оценки θ^* . Этот вопрос решается с помощью неравенства Рао - Крамера.

Пусть $X_1, X_2, \dots, X_n$ выборка из семейства распределений $\{\mathbf{p}_\theta(t), \theta \in \mathfrak{R}^1\}$. Рассмотрим функцию $L(t, \theta) = \ln \mathbf{p}_\theta(t)$ и найдем ее производную по θ :

$$L'_\theta(\theta, t) = \frac{\partial \ln \mathbf{p}_\theta(t)}{\partial \theta} = \frac{1}{\mathbf{p}_\theta(t)} \cdot \frac{\partial \mathbf{p}_\theta(t)}{\partial \theta}$$

Предположим, что выполнено некоторое условие регулярности (R):

а) в случае распределения **непрерывного** типа функции $\mathbf{p}_\theta(t)$ непрерывно дифференцируемы по θ для почти всех t , а интегралы

$$\mathbf{I}(\theta) = \mathbf{E}_\theta \left(L'_\theta(\theta, t) \right)^2 = \int_{-\infty}^{\infty} \left(\frac{\partial \ln \mathbf{p}_\theta(t)}{\partial \theta} \right)^2 \cdot \mathbf{p}_\theta(t) dt = \int_{-\infty}^{\infty} \frac{1}{\mathbf{p}_\theta(t)} \cdot \left(\frac{\partial \mathbf{p}_\theta(t)}{\partial \theta} \right)^2 dt$$

существуют и непрерывны по θ ;

б) в случае распределения **дискретного** типа существуют частные производные $\frac{\partial \mathbf{p}_\theta(t)}{\partial \theta}$,
а ряды

$$\mathbf{I}(\theta) = \mathbf{E}_\theta \left(L'_\theta(\theta, t) \right)^2 = \sum_i \left(\frac{\partial \ln \mathbf{p}_\theta(i)}{\partial \theta} \right)^2 \cdot \mathbf{p}_\theta(i) = \sum_i \frac{1}{\mathbf{p}_\theta(i)} \left(\frac{\partial \mathbf{p}_\theta(i)}{\partial \theta} \right)^2$$

сходятся абсолютно и равномерно в $\mathfrak{R}^1$,

где $\mathbf{p}_\theta(i)$ —вероятности принятия значений i дискретной случайной величины.

$\mathbf{I}(\theta)$ —называется *информационным количеством Фишера*. Без доказательства (его можно найти в [4]) сформулируем следующую теорему.

Теорема (неравенство Рао - Крамера). Пусть выполнено условие (R). Тогда для оценки θ^* неизвестного параметра θ справедливо следующее неравенство

$$\mathbf{D}_\theta \theta^* \geq \frac{(1+h'(\theta))^2}{n \cdot \mathbf{I}(\theta)}.$$

Следствия теоремы.

1. При выполнении условий теоремы справедливо неравенство

$$\mathbf{E}_\theta (\theta^* - \theta)^2 \geq \frac{(1+h'(\theta))^2}{n \cdot \mathbf{I}(\theta)} + (h(\theta))^2,$$

если же θ^* — несмещенная оценка, то справедливо неравенство

$$\mathbf{D}_\theta \theta^* \geq \frac{1}{n \cdot \mathbf{I}(\theta)}.$$

2. Если выполнено условие (R) и в неравенстве Рао-Крамера достигается равенство, то θ^* — эффективная оценка в классе оценок со смещением $h(\theta)$, т.е.

$$\mathbf{E}_\theta (\theta^* - \theta)^2 = \frac{(1+h'(\theta))^2}{n \cdot \mathbf{I}(\theta)} + (h(\theta))^2.$$

Пример. Пусть $X_1, X_2, \dots, X_n$ выборка объема n из нормального распределения Φ_{a, σ^2} .

Предположим, что σ известно, и $\theta = a$ —неизвестный параметр. Вычислим $\mathbf{I}(\theta) = \mathbf{I}(a)$

$$\begin{aligned} \mathbf{I}(\theta) &= \mathbf{E}_\theta \left(L'_\theta(\theta, t) \right)^2 = \int_{-\infty}^{\infty} \left(\frac{\partial \ln \mathbf{p}_\theta(t)}{\partial \theta} \right)^2 \cdot \mathbf{p}_\theta(t) dt = \\ &= \int_{-\infty}^{\infty} \left(\frac{t-a}{\sigma^2} \right)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-a)^2}{2\sigma^2}} dt = \frac{1}{\sigma^4} \cdot \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (t-a)^2 \cdot e^{-\frac{(t-a)^2}{2\sigma^2}} dt = \frac{1}{\sigma^2}, \end{aligned}$$

поскольку

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (t-a)^2 \cdot e^{-\frac{(t-a)^2}{2\sigma^2}} dt = \int_{-\infty}^{\infty} (t - \mathbf{E}\xi)^2 \cdot \mathbf{p}_\xi(t) dt = \mathbf{D}\xi$$

—дисперсия случайной величины ξ , имеющей распределение Φ_{a, σ^2} .

Рассмотрим оценку $\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$. Используя свойства математического ожидания и дисперсии (см. 2.6) мы получаем

$$\begin{aligned} \mathbf{E}_\theta \bar{X} &= \mathbf{E}_\theta \left(\frac{1}{n} \cdot \sum_{i=1}^n X_i \right) = \frac{1}{n} \cdot \sum_{i=1}^n \mathbf{E}_\theta X_i = \frac{1}{n} \cdot n \cdot \mathbf{E}_\theta X_1 = a, \\ \mathbf{D}_\theta \bar{X} &= \mathbf{D}_\theta \left(\frac{1}{n} \cdot \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \cdot \sum_{i=1}^n \mathbf{D}_\theta X_i = \frac{1}{n^2} \cdot n \cdot \mathbf{D}_\theta X_1 = \frac{1}{n} \cdot \sigma^2. \end{aligned}$$

Таким образом, согласно следствия 2 теоремы 1 оценка $\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$ - является эффективной оценкой для a , поскольку в неравенстве Рао-Крамера достигается равенство, т.е. $\mathbf{D}\bar{X} = \frac{1}{n} \cdot \sigma^2$.

3.6. Построение доверительных интервалов

Пусть $X_1, X_2, \dots, X_n$ выборка из семейства распределений $\{\mathbf{p}_\theta(t), \theta \in \mathfrak{R}^1\}$, θ - неизвестный параметр.

Требуется указать интервал $(\theta^-, \theta^+) = (\theta^-(X_1, X_2, \dots, X_n); \theta^+(X_1, X_2, \dots, X_n))$, который с заданной достаточно высокой вероятностью будет покрывать неизвестное значение параметра θ .

Определение 1. Доверительным интервалом для неизвестного параметра θ уровня доверия $1 - \varepsilon$, $0 < \varepsilon < 1$, называется интервал (θ^-, θ^+) , построенный по выборке, и такой, что неизвестное значение параметра θ покрывается этим интервалом с вероятностью $1 - \varepsilon$, т.е.

$$\mathbf{P}(\theta^- < \theta < \theta^+) = 1 - \varepsilon.$$

Пример 1. Пусть у нас имеется одно наблюдение X_1 из равномерно распределения $U_{0,b}$, где b — неизвестный параметр. Необходимо указать интервал (b^-, b^+) такой, что

$$\mathbf{P}(b^- < b < b^+) = 1 - \varepsilon,$$

где $1 - \varepsilon$ — уровень доверия.

