Финансовый университет
при правительстве Российской Федерации

Шамраева
Виктория Викторовна

кандидат физико-математических наук,
доцент кафедры
математики и анализа данных

Теория вероятностей и математическая статистика

НАПРАВЛЕНИЕ ПОДГОТОВКИ: «Прикладная математика - ПМ»

КВАЛИФИКАЦИЯ (СТЕПЕНЬ): бакалавр

Методы построения точечных оценок

Раздел 1. Оценки параметров

Для нахождения оценок неизвестных параметров используют различные методы.

Наиболее распространенными являются

метод моментов,

метод максимального правдоподобия (ММП) и

метод наименьших квадратов (МНК).

Раздел 1. Оценки параметров

Метод моментов

(разработан П.Л. Чебышёвым в 1887 г. в процессе доказательства центральной предельной теоремы и далее обобщенный и развитый К. Пирсоном в 1894 г.)

Раздел 1. Оценки параметров

Суть метода моментов: выразить числовые параметры теоретического распределения через моменты распределения, оцененные по выборки.

Число моментов должно соответствовать числу неизвестных параметров распределения (чаще всего используют первые два момента).

После вычисления приравниваем теоретические и выборочные моменты друг к другу и выражаем оценки параметров.

Раздел 1. Оценки параметров

Данный метод прост в реализации, дает неплохие оценки и удобен для отработки навыков.

Про **свойства оценок**:

состоятельность оценок выполняется при непрерывной зависимости от параметра, асимптотическая эффективность оценок, полученных по методу максимального правдоподобия (ММП) всегда лучше чем у метода моментов (ММ), оценки по ММ чаще всего смещенные (требуется проверка).

Раздел 1. Оценки параметров

Имеется случайная выборка $X_1, X_2, \dots, X_n$ из генеральной совокупности X, распределение которой зависит от вектора параметров $\vec{\theta} = (\theta_1, \dots, \theta_k)$.

Требуется найти оценки $\hat{\theta}_i$ параметров θ_i, $i = 1, 2, \dots, k$.

Рассматриваются выборочные начальные моменты $\hat{\nu}_k$ (или центральные моменты $\hat{\mu}_i$), $i = 1, \dots, k$. Величины ν_i являются функциями неизвестного вектора параметров $\vec{\theta}$, т.е.

$$\nu_i = \nu_i(\vec{\theta}).$$

Замечание. В текущем разделе нижние индексы в обозначениях оценок параметров указывают не объемы выборки, а номера компонент вектора параметров.

Раздел 1. Оценки параметров

Метод моментов заключается в том, что в качестве точечной оценки вектора параметров $\vec{\theta}$ берут статистику, выражение для которой получается в результате решения системы уравнений

$$\nu_i(\vec{\theta}) = \hat{\nu}_k, i = 1, 2, \ldots, k.$$

Очевидным **достоинством** метода моментов является его простота, однако качество оценок, полученных с помощью этого метода, не всегда бывает высоким, особенно при небольших объемах выборки.

Раздел 1. Оценки параметров

Пример. Пусть $X_1, X_2, \dots, X_n$ — случайная выборка из генеральной совокупности с **показательным законом распределения**. Требуется, используя **метод моментов**, найти оценку параметра распределения λ.

Решение.

Раздел 1. Оценки параметров

Пример. Известно, что случайная величина X имеет **равномерное распределение**, но отрезок $[a; b]$, на котором она распределена, не известен. Требуется по случайной выборке $X_1, X_2, \dots, X_n$ объема n найти оценки a и b величин a и b **методом моментов**.

Решение.

Раздел 1. Оценки параметров

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

Следствие.

Для всех основных типов распределений, рассмотренных ранее (**биномиальное**, **геометрическое**, **пуассоновское**, **гипергеометрическое**, **показательное**, **равномерное**, **нормальное** и **логнормальное**), существуют начальные моменты любого порядка.

Следовательно, для этих распределений оценки метода моментов являются **состоятельными**.

Раздел 1. Оценки параметров

Оптимальные статистические оценки

Раздел 1. Оценки параметров

Среднеквадратичная ошибка статистической оценки $\hat{\theta}$ задаётся формулой

$$\Delta = E[(\hat{\theta} - \theta)^2]$$

(MSE – Mean Squared Error).

Систематической ошибкой (или **смещением**) статистической оценки $\hat{\theta}$ называется разность

$$b(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

Раздел 1. Оценки параметров

Пример. Пусть $\hat{\theta} = T(X_1, \ldots, X_n)$ оценка параметра θ, а $b = E(\hat{\theta}) - \theta$ – смещение. Тогда

$$\Delta = Var(\hat{\theta}) + b^2,$$

где $\Delta = E[(\hat{\theta} - \theta)^2]$ – среднеквадратичная ошибка оценки (MSE – Mean Squared Error) .

Решение.

Среднеквадратичная ошибка статистической оценки $\hat{\theta}$ задаётся формулой $\Delta = E[(\hat{\theta} - \theta)^2]$ (MSE – Mean Squared Error).

Систематической ошибкой (или **смещением**) статистической оценки $\hat{\theta}$ называется разность $b(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Раздел 1. Оценки параметров

Статистическая оценка $\hat{\theta}$ называется **оптимальной** в некотором классе статистических оценок $\hat{\Theta}$, если в этом классе она имеет **наименьшую** среднюю квадратическую ошибку.

Среднеквадратичная ошибка статистической оценки $\hat{\theta}$ задаётся формулой $\Delta = E[(\hat{\theta} - \theta)^2]$ (MSE – Mean Squared Error).

Систематической ошибкой (или **смещением**) статистической оценки $\hat{\theta}$ называется разность $b(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Алгоритм
метода максимального правдоподобия

(разработан Р. Фишером в 1912—1922 гг.)

Раздел 1. Оценки параметров

Задача.

Имеется выборка $X_1, X_2, \dots, X_n$ значений случайной величины X, распределение которой зависит от параметра θ, где θ — число или вектор.

Требуется найти оценку $\hat{\theta}$ параметра θ.

Раздел 1. Оценки параметров

Для **дискретной случайной величины** **функция правдоподобия** определяется по формуле

$$L(x_1, x_2, \dots, x_n;\ \theta) = p(x_1;\ \theta) \cdot p(x_2;\ \theta) \cdots p(x_n;\ \theta),$$

где $p(x_i;\ \theta)$ — вероятность события $\{X = x_i\}$, зависящая от θ.

Для **непрерывной случайной величины** **функция правдоподобия** определяется следующим образом

$$L(x_1, x_2, \dots, x_n;\ \theta) = f(x_1;\ \theta) \cdot f(x_2;\ \theta) \cdots f(x_n;\ \theta),$$

где $f(x_i;\ \theta)$ значение плотности распределения случайной величины X в точке x_i.