Пусть B^- и B^+ из $[0, b]$ такие, что $\mathbf{P}(X_1 \in (B^-, B^+)) = 1 - \varepsilon$ (*). Предположим, что

$$\mathbf{P}(X_1 \in [0, B^-]) = \mathbf{P}(X_1 \in [B^+, b]) = \frac{\varepsilon}{2}.$$

С другой стороны

$$\mathbf{P}(X_1 \in [0, B^-]) = \frac{1}{b} \cdot B^- \quad \text{и} \quad \mathbf{P}(X_1 \in [B^+, b]) = \frac{b - B^+}{b},$$

поскольку X_1 имеет равномерно распределение $U_{0,b}$. Тогда мы получим

$$\frac{1}{b} \cdot B^- = \frac{\varepsilon}{2} \quad \text{и} \quad \frac{b - B^+}{b} = \frac{\varepsilon}{2},$$

и отсюда $B^- = \frac{\varepsilon}{2} \cdot b$ и $B^+ = b \left(1 - \frac{\varepsilon}{2}\right)$. Подставляя полученные значения B^- и B^+ в равенство (*) имеем

$$1 - \varepsilon = \mathbf{P}\left(\frac{\varepsilon}{2} \cdot b < X_1 < b \left(1 - \frac{\varepsilon}{2}\right)\right) = \mathbf{P}\left(X_1 \cdot \frac{2}{2 - \varepsilon} < b < X_1 \cdot \frac{2}{\varepsilon}\right) = \mathbf{P}(b^- < b < b^+)$$

где $b^- = X_1 \cdot \frac{2}{2 - \varepsilon}$, $b^+ = X_1 \cdot \frac{2}{\varepsilon}$.

Итак, доверительным интервалом для b будет $\left(X_1 \cdot \frac{2}{2 - \varepsilon}, X_1 \cdot \frac{2}{\varepsilon}\right)$.

Пример 2. Построение доверительных интервалов для параметров нормального распределения.

Пусть $X_1, X_2, \dots, X_n$ выборка объема n из нормального распределения Φ_{a, σ^2} . Здесь мы построим доверительные интервалы для

1) параметра a , если σ^2 известно,

2) параметра a , если σ^2 неизвестно,

3) параметра σ^2 , если a известно,

4) параметра σ^2 , если a неизвестно.

Итак, рассмотрим каждый случай отдельно.

1). Предположим, что σ^2 известно. Согласно центральной предельной теореме (см. 2.7.3)

$$\eta_0 = \sqrt{n} \cdot \frac{\bar{X} - a}{\sigma} \rightarrow \Phi_{0,1},$$

где $\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$.

Предположим, что $k_{\varepsilon} > 0$ таково, что $\mathbf{P}(|\eta_0| < k_{\varepsilon}) = 1 - \varepsilon$, где $1 - \varepsilon$ — уровень доверия.

С другой стороны,

$$\begin{aligned} \mathbf{P}(|\eta_0| < k_{\varepsilon}) &= \mathbf{P}(-k_{\varepsilon} < \eta_0 < k_{\varepsilon}) = \Phi_{0,1}(k_{\varepsilon}) - \Phi_{0,1}(-k_{\varepsilon}) = 2 \cdot \Phi_{0,1}(k_{\varepsilon}) - 1 \\ &= 2 \cdot (0,5 + \Phi_0(k_{\varepsilon})) - 1 = 2 \cdot \Phi_0(k_{\varepsilon}) \quad (\text{см. 2.5.3}). \end{aligned}$$

отсюда $\Phi_0(k_{\varepsilon}) = (1 - \varepsilon)/2$. Значение k_{ε} находим из таблицы значений функции $\Phi_0(k_{\varepsilon})$ (см.

Таблицу 1 Приложения). При найденном k_{ε} имеет место

$$\mathbf{P}\left(-k_{\varepsilon} < \sqrt{n} \cdot \frac{\bar{X} - a}{\sigma} < k_{\varepsilon}\right) = 1 - \varepsilon,$$

и следовательно, доверительным интервалом параметра a будет интервал:

$$\left(\bar{X} - k_{\varepsilon} \cdot \frac{\sigma}{\sqrt{n}}; \bar{X} + k_{\varepsilon} \cdot \frac{\sigma}{\sqrt{n}}\right).$$

2). Предположим, что σ^2 неизвестно. Воспользуемся оценкой s_0^2 параметра σ^2 :

$$s_0^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2,$$

поскольку она несмещенная. Известно (см. [2]), что

$$\eta_1 = \sqrt{n} \cdot \frac{\bar{X} - a}{s_0} \rightarrow \mathbf{G}_{n-1},$$

где, напомним, $\mathbf{G}_{n-1}$ — распределение Стьюдента с $n - 1$ степенями свободы.

Имеется таблица значений функции $\mathbf{G}_{n-1}(x)$ (см. Таблицу 2 Приложения), из которой найдем k_ε .

С другой стороны, $\mathbf{P}(|\eta_1| < k_\varepsilon) = 1 - \varepsilon$, поскольку $\mathbf{P}(|\eta_1| \geq k_\varepsilon) = 2(1 - \mathbf{G}_{n-1}(k_\varepsilon))$ (см.[4]). Отсюда получаем, что

$$\mathbf{P}\left(-k_\varepsilon < \sqrt{n} \cdot \frac{\bar{X} - a}{s_0} < k_\varepsilon\right) = 1 - \varepsilon,$$

и следовательно, доверительным интервалом для параметра a будет

$$\left(\bar{X} - k_\varepsilon \cdot \frac{s_0}{\sqrt{n}}; \bar{X} + k_\varepsilon \cdot \frac{s_0}{\sqrt{n}}\right).$$

3) Теперь оценим σ^2 , если a известно. Воспользуемся оценкой

$$s_1^2 = \frac{1}{n} \cdot \sum_{i=1}^n (X_i - a)^2.$$

Поскольку

$$\mathbf{E}_\theta\left(\frac{X_i - a}{\sigma}\right) = 0 \text{ и } \mathbf{D}_\theta\left(\frac{X_i - a}{\sigma}\right) = 1, i = 1, 2, \dots, n,$$

то согласно 3.2

$$\eta_2 = \frac{n \cdot s_1^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - a}{\sigma}\right)^2 \sim \mathbf{H}_n.$$

У функции $\mathbf{H}_n(x)$ тоже имеется таблица значений (см. Таблицу 3 Приложения).

Найдем интервал $\left(\chi_{n, \frac{\varepsilon}{2}}^2, \chi_{n, 1 - \frac{\varepsilon}{2}}^2\right)$, такой, что

$$\mathbf{P}\left(\chi_{n, \frac{\varepsilon}{2}}^2 < \eta_2 < \chi_{n, 1 - \frac{\varepsilon}{2}}^2\right) = 1 - \varepsilon.$$

Чтобы воспользоваться таблицей, предположим, что

$$\mathbf{P}\left(\eta_2 \leq \chi_{n, \frac{\varepsilon}{2}}^2\right) = \mathbf{P}\left(\eta_2 \geq \chi_{n, 1 - \frac{\varepsilon}{2}}^2\right) = \frac{\varepsilon}{2}.$$

Теперь, из таблицы значений $\mathbf{H}_n(x)$ при заданном ε находим $\chi_{n, \frac{\varepsilon}{2}}^2$, и

воспользовавшись равенством

$$\mathbf{P}\left(\eta_2 \geq \chi_{n, 1 - \frac{\varepsilon}{2}}^2\right) = 1 - \mathbf{P}\left(\eta_2 \leq \chi_{n, \frac{\varepsilon}{2}}^2\right) = 1 - \frac{\varepsilon}{2},$$

находим $\chi^2_{n,1-\frac{\varepsilon}{2}}$. Тем самым, получаем

$$\chi^2_{n,\frac{\varepsilon}{2}} < \frac{n \cdot s_1^2}{\sigma^2} < \chi^2_{n,1-\frac{\varepsilon}{2}},$$

и следовательно, доверительным интервалом для σ^2 будет интервал

$$\left(\frac{n \cdot s_1^2}{\chi^2_{n,1-\frac{\varepsilon}{2}}}, \frac{n \cdot s_1^2}{\chi^2_{n,\frac{\varepsilon}{2}}} \right).$$

4). И наконец, найдем доверительный интервал для σ^2 , если a неизвестно. Воспользуемся статистикой

$$\eta_3 = \frac{(n-1) \cdot s_0^2}{\sigma^2} \sim \mathbf{H}_{n-1}.$$

Аналогично предыдущему пункту находим $\chi^2_{n-1,\frac{\varepsilon}{2}}$ и $\chi^2_{n-1,1-\frac{\varepsilon}{2}}$, и получаем доверительный интервал для σ^2 , когда a неизвестно,

$$\left(\frac{(n-1) \cdot s_0^2}{\chi^2_{n-1,1-\frac{\varepsilon}{2}}}, \frac{(n-1) \cdot s_0^2}{\chi^2_{n-1,\frac{\varepsilon}{2}}} \right).$$

Задачи к 3.6

1. Вам нужно найти пять человек, пользующихся услугами некоторой фирмы. При опросе на улице случайных прохожих оказалось, что 10, 15, 20, 25 и 30-й прохожие пользуются услугами фирмы. Методами моментов и максимального правдоподобия оцените вероятность того, что случайный прохожий пользуется услугами фирмы; найдите для этой вероятности 90%-ные доверительные границы.