Раздел 1. Оценки параметров

Чем больше $L(x_1, x_2, \dots, x_n; \theta)$, тем вероятнее (или **правдоподобнее**) получить при наблюдениях именно *данную к о н к р е т н у ю* в ы б о р к у $x_2, \dots, x_n$.

Раздел 1. Оценки параметров

Оценкой максимального правдоподобия параметра θ называется такая статистика $\hat{\theta}$, значения которой для любой выборки удовлетворяют условию:

$$L\left(x_1, x_2, \dots, x_n; \hat{\theta}\right) = \max_{\theta} L(x_1, x_2, \dots, x_n;\ \theta).$$

Оценку максимального правдоподобия находят, решая **уравнение правдоподобия**

$$\frac{\partial \ln L(x_1, x_2, \dots, x_n;\ \theta)}{\partial \theta} = 0$$

Раздел 1. Оценки параметров

Важность метода максимального правдоподобия связана с его оптимальными свойствами.

Так, если для параметра θ существует **эффективная** оценка $\widehat{\theta}_{э}$, то оценка максимального правдоподобия единственная и равна $\widehat{\theta}_{э}$.

Кроме того, при достаточно общих условиях оценки максимального правдоподобия являются **состоятельными** и имеют **асимптотически нормальное распределение**.

Раздел 1. Оценки параметров

Пример. Пусть $X_1, X_2, \dots, X_n$ - выборка из распределения Бернулли $Bin(1, \theta)$. Требуется найти методом максимального правдоподобия оценку $\hat{\theta}$.

Решение.

Для **дискретной случайной величины** **функция правдоподобия** определяется по формуле $L(x_1, x_2, \dots, x_n;\ \theta) = p(x_1;\ \theta) \cdot p(x_2;\ \theta) \cdots p(x_n;\ \theta)$,
где $p(x_i;\ \theta)$ — вероятность события $\{X = x_i\}$, зависящая от θ.

Раздел 1. Оценки параметров

Пример. Случайная величина X представляет собой продолжительность безотказной работы элемента электронной аппаратуры. Известно, что ее распределение подчиняется **закону Релея** с плотностью распределения

$$f(x) = (2x/\theta)e^{-x^2/\theta}, x \geq 0.$$

Требуется найти оценку $\hat{\theta}$ неизвестного значения параметра θ методом максимального правдоподобия.

Решение.

Для **непрерывной случайной величины** **функция правдоподобия** определяется следующим образом $L(x_1, x_2, \dots, x_n;\ \theta) = f(x_1;\ \theta) \cdot f(x_2;\ \theta) \cdots f(x_n;\ \theta)$,
где $f(x_i;\ \theta)$ значение плотности распределения случайной величины X в точке x_i.

Раздел 1. Оценки параметров

Бывают случаи, когда функция правдоподобия достигает максимума не во внутренней точке, а на границе некоторой области, либо когда она просто не дифференцируема в точке максимума.

Такие случаи называются **нерегулярными**.

Раздел 1. Оценки параметров

Пример. Дана выборка из **распределения Лапласа** с плотностью

$$f(x) = \frac{1}{2} e^{-|x-\theta|}.$$

Найдите оценку параметра θ методом максимального правдоподобия.

Решение.

Для **непрерывной случайной величины** **функция правдоподобия** определяется следующим образом $L(x_1, x_2, \dots, x_n;\ \theta) = f(x_1;\ \theta) \cdot f(x_2;\ \theta) \cdot \dots \cdot f(x_n;\ \theta)$,
где $f(x_i;\ \theta)$ значение плотности распределения случайной величины X в точке x_i.

Раздел 1. Оценки параметров

Пример. Найдите оценки параметров a и b по методу максимального правдоподобия для **равномерного распределения** $Unif([a; b])$.

Решение.

Для **непрерывной случайной величины** **функция правдоподобия** определяется следующим образом $L(x_1, x_2, \dots, x_n;\ \theta) = f(x_1;\ \theta) \cdot f(x_2;\ \theta) \cdot \dots \cdot f(x_n;\ \theta)$,
где $f(x_i;\ \theta)$ значение плотности распределения случайной величины X в точке x_i.

Выборка из двумерного нормального распределения

Раздел 1. Оценки параметров

Пример. Пусть $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ - выборка **из двумерного нормального распределения**:

$$N\left(\big(E(X); E(Y)\big); \begin{Vmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{Vmatrix}\right) =$$

$$= N\left((0;0); \begin{Vmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{Vmatrix}\right)$$

неизвестными параметрами $\theta_1 = \sigma^2 > 0$ и $\theta_2 = \rho \in (-1; 1)$. Построить ОМП $\hat{\sigma}^2$ и $\hat{\rho}$.

Решение.

Раздел 1. Оценки параметров

Итак, для выборки $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ - **из двумерного нормального распределения**:

$$N\left(\left(E(X); E(Y)\right); \begin{Vmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{Vmatrix}\right) =$$

$$= N\left((0;0); \begin{Vmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{Vmatrix}\right)$$

С неизвестными параметрами $\theta_1 = \sigma^2 > 0$ и $\theta_2 = \rho \in (-1; 1)$.

Оценки максимального правдоподобия

$$\hat{\sigma}^2 = \frac{1}{2n}\sum_{k=1}^{n}\left(X_k^2 + Y_k^2\right) \quad \text{и} \quad \hat{\rho} = \frac{2\sum_{k=1}^{n} X_k Y_k}{\sum_{k=1}^{n}\left(X_k^2 + Y_k^2\right)}$$

неизвестных параметров σ^2 и ρ.

Раздел 4. Интервальные статистические оценки

Интервальные оценки и доверительные области.

Распределение χ^2

ЗАКОН РАСПРЕДЕЛЕНИЯ ПИРСОНА

Пусть Z_1 , Z_2 , …, Z_k — независимые с.в., распределенные по нормальному закону с параметрами $a = 0,\ \sigma^2 = 1\ [Z_i \sim N(0;\ 1), i = 1,2, \dots, k]$.

Закон распределения с.в.

$$\chi_k^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

по определению называется **законом распределения Пирсона** с k степенями свободы или **законом распределения «Хи квадрат»** с k степенями свободы.

Числом степеней свободы k распределения называется число независимых значений случайной величины. Это число равно числу наблюдений (вариантов) n за вычетом числа уравнений связи L, которые накладываются на эти наблюдения.