2. Фирма с целью установления известности её продукции опросила на каждой из пяти улиц по 40 человек. Количество знакомых с продукцией фирмы оказалось таким: 20, 10, 30, 10, 15.

а) Методами моментов и максимального правдоподобия оцените степень известности и продукции фирмы;

б) постройте 90%-ный и 95%-ный доверительные интервалы для степени известности продукции. Какой из интервалов шире и почему?

в) пользуясь 95%-ным доверительным интервалом, оцените число жителей среди 2000, знакомых с продукцией фирмы.

3. Из 200 работников банка случайным образом отобрано 20 человек, средняя зарплата которых составила 600 у.е., а среднеквадратическое отклонение 100 у.е. Предположив, что зарплата распределена по нормальному закону, определите с 95%-ной надёжностью среднюю зарплату в банке и суммарные затраты банка на зарплату в месяц.

4. При проверке двух предприятий розничной торговли ревизор установил, что в одном магазине для случайной выборки $n=10$ счетов среднее сальдо счёта равно 54 у.е., а в другом, при таком же объёме выборки, 45 у.е. Используя 95%-ные доверительные границы, оцените разность средних сальдо счетов для двух магазинов, если среднее квадратичное отклонение сальдо для первого магазина $\sigma_1=3$ у.е., а для второго $\sigma_2=2$ у.е. Предполагается нормальное распределение сальдо счёта.

5. На овцеводческой ферме из стада произведена выборка для взвешивания 36 овец. Их средний вес оказался равным 50 кг. Предположив распределение веса нормальным и определив несмещённую оценку выборочной дисперсии $s^2=16$, найти доверительный интервал для оценки математического ожидания с надёжностью а) 0,8; б) 0,9; в) 0,95.

3.7. Проверка статистических гипотез

3.7.1. Основные понятия

Пусть $X_1, X_2, \dots, X_n$ выборка объема n из генеральной совокупности с функцией распределения $F(x)$. В задачах проверки статистических гипотез $F(x)$ называют *теоретической* функцией распределения, выборку для удобства обозначают через $X = (X_1, X_2, \dots, X_n)$. В связи с этим отметим еще раз, что выборку можно рассматривать как n -мерный случайный вектор, где $X_1, X_2, \dots, X_n$ независимые (в совокупности) случайные величины, имеющие одну и ту же функцию распределения $F(x)$.

В настоящем параграфе речь пойдет о проверке каких-либо предположений (гипотез) относительно распределения $F(x)$. Например, « X —выборка из генеральной совокупности с нормальным распределением» или « X —выборка из генеральной совокупности с равномерным распределением $U_{0,b}$ » и т.д.

Определение 1. *Статистической гипотезой (или просто гипотезой) называется любое утверждение относительно свойств генеральной совокупности, об истинности (справедливости) которого мы судим по выборочным данным $X_1, X_2, \dots, X_n$.*

Гипотезы бывают простыми и сложными.

Определение 2. *Если гипотеза полностью (однозначно) определяет распределение генеральной совокупности, то она называется простой, в противном случае—сложной гипотезой.*

Например, гипотеза « X —выборка из генеральной совокупности с нормальным распределением $\Phi_{0,1}$ »—простая, а гипотеза « X —выборка из генеральной совокупности с нормальным распределением»— сложная.

По смысловому содержанию выделим некоторые типы гипотез.

1. Гипотеза согласия. Пусть $X_1, X_2, \dots, X_n$ выборка объема n из генеральной совокупности с функцией распределения $F(x)$. Гипотеза согласия — это предположение о том, что неизвестная функция распределения (теоретическая) $F(x)$ совпадает с функцией распределения (гипотетической) $F_0(x)$ которая точно известна, и гипотеза выражается (обозначается) так:

$$H: F(x) = F_0(x).$$

Пример 1. Пусть X — выборка объема n из генеральной совокупности с функцией распределения $F(x)$. Проверяем гипотезу о том, что эта выборка из показательного распределения с параметром λ , т.е.

$$H : F(x)=\Gamma_{\lambda,1}(x),$$

где, как известно, $\Gamma_{\lambda,1}(x)$ — функция распределения показательного закона с параметром λ .

2. Гипотеза однородности (двух выборок). Пусть $X = (X_1, X_2, \dots, X_n)$ и $Y = (Y_1, Y_2, \dots, Y_n)$ две выборки из генеральных совокупностей с функциями распределений $F_0(x)$ и $F_1(x)$, соответственно. Гипотеза однородности состоит в следующем:

$$H: F_0(x) = F_1(x).$$

Пример 2. если X — выборка из генеральной совокупности с функцией распределения $\Phi_{a_1, \sigma^2}(x)$, а Y — выборка из генеральной совокупности с функцией распределения $\Phi_{a_2, \sigma^2}(x)$, где параметр σ^2 —один и тот же, то гипотезой однородности будет следующая гипотеза:

$$H: a_1 = a_2.$$

3. Гипотеза некоррелированности. Предположим, что мы имеем выборку n пар значений $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ двумерной случайной величины $\xi = (\xi_1, \xi_2)$ (см. 3.1.3). Рассмотрим следующую величину

$$r(X, Y) = \frac{\frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_X \cdot s_Y},$$

которая называется *выборочным коэффициентом корреляции*, где

$$s_X^2 = \frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})^2, \quad s_Y^2 = \frac{1}{n} \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Тогда гипотеза $H : r(X, Y) = 0$ будет гипотезой некоррелированности случайных величин ξ_1 и ξ_2 .

3.7.2. Принцип Неймана-Пирсона построения критериев.

Лемма Неймана - Пирсона

Во многих приложениях возникают задачи проверки многих гипотез. Эту задачу можно описать следующим образом.

Пусть задано конечное разбиение параметрического множества $\Theta = \Theta_1 \cup \Theta_2 \cup \dots \cup \Theta_m$. Мы проверяем, какому из подмножеств Θ_j принадлежит неизвестный параметр θ . Если проверка покажет, что $\theta \in \Theta_k$, решение интерпретируется как *принятие* гипотезы $H_k : \{\theta \in \Theta_k\}$ и отвержение остальных $m - 1$ гипотез $H_j : \{\theta \in \Theta_j\}$, $j=1, \dots, m, j \neq k$. Гипотезу H_k называют *основной*, а гипотезы $H_j : j \neq k$ — *альтернативными* или *конкурирующими*.

Рассмотрим теперь задачу проверки двух простых гипотез— основной H и альтернативной $\bar{H}$.

Принцип Неймана-Пирсона построения критериев для проверки двух простых гипотез основан на понятиях ошибок.

Критерием будем называть любую процедуру (правило) проверки гипотез. Критерии делятся на параметрические и непараметрические. Параметрические критерии строятся на основе параметров выборочной совокупности и представляют функции этих параметров, а непараметрические критерии — функции от выборочных значений. Параметрические критерии применяются только в том случае, когда генеральная совокупность нормальная, и при условии, что генеральные параметры сравниваемых групп равны между собой, т.е. $\alpha_1 = \alpha_2, \mu_1 = \mu_2$

Пусть X — пространство выборок X , и предположим, что X разбито на непересекающиеся множества S и D , т.е. $X = S \cup D, S \cap D = \emptyset$. Задаем некоторое достаточно малое число $\varepsilon > 0$, которое называется *уровнем значимости*. Допустим, что процедура T (обозначение) проверки гипотез H и $\bar{H}$ заключается в том, что если $X \in S$, то гипотезу H отвергаем, если же $X \in D$, то гипотезу H не отвергаем. Множество S — называется *критической областью*, а D — *доверительной областью*.

Любой критерий T в случае проверки двух простых гипотез H и $\bar{H}$ можно характеризовать числовой функцией $T(X) = \mathbf{P}(X \in S)$, которая называется *критической функцией критерия T* . Тогда, если $X \in S$, то $T(X) = 1$ и гипотеза, H отвергается, если же $X \in D$, $T(X) = 0$ и гипотеза H не отвергается.

При проверке гипотез возможны следующие ошибки.

Определение 1. *Ошибка I (первого) рода — это ошибка состоящая в том, что мы отвергаем гипотезу, которая верна.*

Вероятность ошибки I рода $\mathbf{P}(X \in S/H) = \alpha(T)$ — называется *значимостью критерия T* . Для любого критерия T должно выполняться $\alpha(T) < \varepsilon$.

Определение 2. *Ошибка II (второго) рода — это ошибка, состоящая в том, что мы принимаем гипотезу, которая не верна.*

Вероятность $1 - \mathbf{P}(X \in D/\bar{H}) = \delta(T)$ называется *мощностью критерия T* ,

где $\mathbf{P}(X \in D/\bar{H}) = \beta(T)$ — вероятность ошибки II рода.

Какой критерий наиболее предпочтительнее? Любой критерий характеризуется двумя величинами: $\varepsilon(T)$ и $\beta(T)$. Для выбора лучшего критерия рассмотрим несколько известных подходов.

1. Минимаксный подход.

Критерий T_1 не хуже, чем критерий T_2 , если

$$\max \{ \varepsilon(T_1), \beta(T_1) \} \leq \max \{ \varepsilon(T_2), \beta(T_2) \}.$$

Критерий T называется *минимаксным* критерием, если он не хуже всех других критериев в смысле минимаксного подхода.

2. Байесовский подход.

Предположим, что нам известно, что гипотеза H справедлива с вероятностью p , а $\bar{H}$ справедлива с вероятностью $q=1-p$. Критерий T_1 не хуже, чем критерий T_2 в смысле байесовского подхода, если

$$p \cdot \varepsilon(T_1) + q \cdot \beta(T_1) \leq p \cdot \varepsilon(T_2) + q \cdot \beta(T_2).$$

Критерий T называется *байесовском* критерием, если он не хуже всех остальных критериев в смысле байесовского подхода.

3. Подход Неймана – Пирсона.

Обозначим через $K_\varepsilon = \{T: \varepsilon(T) \leq \varepsilon\}$. Критерий $T \in K_\varepsilon$ назовем наиболее мощным критерием уровня ε , если вероятность ошибки II рода $\beta(T) \leq \beta(T')$ для любого критерия T' из этого класса, $T' \in K_\varepsilon$.

Рассмотрим гипотезу $H: X = (X_1, X_2, \dots, X_n)$ – выборка из семейства распределений $\{p_{\theta_1}(t)\}$ против альтернативы $\bar{H}: X = (X_1, X_2, \dots, X_n)$ – выборка из семейства распределений $\{p_{\theta_2}(t)\}$.

Обозначим функции правдоподобия

$$f(\theta_1) = \prod_{i=1}^n p_{\theta_1}(X_i); \quad f(\theta_2) = \prod_{i=1}^n p_{\theta_2}(X_i).$$

Лемма Неймана-Пирсона. Для любого $\varepsilon, \varepsilon \in [0,1]$ – наиболее мощный критерий уровня ε существует и совпадает с критерием отношения правдоподобия:

$$\left\{ \begin{array}{l} \text{Если } \frac{f(\theta_2)}{f(\theta_1)} < c, \text{ то не отвергаем гипотезу } H; \\ \text{Если } \frac{f(\theta_2)}{f(\theta_1)} > c, \text{ то отвергаем гипотезу } H; \\ \text{Если } \frac{f(\theta_2)}{f(\theta_1)} = c, \text{ то гипотезу } H \text{ с вероятностью } 1-p \text{ не отвергаем, с} \\ \text{вероятностью } p \text{ отвергаем.} \end{array} \right.$$

При этом c и p определяется из уравнения $\varepsilon(T) = \varepsilon$ или

$$\mathbf{P}\left(\frac{f(\theta_2)}{f(\theta_1)} < c / H\right) + p \cdot \mathbf{P}\left(\frac{f(\theta_2)}{f(\theta_1)} = c / H\right) = \varepsilon.$$

Пример. Проводится эксперимент по проверке телепатических способностей человека. Эксперимент заключается в угадывании заряженности или не заряженности n пробирок. Если человек угадывает, то записывается «1», если не угадывает, то «0». Мы получаем выборку $X = (X_1, X_2, \dots, X_n)$, объема n , состоящей из нулей и единиц. Пусть p – вероятность события {человек обладает телепатическими способностями}.

Проверим гипотезу $H: p = 1/2$ (т.е. эксперимент ничего не дал) против альтернативы $\bar{H}: p \neq 1/2$. Согласно гипотезе H считаем, что X – выборка из генеральной совокупности с распределением Бернулли $\mathbf{B}_p^1$ при $p = 1/2$. Пусть $\varepsilon > 0$ – уровень значимости, а $\varepsilon(T) = \mathbf{P}(T(X) = 1/H)$, и нам надо построить критерий T . Используя ЦПТ (см. 2.7.3.) мы получим

$$\eta = \sqrt{n} \cdot \frac{\bar{X} - \frac{1}{2}}{\sqrt{\frac{1}{4}}} = 2 \cdot \sqrt{n} \left(\bar{X} - \frac{1}{2} \right) \sim \Phi_{0,1},$$

поскольку $\mathbf{E}_\theta X = \frac{1}{2}, \mathbf{D}_\theta X = \frac{1}{4}, i = 1, 2, \dots, n$.

Предположим, что $\mathbf{P}(|\eta_1| < k_\varepsilon) = 1 - \varepsilon$ и из таблицы значений функции Φ_0 (см. Таблицу 1 Приложения) находим k_ε . Получаем следующий критерий:

$$T(X) = \begin{cases} 1, & \text{если } 2 \cdot \sqrt{n} \left| \bar{X} - \frac{1}{2} \right| \geq k_\varepsilon, \\ 0, & \text{если } 2 \cdot \sqrt{n} \left| \bar{X} - \frac{1}{2} \right| < k_\varepsilon. \end{cases}$$

Итак, если $X \in D \equiv \left\{ X = (X_1, X_2, \dots, X_n) : \left| \bar{X} - \frac{1}{2} \right| < k_\varepsilon \right\}$, то гипотезу H не отвергаем, если же $X \in S \equiv \left\{ X = (X_1, X_2, \dots, X_n) : \left| \bar{X} - \frac{1}{2} \right| \geq k_\varepsilon \right\}$, то гипотезу H отвергаем.

3.7.3. Примеры критериев для проверки гипотез

1. Критерий Колмогорова-Смирнова

Пусть $X = (X_1, X_2, \dots, X_n)$ выборка объема n из генеральной совокупности с функцией распределения $F(x)$. Рассмотрим гипотезу

$$H: F(x) = F_0(x),$$

где $F_0(x)$ — непрерывная функция распределения, которая известна (или мы задаем).

Введем следующую величину $K_n = \max_{-\infty < x < \infty} |F_n(x) - F_0(x)|$, где $F_n(x)$ — эмпирическая функция распределения выборки X .

Без доказательства (его можно найти в [2]) приведем формулировку следующей теоремы.

Теорема. При $n \rightarrow \infty$ имеет место следующее соотношение

$$\mathbf{P}(\sqrt{n} \cdot K_n < z) \rightarrow Q(z) \text{ при } n \rightarrow \infty,$$

$$\text{где } Q(z) = \sum_{k=1}^n (-1)^k \cdot e^{-2k^2 \cdot z^2}.$$

Заметим, что функцию $Q(z)$ называют распределением Колмогорова-Смирнова и ее значения табулированы (см. Таблицу 4 Приложения). Для заданного уровня значимости ε предлагается следующий критерий Колмогорова-Смирнова:

$$T(X) = \begin{cases} 1, & \text{если } \sqrt{n} \cdot K_n \geq z_0, \\ 0, & \text{если } \sqrt{n} \cdot K_n < z_0, \end{cases}$$

где z_0 находится из уравнения $Q(z_0) = 1 - \varepsilon$ с помощью таблицы значений функции $Q(z)$.