Раздел 4. Интервальные статистические оценки

С.в.

$$\chi_k^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

где $Z_i \sim N(0;\ 1)$ – независимые с.в., $(i = 1,2, \dots, k)$

обозначается $\mathcal{X}^2(k)$, то есть

$$\chi_k^2 \sim \mathcal{X}^2(k).$$

С.в. $\chi_k^2 \sim \mathcal{X}^2(k)$, может принимать **только неотрицательные значения**.

Раздел 4. Интервальные статистические оценки

Плотность этого распределения имеет вид

$$f(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}} \cdot \Gamma(\frac{n}{2})} \cdot x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}, x > 0; \\ 0, \qquad\qquad x \leq 0. \end{cases}$$

Функция плотности распределения хи-квадрат зависит лишь только от одного параметра – числа степеней свободы.

Упражнение. Доказать, что $f(x)$ является плотностью.

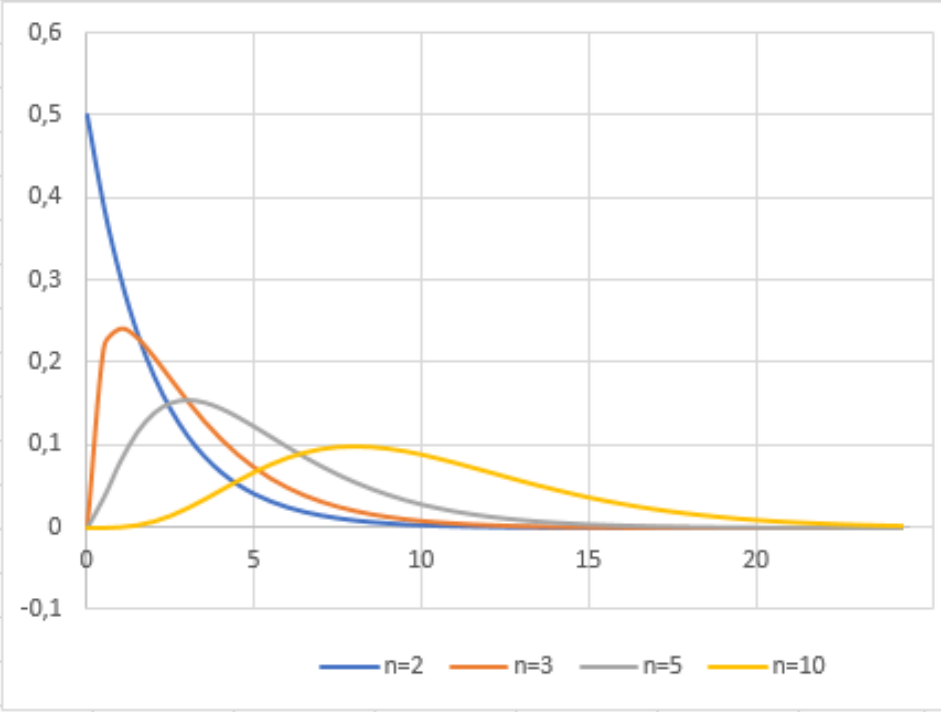
Раздел 4. Интервальные статистические оценки

Графики плотностей $f(x)$ распределения хи-квадрат:

MS Excel:

	A	B	C	D	E
1	x	y			
2	n=	2	3	5	10
3	0	0,5	0	0	0
4	0,5	0,3894	0,219696	0,036616	6,34E-05
5	1	0,303265	0,241971	0,080657	0,00079
6	1,5	0,236183	0,230799	0,1154	0,003114
7	2	0,18394	0,207554	0,138369	0,007664
8	2,5	0,143252	0,180722	0,150602	0,014572
9	3	0,111565	0,15418	0,15418	0,023533
10	3,5	0,086887	0,129697	0,151313	0,033954
11	4	0,067668	0,107982	0,143976	0,045112
12	4,5	0,0527	0,089198	0,133797	0,056276
13	5	0,041042	0,073225	0,122042	0,066801
14	5,5	0,031964	0,059811	0,109654	0,076169
15	6	0,024894	0,048652	0,097304	0,084016
16	6,5	0,019387	0,039438	0,085448	0,090123
17	7	0,015099	0,031873	0,074371	0,094406
18	7,5	0,011759	0,025694	0,064236	0,09689
19	8	0,009158	0,020667	0,055112	0,097683
20	8,5	0,007132	0,016591	0,047007	0,096953
21	9	0,005554	0,013296	0,039887	0,094904
22	9,5	0,004326	0,010638	0,033688	0,091756
...	...	...	...	...	...
52	24,5	2,39E-06	9,45E-06	7,72E-05	0,002245
53	25	1,86E-06	7,43E-06	6,19E-05	0,001895

=ХИ2.РАСП(A3;E2;0)



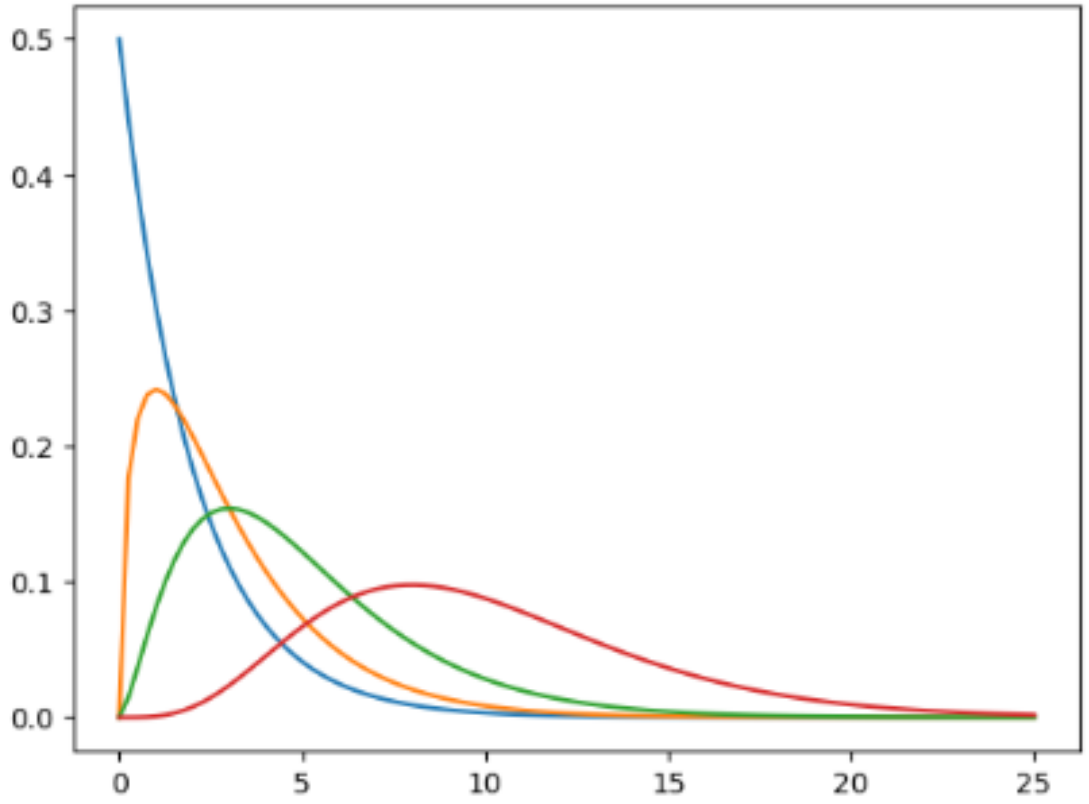
Основные распределения статистики

Графики плотностей $f(x)$ распределения хи-квадрат:

 # Python:

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as sts
```

```
x = np.linspace(0, 25, 100)
for n in 2, 3, 5, 10:
    y = sts.chi2(df=n).pdf(x)
    plt.plot(x, y)
```



Раздел 4. Интервальные статистические оценки

Замечание. Распределение «хи-квадрат» с двумя степенями свободы совпадает с **показательным распределением** с параметром $\lambda = \frac{1}{2}$.

Раздел 4. Интервальные статистические оценки

Функция распределения:

$$F(x) = I\left(\frac{x}{2}, \frac{n}{2}\right).$$

Коэффициент асимметрии:

$$S_k = \sqrt{\frac{8}{n}}.$$

Коэффициент эксцесса:

$$E_x = \frac{12}{n}.$$

$$I(x,n) = \frac{\Gamma(x,n)}{\Gamma(n)}, \Gamma(n) = \int_0^{\infty} t^{n-1}e^{-t}dt\,, \Gamma(x,n) = \int_0^{x} t^{n-1}e^{-t}dt\,, n > 0.$$

Раздел 4. Интервальные статистические оценки

Теорема 1.

В выборке X_1 , …, X_1 из нормально распределенной генеральной совокупности выборочное среднее $\bar{X}$ и исправленная выборочная дисперсия S^2 взаимно независимы. Величина $\frac{(n-1)S^2}{\sigma^2}$ имеет распределение $\mathcal{X}^2(n-1)$.

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots X_n); \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Раздел 4. Интервальные статистические оценки

Теорема 2. Если случайные величины X и Y независимы, распределенные по законам:

$$X \sim \mathcal{X}^2(n), Y \sim \mathcal{X}^2(n),$$

то сумма этих случайных величин также имеет распределение $\mathcal{X}^2$:

$$X + Y \sim \mathcal{X}^2(n + k).$$

(То есть распределение $\mathcal{X}^2$ **устойчиво относительно суммирования**).

Раздел 4. Интервальные статистические оценки

Упражнение. Доказать, что если $X_1, \ldots, X_n \sim N(a, \sigma^2)$ и независимы, то с.в.

$$\chi_k^2 = \sum_{i=1}^{n} \left(\frac{X_i - a}{\sigma} \right)^2 \sim \chi^2(n).$$

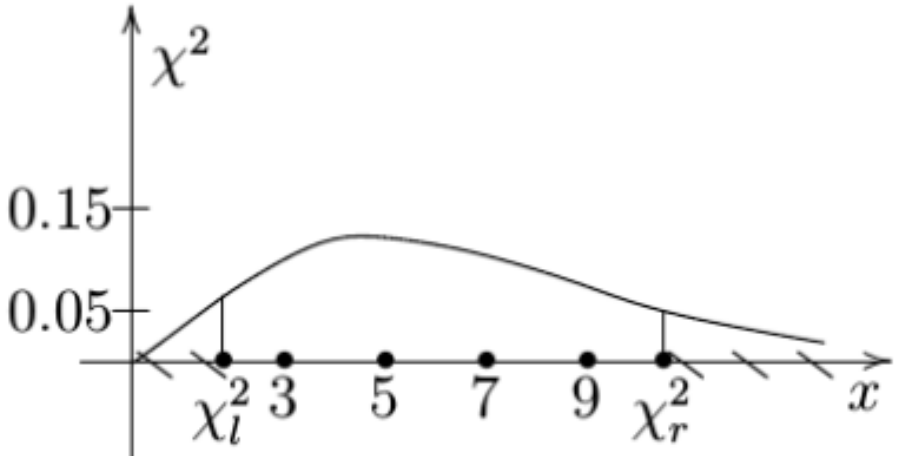
Раздел 4. Интервальные статистические оценки

Замечание. Распределение хи-квадрат стремится к нормальному распределению при $k \to \infty$. Математическое ожидание и дисперсия с.в. χ^2_k равны соответственно

$$\mathbf{E}\left(\chi^2_k\right) = k\,,$$

$$\mathbf{Var}\left(\chi^2_k\right) = 2k\,.$$

Математическое ожидание больше моды этого распределения, потому что правый хвост "тяжелее" левого.



Распределение Стьюдента

Раздел 4. Интервальные статистические оценки

В начале 20-го века статистик Уильям С. Госсет, сотрудник ирландского отделения пивоваренной компании Guinness, заинтересовался проблемой оценки математического ожидания при неизвестном стандартном отклонении.

Поскольку компания Guinness запрещала своим сотрудникам публиковать работы под собственными именами, Госсет взял псевдоним Стьюдент. По этой причине распределение, предложенное Госсетом, называется **t-распределением Стьюдента**.

ЗАКОН РАСПРЕДЕЛЕНИЯ СТЬЮДЕНТА

Пусть $Z \sim N(0;\ 1)$ (с.в., распределенная **по нормальному закону** с параметрами $a = 0, \sigma^2 = 1$), а $\chi_k^2 \sim \mathcal{X}^2(k)$, причем с.в. Z и χ_k^2 независимы.

Закон распределения случайной величины

$$T_k = \frac{Z}{\sqrt{\frac{\chi_k^2}{k}}}$$

по определению называется **законом распределения Стьюдента** (или ***t*-распределением**) с k степенями свободы.

Раздел 4. Интервальные статистические оценки