Пример 1. В продовольственном магазине сделаны контрольные замеры проданной колбасы, отклонения от истинного веса даны в таблице.

$X_i, \text{ гр.}$	1	2	3	4	5	6	7	8	9
n_i	16	15	17	18	18	11	6	29	20

С помощью критерия Колмогорова – Смирнова при уровне значимости $\varepsilon=0,05$ установим - согласуются ли данные выборки с законом равномерного распределения на отрезке $[0;10]$. В нашем примере $F_0(x)$ – функция равномерного распределения на отрезке $[0;10]$. Вычислим значения эмпирической функции распределения $F_n(x)$:

$$F_n(X_1)=F_n(1)=0; F_n(X_2)=F_n(2)=\frac{16}{150}=0,107; F_n(X_3)=F_n(3)=\frac{31}{150}=0,207;$$

$$F_n(X_4)=F_n(4)=\frac{48}{150}=0,32; F_n(X_5)=F_n(5)=\frac{66}{150}=0,44; F_n(X_6)=F_n(6)=\frac{77}{150}=0,51;$$

$$F_n(X_7)=F_n(7)=\frac{95}{150}=0,63; F_n(X_8)=F_n(8)=\frac{101}{150}=0,67; F_n(X_9)=F_n(9)=\frac{130}{150}=0,87;$$

$$F_n(X_{10})=F_n(10)=\frac{150}{150}=1.$$

Вычислим теперь значения $F_0(x)$ используя

$$F_0(x)=\begin{cases} 0, & \text{если } x \leq 0, \\ \frac{x}{10}, & \text{если } x \in [0,10], \\ 1, & \text{если } x > 10. \end{cases}$$

получим

$$F_0(X_1)=0,1; F_0(X_2)=0,2; F_0(X_3)=0,3; F_0(X_4)=0,4; F_0(X_5)=0,5;$$

$$F_0(X_6)=0,6; F_0(X_7)=0,7; F_0(X_8)=0,8; F_0(X_9)=0,9; F_0(X_{10})=1.$$

Вычислим разности $|F_n(X_i)-F_0(X_i)|$ и $|F_n(X_i)-F_0(X_{i-1})|$ при $i = 1, \dots, 10$. Из них выберем наибольшую и получим

$$K_n = \max |F_n(x) - F_0(x)| = 0,13.$$

(Проверьте самостоятельно!). Следовательно, $\sqrt{n} \cdot K_n = \sqrt{150 \cdot 0,13} = 1,59$. Величина z_0 находится из таблицы распределения Колмогорова-Смирнова, при $\varepsilon=0,05$, $z_0 = 1,358$. Сравнение нам дает $\sqrt{n} \cdot K_n > z_0$, что отвергает основную гипотезу, значит распределение неравномерное.

2. Критерий Пирсона (хи-квадрат)

Пусть $X = (X_1, X_2, \dots, X_n)$ — выборка объема, n из генеральной совокупности с функцией распределения $F(x)$.

Рассмотрим гипотезу $H: F(x) = F_0(x)$, где $F_0(x)$ — функция распределения, которая точно известна (или мы задаем).

Множество возможных значений случайной величины ξ с функцией распределения $F(x)$ разбиваем на k непересекающихся подмножеств $\Delta_1, \Delta_2, \dots, \Delta_k$, где $\Delta_i = [a_i, b_i)$. Обозначим через n_i — число выборочных значений X_i , попавших в Δ_i , и

$$p_i = P(\xi \in \Delta_i / H) = F_0(b_i) - F_0(a_i).$$

Для проверки гипотезы H предлагается следующая статистика:

$$\chi^2 = \sum_{i=1}^n \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^n \frac{n_i^2}{np_i} - n,$$

поскольку $\sum_{i=1}^n n_i = n$, $\sum_{i=1}^n p_i = 1$.

Воспользуемся тем, что $\chi^2 \sim \mathbf{H}_{k-1}$ (доказательство этого факта можно найти в [2]).

Как мы указывали ранее, имеется таблица значений функции $\mathbf{H}_n$ (см. Таблицу 3 Приложения). Из этой таблицы при заданном ε находим χ_ε^2 .

Для проверки гипотезы предлагается следующий критерий Пирсона:

$$T(X) = \begin{cases} 1, & \text{если } \chi^2 \geq \chi_\varepsilon^2, \\ 0, & \text{если } \chi^2 < \chi_\varepsilon^2. \end{cases}$$

Пример 2. На экзамене по теории вероятностей экзаменатор задает только один вопрос по одному из 4-х частей курса. Из 100 студентов 26 получили вопрос по 1-й части, 32 студента — по 2-й, 17 студентов по 3-й и остальные 25 — по 4-й. Можно ли по этим результатам принять гипотезу о том, что для пришедшего на экзамен имеется одинаковая вероятность получить вопрос по любой из 4-х частей.

Предположим, что $\varepsilon = 0,05$. Пусть $X_1, \dots, X_{100}$ — номера вопросов, $\Delta_1, \Delta_2, \Delta_3, \Delta_4$ — интервалы «вопросов» по соответствующей части курса.

Δ_i	Δ_1	Δ_2	Δ_3	Δ_4
------------	------------	------------	------------	------------

n_i	2	32	17	25
-------	---	----	----	----

Проверим гипотезу $H: F(x)=F_0(x)$, где по условиям задачи таблица распределения $F_0(x)$ имеет вид:

1	2	3	4
0,25	0,25	0,25	0,25

Вычислим

$$\chi^2 = \sum_{i=1}^4 \frac{n_i^2}{25} - 100 = \frac{1}{25} \sum_{i=1}^4 n_i^2 - 100 = 4,56.$$

Число степеней свободы $\nu = \kappa - 1 = 3$, $\varepsilon = 0,05$, из таблицы распределения χ^2 находим χ_{ε}^2 , где $\chi_{\varepsilon}^2 = 7,81$. Сравним $\chi^2 < \chi_{\varepsilon}^2$, гипотеза не отвергается.

3. Критерий Стьюдента

Продemonстрируем этот критерий для проверки гипотез однородности для нормальных совокупностей. Пусть $X_1, X_2, \dots, X_n$ выборка из нормального распределения с параметрами a_1 и σ_1^2 ; $Y_1, Y_2, \dots, Y_n$ выборка из нормального распределения с параметрами a_2 и σ_2^2 ;

Принадлежат ли эти выборки одной и той же генеральной совокупности?

Рассмотрим варианты с двусторонней альтернативой.

1) . Предположим, что $\sigma_1 = \sigma_2 = \sigma$ известно.

Проверяем гипотезу $H: a_1 = a_2$ против альтернативы $\bar{H}: a_1 \neq a_2$ Рассмотрим статистику $\xi = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$. Эта случайная величина подчиняется стандартному нормальному распределению.

Если будет выполнено условие $|\xi| > t_\varepsilon$, то основная гипотеза отвергается, где t_ε находится из таблицы значений функции $\Phi_0(t_\varepsilon) = \frac{1-\varepsilon}{2}$.

2). Предположим, что $\sigma_1 = \sigma_2 = \sigma$ неизвестно.

Рассмотрим статистику

$$t = \frac{\bar{Y} - \bar{X}}{\sqrt{n_1 S_x^2 + n_2 S_y^2}} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2} (n_1 + n_2 - 2)}$$

где

$$S_x^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2; \quad \bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i; \quad S_y^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2; \quad \bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i.$$

Эта статистика подчиняется распределению Стьюдента с числом степеней свободы $n_1 + n_2 - 2$. Из таблицы распределения Стьюдента находим t_ε . Если $|t| > t_\varepsilon$, то основная гипотеза отвергается.

Теперь рассмотрим варианты с односторонней альтернативой.

1) Предположим, что $\sigma_1 = \sigma_2 = \sigma$ известно.

А) Проверяем гипотезу $H: a_1 = a_2$ против альтернативы $\bar{H}: a_1 > a_2$.

Вычислим статистику ξ (см. выше). Если $\xi < -t_\varepsilon$, то основная гипотеза отвергается, где t_ε находится из таблицы значений функции $\Phi_0(t_\varepsilon) = \frac{1-2\varepsilon}{2}$.