$Z \sim N(0;\ 1), \chi_k^2 \sim \mathcal{X}^2(k)$ - независимые с.в.

Закон распределения Стьюдента с k степенями свободы

$$T_k = \frac{Z}{\sqrt{\frac{\chi_k^2}{k}}}$$

обозначается $t(k)$, то есть

$$T_k \sim t(k).$$

Раздел 4. Интервальные статистические оценки

Плотность распределения Стьюдента с n степенями свободы имеет вид

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

$\Gamma(n) = \int_0^\infty t^{n-1}e^{-t}dt\,, n > 0; B(u,v) = \int_0^1 t^{u-1}(1-t)^{v-1}dt$, где $u, v > 0$

Раздел 4. Интервальные статистические оценки

Функция распределения:

$$F(x) = \int_{-\infty}^{x} f(t)dt = 1 - \frac{1}{2} I_{x(t)}\left(\frac{n}{2}, \frac{1}{2}\right),$$

где $x(t) = \frac{n}{t^2+n}$.

Коэффициент асимметрии:

$$S_k = 0, n > 3.$$

Коэффициент эксцесса:

$$E_x = \frac{6}{n-4}, n > 4.$$

$I_x(u,v) = \frac{B_x(u,v)}{B(u,v)}$;

$B(u,v) = \int_0^1 t^{u-1}(1-t)^{v-1}dt,\ B_x(u,v) = \int_0^x t^{x-1}(1-t)^{v-1}dt,$

где $u, v > 0$

Раздел 4. Интервальные статистические оценки

Основные частные случаи:

Распределение Стьюдента с одной степенью свободы ($n = 1$) это **стандартное распределение Коши**

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\operatorname{arctg}(x), \qquad f(x) = \frac{1}{\pi}\frac{1}{(1 + x^2)};$$

Распределение Стьюдента с двумя степенями свободы ($n = 2$)

$$F(x) = \frac{1}{2} + \frac{x}{2\sqrt{2 + x^2}}, f(x) = \frac{1}{(1 + x^2)^{3/2}};$$

Распределение Стьюдента с бесконечным числом степеней свободы ($n = \infty$)

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

совпадает с плотностью вероятности **стандартного нормального распределения.**

Раздел 4. Интервальные статистические оценки

С.в., распределенная по закону Стьюдента, может принимать **любые значения в интервале** $(-\infty;\ \infty)$.

Распределение Стьюдента симметрично относительно 0: если $X\ \sim\ t(k)$, то $-X\ \sim\ t(k)$.

Графики функции $f(x)$, называемые **кривыми Стьюдента**, симметричны при всех $n = 1,2,\ldots$ относительно оси ординат.

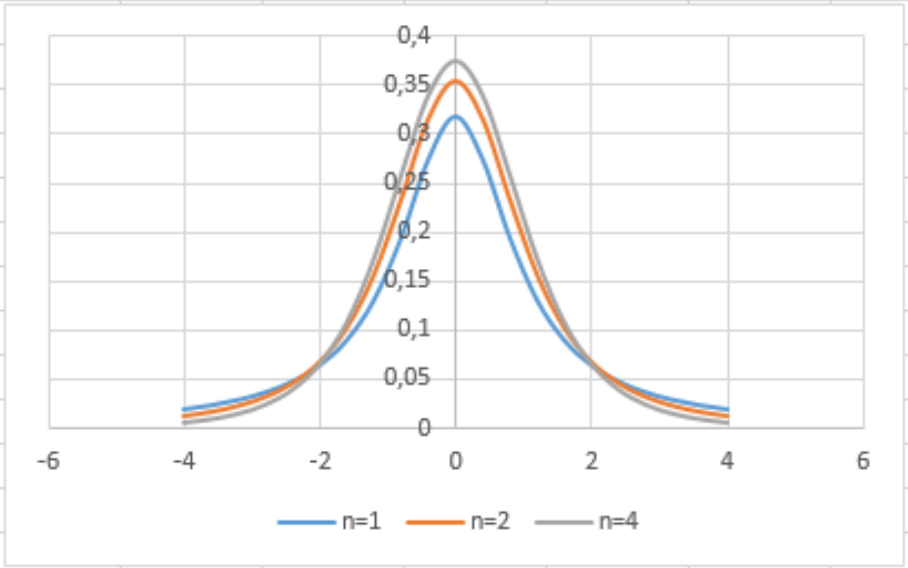
Раздел 4. Интервальные статистические оценки

Графики плотностей $f(x)$ распределения Стьюдента:

MS Excel:

	A	B	C	D
1	x	y		
2	n=	1	2	4
3	-4	0,018724	0,013095	0,006708
4	-3,6	0,022802	0,017282	0,01013
5	-3,2	0,028319	0,023352	0,015682
6	-2,8	0,036008	0,032397	0,024877
7	-2,4	0,047087	0,04626	0,040323
8	-2	0,063662	0,068041	0,066291
9	-1,6	0,089413	0,102696	0,108873
10	-1,2	0,130455	0,156734	0,173854
11	-0,8	0,194091	0,233128	0,258754
12	-0,4	0,274405	0,315006	0,339976
13	0	0,31831	0,353553	0,375
14	0,4	0,274405	0,315006	0,339976
15	0,8	0,194091	0,233128	0,258754
16	1,2	0,130455	0,156734	0,173854
17	1,6	0,089413	0,102696	0,108873
18	2	0,063662	0,068041	0,066291
19	2,4	0,047087	0,04626	0,040323
20	2,8	0,036008	0,032397	0,024877
21	3,2	0,028319	0,023352	0,015682
22	3,6	0,022802	0,017282	0,01013
23	4	0,018724	0,013095	0,006708

=СТЬЮДЕНТ.РАСП(A3;D2;0)



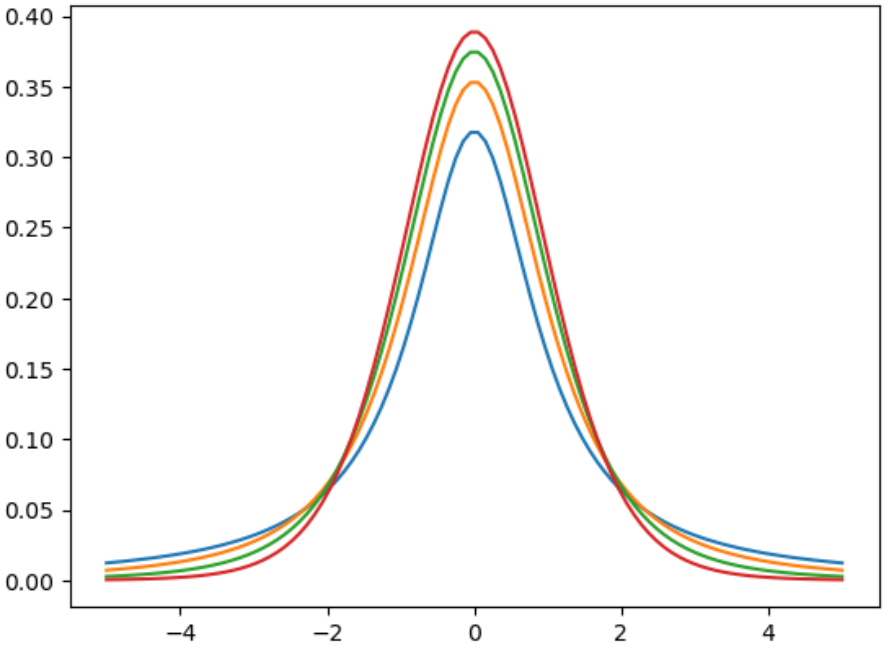
Раздел 4. Интервальные статистические оценки

Графики плотностей $f(x)$ распределения Стьюдента :

 # Python:

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as sts
```

```
x = np.linspace(-5, 5, 100)
for n in 1, 2, 4, 10:
    y = sts.t(df=n).pdf(x)
    plt.plot(x, y)
```



Раздел 4. Интервальные статистические оценки

При $k > 2$ в силу симметрии плотности **математическое ожидание** с.в., имеющей t–распределение, ($T_k \sim t(k)$) равно нулю, т.е. $E(T_k) = 0$, а **дисперсия** равна

$$Var(T_k) = \frac{k}{k-2}.$$

$T_k = \frac{Z}{\sqrt{\frac{\chi_k^2}{k}}} \sim t(k),\ Z \sim N(0;\ 1), \chi_k^2 \sim \mathcal{X}^2(k)$ - независимые с.в.

Раздел 4. Интервальные статистические оценки

Теорема. Если $X_1, \ldots, X_n$ независимы и распределены по нормальному закону $N(a; \sigma^2)$, то с.в.

$$T = \frac{\bar{X} - a}{S/\sqrt{n}} \sim t(n-1).$$

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots X_n); \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Распределение Фишера

ЗАКОН РАСПРЕДЕЛЕНИЯ ФИШЕРА — СНЕДЕКОРА

Пусть теперь с.в. $\chi^2_{k_1} \sim \mathcal{X}^2(k_1)$ и $\chi^2_{k_2} \sim \mathcal{X}^2(k_2)$ независимы. Закон распределения случайной величины

$$F_{k_1;k_2} = \frac{\chi^2_{k_1}/k_1}{\chi^2_{k_2}/k_2}$$

по определению называется **законом распределения Фишера — Снедекора** (или **Фишера,** или ***F*-распределением**) с k_1 и k_2 степенями свободы.

Раздел 4. Интервальные статистические оценки

$\chi^2_{k_1} \sim \mathcal{X}^2(k_1)$ и $\chi^2_{k_2} \sim \mathcal{X}^2(k_2)$ независимые с.в.

С.в. $F_{k_1;k_2} = \frac{\chi^2_{k_1}/k_1}{\chi^2_{k_2}/k_2}$, имеющая распределение Фишера обозначается $F(k_1; k_2)$, то есть

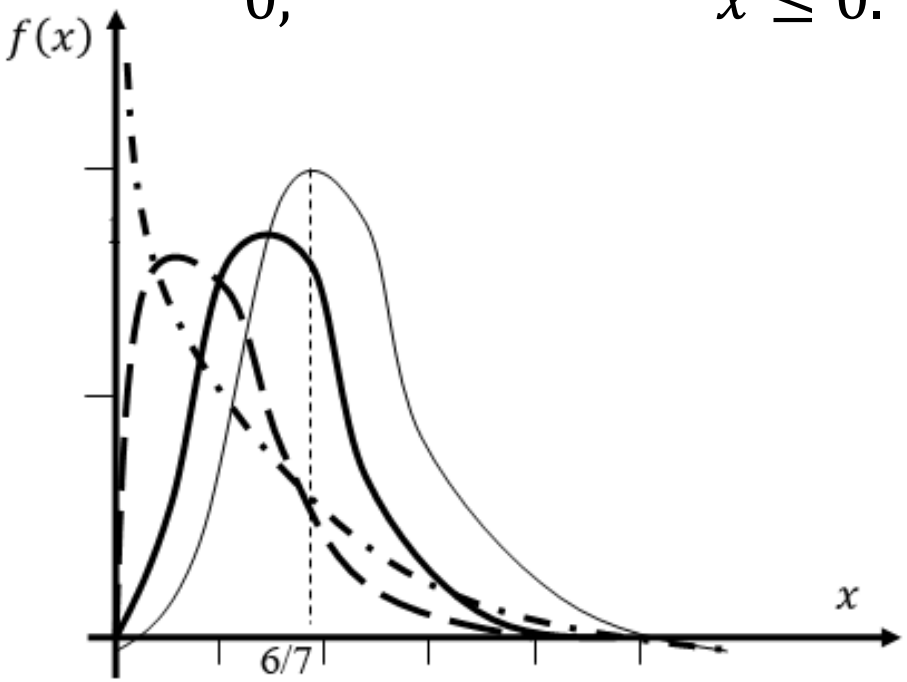
$$F_{k_1;k_2} \sim F(k_1; k_2).$$

Случайная величина $F_{k_1;k_2} \sim F(k_1; k_2)$ может принимать **только неотрицательные значения**.

Раздел 4. Интервальные статистические оценки

Плотность распределения Фишера имеет вид:

$$f(x) = \begin{cases} \frac{1}{B(\frac{k_1}{2};\frac{k_2}{2})} \cdot \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} x^{\frac{k_1}{2}-1} \cdot \left(1+\frac{k_1}{k_2}x\right)^{-\frac{k_1+k_2}{2}}, x > 0; \\ 0, \qquad x \le 0. \end{cases}$$



$$B(u,v) = \int_0^1 t^{u-1}(1-t)^{v-1}dt, \text{ где } u, v > 0$$

Раздел 4. Интервальные статистические оценки

Функция распределения:

$$F(x) = \int_{-\infty}^{x} f(t)dt = I_{\frac{k_1+x}{k_2+k_1x}}\left(\frac{k_1}{2}, \frac{k_2}{2}\right).$$

Коэффициент асимметрии:

$$S_k = \frac{2(2k_1 + k_2 - 1)}{(k_2 - 6)}\sqrt{\frac{2(k_2 - 4)}{k_1(k_1 + k_2 - 2)}}.$$

Коэффициент эксцесса:

$$E_x = \frac{12[(k_2 - 2)^2(k_2 - 4) + k_1(5k_2 - 22)(k_1 + k_2 - 2)]}{k_1(k_1 + k_2 - 2)(k_2 - 6)(k_2 - 8)},$$

$k_2 > 8.$

$I_x(u, v) = \frac{B_x(u, v)}{B(u, v)}$;