Б) Проверяем гипотезу $H: a_1 = a_2$ против альтернативы $\bar{H}: a_1 < a_2$.

Также вычислим ξ . Если $\xi > t_\varepsilon$, то основная гипотеза отвергается, где t_ε находится из таблицы значений функции $\Phi_0(t_\varepsilon) = \frac{1-2\varepsilon}{2}$.

2) Предположим, что $\sigma_1 = \sigma_2 = \sigma$ неизвестно.

А) Проверяем гипотезу $H: a_1 = a_2$ против альтернативы $\bar{H}: a_1 > a_2$

Вычислим также статистику t (см. выше). Если $t > t_\varepsilon$, то основная гипотеза отвергается, где t_ε находится из таблицы распределения Стьюдента.

Б) Проверяем гипотезу $H: a_1 = a_2$ против альтернативы $\bar{H}: a_1 < a_2$.

Вычислим статистику t . Если $t < -t_\varepsilon$, то основная гипотеза отвергается.

Пример 3. В двух фирмах выпускающих детское питание производилась оценка качества продукции. В фирме А, где проверялось 30 единиц продукции, средняя сумма

баллов оказалась равной 52. В фирме B , где проверялось 36 единиц продукции их средняя сумма баллов оказалась равной 47. Среднеквадратическое отклонение сумм баллов, вычисленная для нескольких единиц продукции $\sigma = 12$. Лучшее ли питание выпускается фирмой A , чем фирмой B ?

Проверяем гипотезу $H: a_1 = a_2$ против альтернативы $\bar{H}: a_1 > a_2$.

Здесь a_1 и a_2 средние баллы оценки качества продуктов фирм A и B , соответственно. Пусть $\varepsilon=0.05$. По условиям задачи $\sigma_1 = \sigma_2 = 12$ – известно, а также $\bar{Y} = 47$; $\bar{X} = 52$; $n_1=30$; $n_2=36$. Вычислим

$$\xi = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{47 - 52}{\sqrt{\frac{(12)^2}{30} + \frac{(12)^2}{36}}} = \frac{-5}{\sqrt{\frac{2}{5} + \frac{2}{6}}} = \frac{-5}{\sqrt{\frac{12+10}{30}}} = \frac{-5}{\sqrt{\frac{11}{15}}} = -1,686$$

Из таблицы значений функции $\Phi_0(t_\varepsilon) = \frac{1-2\varepsilon}{2}$ при заданном $\varepsilon = 0,05$ находим $t_\varepsilon = 1,65$.

Сравнение дает $\xi < -t_\varepsilon$, следовательно основная гипотеза отвергается. Действительно, фирмой A производится лучшее детское питание.

Замечание. В случае $\sigma_1 \neq \sigma_2$ задача известна как проблема Беренса – Фишера. Здесь мы этот случай рассматривать не будем..

Задачи к 3.7

1. Имеются две выборки значений (в усл.ед.) объёмов 125 и 80 показателя качества однотипной продукции, изготовленной двумя фирмами:

X_i	14	17	20	23	26	29	32	35	38	41
n_i	2	4	10	15	20	27	18	16	8	5

Y_i	16	20	24	28	32	36	40	44
n_i	3	9	12	17	16	13	7	3

С помощью критерия Стьюдента выяснить, можно ли на уровне значимости 0,05 считать, что рассматриваемый показатель качества продукции двух фирм описывается одной и той же функцией распределения (т.е. выборки извлечены из одной генеральной совокупности).

2. Фирма поставляет радары для измерения скорости движения автомобилей. Для закупки большой партии проведены испытания приборов, изготовленных на заводе A и на заводе B . Измерения проводили на одной и той же машине и на одной и той же дороге. Определены величины отклонений между показаниями спидометра автомобиля и радара:

Завод A

Отклонение, км/ч	Δx_i	-0,7	-0,3	-0,1	0,5	0,8	0,9	1	1,2	1,3
Число измерений	n_i	5	4	2	6	3	1	3	1	1

Завод B

Отклонение, км/ч	Δy_i	-0,6	-0,1	0,4	0,7	1,0	1,4
Число измерений	m_i	4	5	3	2	2	1

Полагая показания спидометра автомобиля эталоном, проверить гипотезу об одинаковой точности измерений, проводимых радаром завода A и B , при уровне значимости 0,1.

3. Результаты исследования числа покупателей в универсаме в зависимости от времени работы приведены ниже:

Часы работы	9-10	10-11	11-12	12-13
Число покупателей	41	82	117	72

Можно ли утверждать при уровне значимости $\alpha=0,05$, что случайная величина X – число покупателей – подчинена нормальному закону?

4. При принятии на работу фирма предлагает 4 теста. Результаты решения этих тестов десятью претендентами приведены ниже:

Число верно	0	1	2	3	4
-------------	---	---	---	---	---

решённых тестов					
Число участников	1	2	2	3	2

Проверить гипотезу о биномиальном распределении случайной величины X – числа успешно решённых тестов – при $\alpha=0,05$.

3.8. Линейная регрессия между двумя случайными величинами

Рассмотрим две величины ξ и η , которые как-то связаны. Пусть имеется выборка, состоящая из n пар наблюдаемых значений ξ и η : $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. Наша цель — построить по выборке функцию $Y = a + kX$, которая бы позволяла для любого значения аргумента X «угадывать» значение случайной величины η , соответствующее значению « $\{\xi = X\}$ ».

Чтобы точно поставить задачу, нужно расшифровать понятие «угадать». Мы приведем две строгие постановки, которые, впрочем, приводят к общему решению.

3.8.1. Задача о наименьших квадратах

Нужно построить линейную функцию $Y = a^* + k^*X$, где коэффициенты a^* , k^* являются функциями элементов выборки $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, такую, что

$$g(a^*, k^*) = \min_{a, k} g(a, k),$$

$$\text{где } g(a, k) = \sum_{i=1}^n (Y_i - (a + k \cdot X_i))^2.$$

Прежде чем решать поставленную задачу, напомним (см. 3.1.3):

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2; \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i; \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad S_y^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

$$r(X, Y) = \frac{\frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{S_x \cdot S_y}.$$

Для нахождения точки минимума функции $g(a, k)$ составим систему уравнений:

$$\begin{cases} \frac{\partial g(a,k)}{\partial a} = 0, \\ \frac{\partial g(a,k)}{\partial k} = 0. \end{cases}$$

Решая эту систему, без труда находим:

$$k^* = \frac{\frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_x^2}, \quad a^* = \bar{Y} - \frac{\frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_x^2}.$$

Таким образом, поставленная задача о наименьших квадратах решена; получена искомая линейная функция:

$$Y = \bar{Y} + \frac{s_y}{s_x} r(X, Y) (X - \bar{X}),$$

или

$$Y = \bar{Y} + \frac{\frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_x^2} (X - \bar{X}).$$

Обратимся теперь к другой постановке проблемы.

3.8.2. Задача о линейной регрессии

Предположим, что в рассматриваемой выборке $X_1, X_2, \dots, X_n$ — произвольные вещественные числа, а числа $Y_1, Y_2, \dots, Y_n$ — значения нормально распределенных случайных величин, причем Y_i имеет нормальное распределение с параметрами $\alpha + \beta(X_i - \bar{X})$ и σ^2 , где α, β, σ^2 не известны. Напомним, что плотность нормального распределения с указанными параметрами имеет вид:

$$p_i(\alpha, \beta, \sigma^2; t) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp \left\{ -\frac{(t - (\alpha + \beta(X_i - \bar{X})))^2}{2\sigma^2} \right\}.$$

Обозначим η_i - случайную величину с такой плотностью. Тогда математическое ожидание случайной величины η_i имеет вид :

$$E\eta_i = \alpha + \beta(X_i - \bar{X}),$$

поэтому искомую линейную функцию следует искать в виде:

$$Y = \alpha^* + \beta^* (X_i - \bar{X}),$$

где α^* , β^* — какие-то оценки для параметров α , β . Поскольку оценки максимального правдоподобия обладают многими хорошими свойствами, остановимся на них.

Напомним, что для построения оценок максимального правдоподобия следует рассмотреть функцию правдоподобия (см. 3.4.2)

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \sum_{i=1}^n \ln \mathbf{p}_i(\alpha, \beta, \sigma^2; Y_i) \\ &= -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \sum_{i=1}^n \frac{(Y_i - (\alpha + \beta(X_i - \bar{X})))^2}{2\sigma^2}. \end{aligned}$$

Тогда оценки α^* , β^* , σ^2 являются оценками максимального правдоподобия неизвестных параметров α , β , σ^2 , если

$$L(\alpha^*, \beta^*, (\sigma^2)^*) = \max_{\alpha, \beta, \sigma^2} L(\alpha, \beta, \sigma^2).$$

Для нахождения точки максимума следует составить систему уравнений

$$\begin{cases} \frac{\partial L(\alpha, \beta, \sigma^2)}{\partial \alpha} = 0, \\ \frac{\partial L(\alpha, \beta, \sigma^2)}{\partial \beta} = 0, \\ \frac{\partial L(\alpha, \beta, \sigma^2)}{\partial \sigma^2} = 0. \end{cases}$$

Решая эту систему, без труда находим (проверьте самостоятельно):

$$\alpha^* = \bar{Y}, \quad \beta^* = \frac{\frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_x^2}, \quad (\sigma^2)^* = \frac{1}{n} \cdot \sum_{i=1}^n (Y_i - (\alpha^* + \beta^*(X_i - \bar{X})))^2.$$

Таким образом, искомая линейная функция найдена:

$$Y = \bar{Y} + \frac{s_y}{s_x} r(X, Y) (X - \bar{X}),$$

или

$$Y = \bar{Y} + \frac{\frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_x^2} (X - \bar{X}).$$

Заметим, что ответы в задачах 1, 2 совпадают.

Пример. Пусть η — стоимость минимального набора продуктов питания городе Новосибирск за месяц (в рублях, по данным Новосибирсстата), а ξ — средний курс доллара США к рублю (за месяц). Ниже приведена таблица значений этих случайных величин по месяцам 2014 года.

	1	2	3	4	5	6	7	8	9	10	11	12
X_i	35,24	36,05	35,69	35,70	34,74	33,63	35,73	36,93	39,49	43,39	49,32	56,26
Y_i	3166	3214	3324	3405	3477	3483	3459	3342	3266	3285	3367	3590

Вычислим выборочные характеристики (например, в Excel):

$$\bar{X}=39,3475; S_x=1565,403; \bar{Y}=3364,833; S_y=118,4433; r(X,Y)=0,353039.$$

Используя эти оценки получаем следующее уравнение

$$Y=\bar{Y}+\frac{s_y}{s_x}r(X,Y)(X-\bar{X})=3364,833+\frac{118,4433}{1565,403}\cdot 0,353039(X-39,3475),$$

т.е. имеем следующее уравнение регрессии:

$$Y=0,026712\cdot X+3363,762.$$

С помощью этого уравнения можно предсказать среднюю стоимость минимального набора продуктов по курсу доллара США на соответствующий момент (месяц).

Задачи к 3.8

1. Туристическая компания предлагает места в гостиницах приморского курорта. Менеджера компании интересует, насколько возрастает привлекательность гостиницы в зависимости от её расстояния до пляжа. С этой целью по 10 гостиницам города была выяснена среднегодовая наполняемость номеров и расстояние в километрах до пляжа:

Расстояние, км	0,1	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
Наполняемость, %	92	95	96	90	89	90	83	80	78	75

Постройте график исходных данных и определите по нему характер зависимости. Постройте уравнение регрессии и дайте интерпретацию полученных результатов.

2. Компанию по прокату автомобилей интересует зависимость между пробегом автомобилей X и стоимостью ежемесячного обслуживания Y . Для выяснения характера этой связи было отобрано 15 автомобилей.

X	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y	13	16	15	20	19	21	26	24	30	32	30	35	34	40	39

Постройте график исходных данных и определите по нему характер зависимости. Постройте уравнение регрессии и дайте интерпретацию полученных результатов.

3. Некоторая компания недавно провела рекламную кампанию в магазинах с демонстрацией антисептических качеств своего нового моющего средства. Через 10 недель компания решила проанализировать эффективность этого вида рекламы, сопоставив еженедельные объёмы продаж (в млн.руб) с расходами (в млн.руб) на рекламу:

Объём продаж,	72	76	78	70	68	80	82	65	62	90
Расходы на рекламу,	5	8	6	5	3	9	12	4	3	10

Постройте график исходных данных и определите по нему характер зависимости. Постройте уравнение регрессии и дайте интерпретацию полученных результатов.

3.9. Дополнительные задачи к Главе 3

1. Пусть $X_1, X_2, \dots, X_n$ выборка из непрерывного распределения с плотностью распределения $p(x, \theta) = \begin{cases} \left(\frac{1}{\theta}\right) \cdot x^{(1-\theta)/\theta} & \text{при } 0 < x < 1, \\ 0, & \text{иначе} \end{cases}$ где $\theta > 0$. Найти оценку θ^* параметра θ методом моментов.

2. Пусть $X_1, X_2, \dots, X_n$ выборка из непрерывного распределения с плотностью распределения $p(t, \theta) = \begin{cases} \frac{\theta}{\sqrt{2t}} \cdot e^{-\theta\sqrt{t}}, & \text{при } t > 0, \\ 0, & \text{иначе.} \end{cases}$ где $\theta > 0$. Найти оценку максимального правдоподобия для параметра θ .

3. Пусть X случайная величина, имеющая биномиальное распределение с параметрами n и p , и пусть $p^* = \frac{x}{n}$ несмещенная оценка для параметра p . Найти несмещенную оценку для $p(1-p)$.

4. Пусть $X_1, X_2, \dots, X_n$ выборка из непрерывного распределения с плотностью распределения $p(t) = \begin{cases} \lambda \cdot e^{-\lambda t} & \text{при } t > 0, \\ 0, & \text{иначе} \end{cases}$. Для несмещенной оценки λ^* параметра λ определить нижнюю границу дисперсии по неравенству Рао-Крамера.

5. Пусть $X_1, X_2, \dots, X_n$ выборка из нормального распределения со средним μ и дисперсией σ^2 . Рассматривается основная гипотеза $H: \mu = 50$ против альтернативы $\bar{H}: \mu > 50$. Пусть уровень значимости 0,25. Если $\bar{X} = 52,53$ и $\frac{1}{8} \cdot \sum_{i=1}^9 (X_i - \bar{X})^2 = 3,3^2$. Вычислить t -статистику Стьюдента.

П Р И Л О Ж Е Н И Я

Таблица 1. Нормальное распределение

Значения функции Лапласа $\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^x e^{-\frac{t^2}{2}} dt$.

x	сотые доли									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0200	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1874
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3437	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,44.63	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4764	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4985	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4988	0,4989	0,4989	0,4990	0,4990

Т а б л и ц а 2. Распределение Стьюдента $G_n(t)$

Значения функции t_ε , которая определяется равенством

$$G_n(t_\varepsilon) = P(|t| < t_\varepsilon) = 1 - \varepsilon$$

n / ε	0,001	0,002	0,005	0,01	0,02	0, 05	0,1	0,2
1	636,619	318,306	127,656	63,657	31,821	12,706	6,314	3,078
2	31,599	22,327	14,089	9,925	6, 965	4,303	2,920	1,886
3	12,924	10,214	7,458	5,841	4, 541	3,182	2,353	1,638
4	8,610	7,173	5,597	4,604	3, 747	2,776	2,132	1,533
5	6,863	5,893	4,773	4,032	3,365	2,571	2,015	1,476
6	5,958	5,207	4,316	3,707	3,143	2, 447	1,943	1,440
7	5,4079	4,785	4,2293	3,499	2,998	2,365	1,895	1,415
8	5,0413	4,5008	3,832	3,355	2,896	2, 306	1,860	1,397
9	4,78	4,2968	3,6897	3,250	2,821	2,262	1,833	1,383
10	4,5869	4,1437	3,5814	3,169	2,764	2,228	1,812	1,372
11	4,437	4,024	3,496	3,106	2,718	2,201	1,796	1,363
12	4,178	3,929	3,4284	3,055	2,681	2,179	1,782	1,356
13	4,22	3,852	3,3725	3,012	2,650	2,160	1,771	1,350
14	4,14	3,787	3,3257	2,977	2,624	2,145	1,761	1,345
15	4,072	3,732	3,286	2,947	2,602	2,131	1,753	1,341
16	4,015	3,686	3,252	2,921	2,583	2,120	1,746	1,337
17	3,965	3,6458	3,2224	2,898	2,567	2,110	1,740	1,333
18	3,9216	3,6105	3,1966	2,878	2,552	2,101	1,734	1,330
19	3,8834	3,5794	3,1737	2,861	2,539	2,093	1,729	1,328
20	3,8495	3,5518	3,1534	2,845	2,528	2,086	1,725	1,325
21	3,819	3,527	3,135	2,731	2,518	2, 080	1,721	1,323
22	3,7921	3,505	3,1188	2,819	2,508	2, 074	1,717	1,321
23	3,7676	3,485	3,104	2,807	2,500	2,069	1,714	1,319
24	3,7454	3,4668	3,0905	2,797	2,492	2,064	1,711	1,318
25	3,7251	3,4502	3,0782	2,787	2,485	2,060	1,708	1,316
26	3,706	3,436	3,066	2,779	2,479	2,056	1,706	1,315
27	3,6896	3,421	3,0565	2,771	2,473	2,052	1,703	1,314
28	3,6739	3,4082	3,0469	2,763	2,767	2,048	1,701	1,313
29	3,8494	3,3962	3,036	2,756	2,462	2,045	1,699	1,311
30	3,646	3,3852	3,0298	2,750	2,457	2,042	1,697	1,310
40	3,551	3,3069	2,9712	2,705	2,423	2,021	1,684	1,303
60	3,4602	3,2317	2,9146	2,660	2,390	2,000	1,671	1,296
100	3,3905	3,1737	2,8707	2,6259	2,3642	1,984	1,6602	1,2901

Т а б л и ц а 3. Распределение хи-квадрат $\mathbf{H}_n(x)$

Значения функции χ^2_ε , которая определяется равенством

$$\mathbf{H}_n(\chi^2_\varepsilon) = \mathbf{P}(\chi^2 < \chi^2_\varepsilon) = 1 - \varepsilon.$$

$n \quad \varepsilon$	0.005	0.01	0.025	0.05	0.10	0.90	0.95	0.975	0.99	0.995
1	7,88	6,63	5,02	3,84	2,71	0,02	0,024	0,039	0,031	0,044
2	10,60	9,21	7,38	5,99	4,61	0,21	0,103	0,051	0,02	0,01
3	12,84	11,34	9,35	7,81	6,25	0,58	0,584	0,352	0,216	0,115
4	14,86	13,28	11,14	9,49	7,78	1,06	0,711	0,484	0,297	0,207
5	16,75	15,09	12,84	11,07	9,24	1,61	1,145	0,831	0,554	0,412
6	18,55	16,81	14,45	12,59	10,64	2,20	1,635	1,237	0,872	0,676
7	20,28	18,48	16,01	14,07	12,02	2,83	2,167	1,690	1,239	0,989
8	21,96	20,09	17,53	15,51	13,36	3,49	2,733	2,180	1,646	1,344
9	23, 59	21,67	19,02	16,92	14,68	4,17	3,325	2,700	2,088	1,735
10	25,19	23,21	20,48	18,31	15,99	4,87	3,940	3,247	2,558	2,156
11	26,76	24, 73	21,92	19,68	17,28	5,58	4,575	3,816	3,053	2,603
12	28,30	26,22	23, 34	21,03	18,55	6,30	5,226	4,404	3,571	3,074
13	29,82	27,69	24,74	22,36	19,81	7,04	5,892	5,009	4,107	3,565
14	31,32	29,14	26,12	23,68	21,06	7,79	6,571	5,629	4,660	4,075
15	32, 80	30,58	27,19	25, 00	22,31	8,55	7,261	6,262	5,229	4,601
16	34,27	32,00	28,85	26,30	23,54	9,31	7,962	6,908	5,812	5,142
17	35,72	33,41	30,19	27,59	24,77	10,09	8,672	7,564	6,408	5,697
18	37,16	34,81	31,53	28,87	25, 99	10,87	9,390	8,231	7,015	6,265
19	38,58	36,19	32,85	30,14	27,20	11,65	10,12	8,907	7,633	6,844
20	40,00	37,57	34,17	31,41	28,41	12,44	10,85	9,591	8,260	7,434
21	41,40	38,93	35,48	32,67	29,62	13,24	11,59	10,28	8,897	8,034
22	42,80	40,29	36,78	33,92	30,81	14,04	12,34	10,98	9,542	8,643
23	44,18	41,64	38,08	35,17	32,01	14,85	13,09	11,69	10,20	9,260
24	45,56	42,98	39, 36	36,42	33,20	15,66	13,85	12,40	10,86	9,886
25	46,93	44,31	40,65	37,65	34,38	16,47	14,61	13,12	11,52	10,52
26	48,29	45,64	41,92	38,88	35, 56	17,29	15,38	13,84	12,20	11,16
27	49,64	46,96	43,19	40,11	36,74	18,11	16,15	14,57	12,88	11,81
28	50,99	48,28	44,46	41,34	37,92	18,94	16,93	15,31	13,57	12,46
29	52,34	49,59	45,72	42, 56	39,09	19,77	17,71	16,05	14,26	13,12
30	53,67	50,89	46,98	43, 77	40,26	20,60	18,50	16,80	14,95	13,79
40	66,67	63,69	59,34	55,76	51,81	29,05	26,51	24,43	22,16	20,71
60	91,95	88,38	83,30	79,08	74,40	46,46	43,19	40,48	37,49	35,54
100	140,2	135,8	129,6	124,4	118,5	82,36	77,93	74,22	70,07	67,33

Т а б л и ц а 4. Распределение Колмогорова-Смирнова $Q(x)$.

Значения функции z_ε , которая определяется равенством

$$Q(z) = \mathbf{P}(z < z_0) = 1 - \varepsilon.$$

ε	0,10	0,05	0,025	0,02	0,01	0,005
z_0	1,224	1,358	1,480	1,520	1,627	1,730

СПИСОК ЛИТЕРАТУРЫ

1. **Боровков А.А.** Теория вероятностей: Учебное пособие для вузов. М., 1999. 472 с.
2. **Боровков А.А.** Математическая статистика. Учебник. СПб., 2010. 704 с.
3. **Гнеденко Б.В.** Курс теории вероятностей. М., 2005. 448 с.
4. **Крамер Г.** Математические методы статистики. М., 1975. 648 с.
5. **Гурский Е.И.** Теория вероятностей с элементами математической статистики. М., 1971. 382 с.
6. **Гмурман В.Е.** Руководство к решению задач по теории вероятностей и математической статистике. М., 2006. 476 с.
7. **Вентцель Е.С., Овчаров Л.А.** Сборник задач по теории вероятностей. М., 1969. 368 с.
8. **Емельянов Г.В., Скитович В.П.** Сборник задач по теории вероятностей и математической статистике. Л., 1967. 332 с.
9. **Зубков А.М., Севастьянов Б.А., Чистяков В.П.** Сборник задач по теории вероятностей: Пособие для вузов. М., 1989. 320 с.
10. **Свешников А.А. и др.** Сборник задач по теории вероятностей, математической статистике и теории случайных функций. СПб., 2008. 446 с.
11. **Бородихин В.М., Джафаров К.А., Путинцева А.П.** Теория вероятностей и математическая статистика. Учебное пособие. Н., 1997. 154 с.
12. **Кремер Н.Ш.** Теория вероятностей и математическая статистика. Учебник для вузов. М., 2007. 543 с.
13. Сборник задач по высшей математике для экономистов. Учебное пособие /Под ред. **В.И.Ермакова**. М., 2002.-575 с.
14. **Калинина В.Н.** Теория вероятностей и математическая статистика. Компьютерно-ориентированный курс. Учебное пособие. М., 2008. 336 с.
15. **Климов Г.П.** Теория вероятностей и математическая статистика. М., 2011. 368 с.
16. Теория статистики с основами теории вероятностей. Учебное пособие для вузов /Под ред. **И.И.Елисеевой**. М., 2001. 446 с.