$B(u, v) = \int_0^1 t^{u-1}(1 - t)^{v-1}dt,\ B_x(u, v) = \int_0^x t^{x-1}(1 - t)^{v-1}dt,$

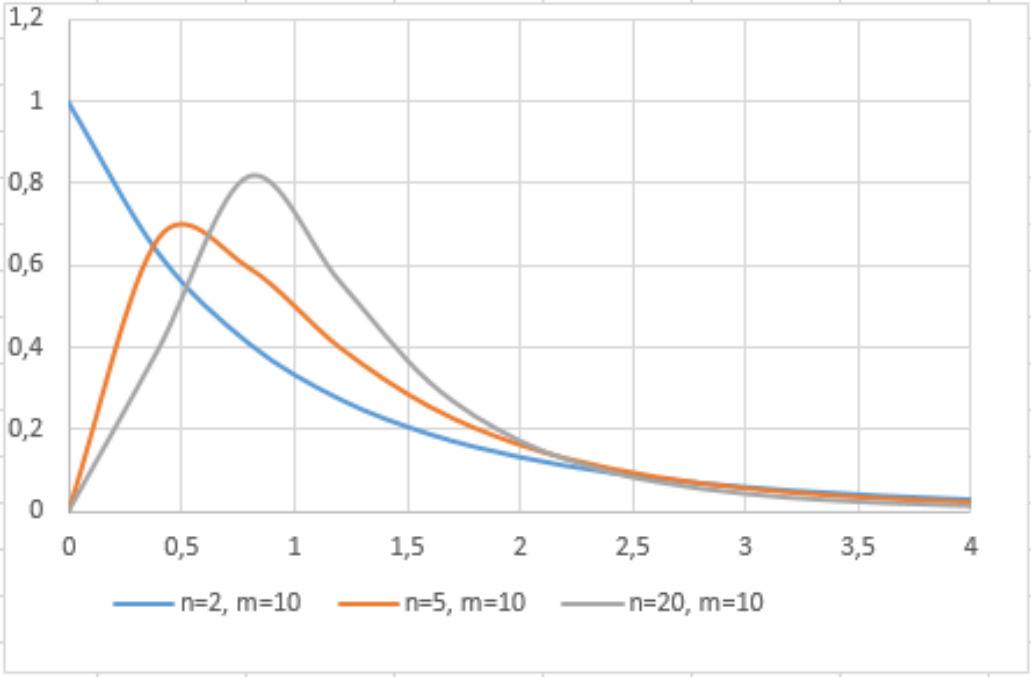
где $u, v > 0$

Раздел 4. Интервальные статистические оценки

Графики плотностей $f(x)$ распределения Фишера:

MS Excel:

	A	B	C	D	E	F	G	H	I	J	K
1	**x**	**y**									
2	**n=**	**2**	**5**	**20**							
3	**m=**	**10**									
4	0	1	0	0	**=F.РАСП($A4;D$2;B3;0)**						
5	0,4	0,63017	0,668253	0,398278							
6	0,8	0,410442	0,594819	0,820245							
7	1,2	0,275087	0,401405	0,563873							
8	1,6	0,189041	0,255474	0,315555							
9	2	0,13281	0,162006	0,17197							
10	2,4	0,095155	0,104197	0,095767							
11	2,8	0,069383	0,06837	0,055207							
12	3,2	0,051397	0,045828	0,033007							
13	3,6	0,038622	0,031367	0,020429							
14	4	0,029401	0,021897	0,013051							
15											
16											
17											
18											
19											
20											



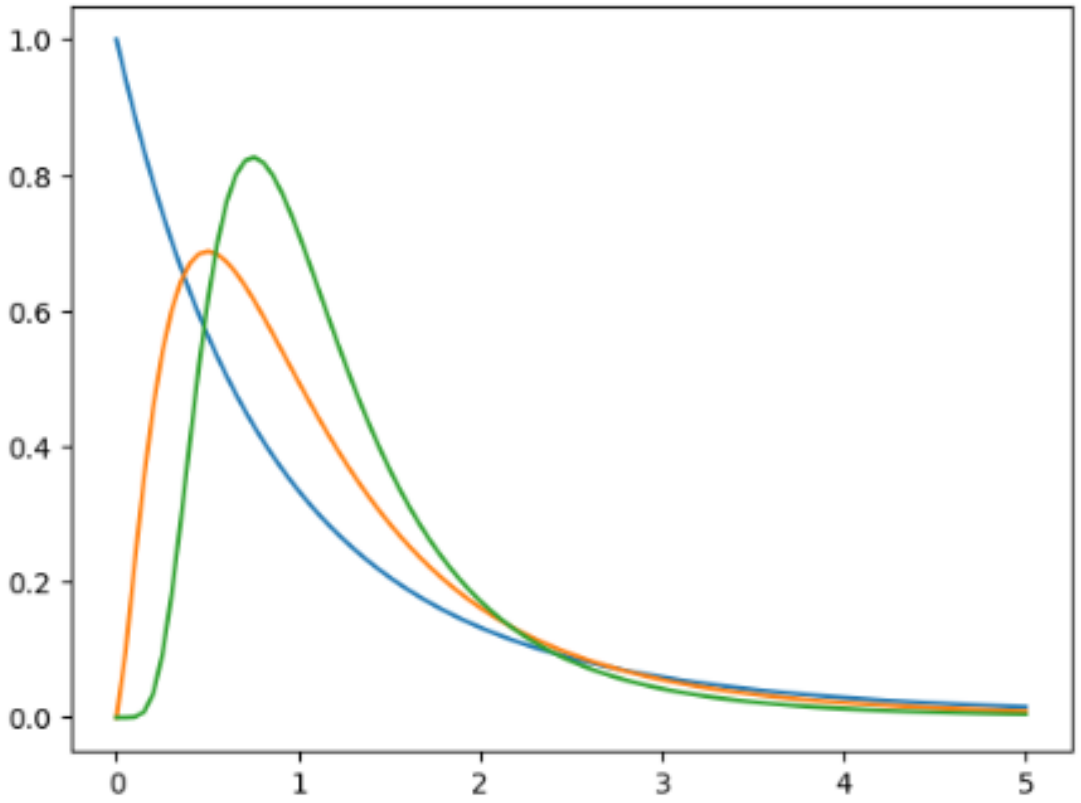
Раздел 4. Интервальные статистические оценки

Графики плотностей $f(x)$ распределения Фишера:

 # Python:

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as sts
```

```
x = np.linspace(0, 5, 100)
for n in 2, 5, 20:
    y = sts.f(n, 10).pdf(x)
    plt.plot(x, y)
```



Раздел 4. Интервальные статистические оценки

Упражнение 1. Доказать, что если $F \sim F(k_1; k_2)$, то

$$\frac{1}{F} \sim F(k_2; k_1).$$

Упражнение 2. Доказать, что если $t_n \sim t(n)$, то

$$t_n^2 \sim F(1; n).$$

Раздел 4. Интервальные статистические оценки

Математическое ожидание и **дисперсия** с.в. $F \sim F(k_1; k_2)$, имеющей F–распределение, равны соответственно:

$$E(F) = \frac{k_2}{k_2 - 1}, k_2 > 2;$$

$$Var(F) = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}, k_2 > 4.$$

Квантили и процентные точки с.в.

Квантилью (или левосторонней критической границей) уровня α с.в. X называется такое число x_α, что

$$F_X(x_\alpha) = \alpha,$$

(т. е. $\boldsymbol{P}\{X \le x_\alpha\} = \alpha$),

а **100α%-ной точкой (или правосторонней критической границей уровня** α) с.в. X называется такое число ω_α, что

$$F_X(\omega_\alpha) = 1 - \alpha$$

(т. е. $\boldsymbol{P}\{X > \omega_\alpha\} = \alpha$).

Левосторонняя и правосторонняя критические границы одного и того же уровня связаны между собой очевидным соотношением $x_\alpha = \omega_{1-\alpha}$

Раздел 4. Интервальные статистические оценки

Замечание. Эти определения корректны только для случая абсолютно непрерывной с.в.

Если же с.в. *X* является **дискретной** или **смешанной**, то может оказаться так, что либо числа, определяемого этими формулами не существует, либо таких чисел бесконечно много.

Раздел 4. Интервальные статистические оценки

Поскольку для любого $\alpha \in (0;\ 1)$ можно найти два таких числа x'_α и x''_α, что $F_X(x'_\alpha) \leq \alpha$, а $F_X(x''_\alpha) > \alpha$,
причем

$$F_X(x) \geq \alpha, \forall x > x'_\alpha$$

и

$$F_X(x) \leq \alpha\ , \forall\ x < x''_\alpha,$$

то в этом случае **квантилью** *уровня* α с.в. X называется любое число $x_\alpha \in [x'_\alpha; x''_\alpha]$,

а **100α%-*ной точкой*** этой с.в. — любая ее квантиль уровня $(1 - \alpha)$.

Раздел 4. Интервальные статистические оценки

Двусторонними критическими границами уровня α случайной величины X называются такие числа $\underline{x}_\alpha$, $\overline{x}_\alpha$, что одновременно

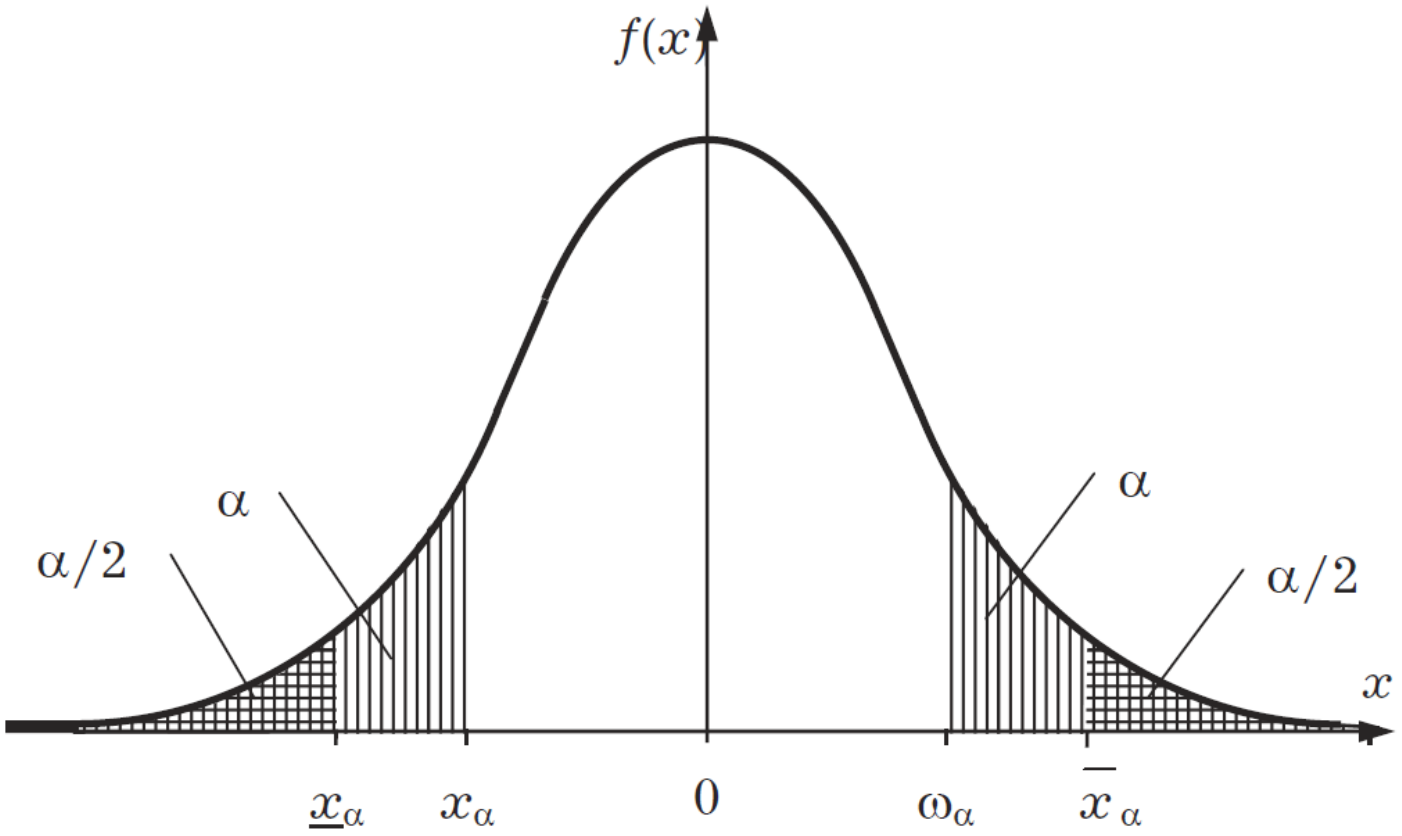
$$F_X\left(\underline{x}_\alpha\right) = \frac{\alpha}{2}, \qquad F_X\left(\overline{x}_\alpha\right) = 1 - \frac{\alpha}{2}$$

т.е. $P\left(X \leq \underline{x}_\alpha\right) = P\left(X > \overline{x}_\alpha\right) = \frac{\alpha}{2}$.

Односторонние и двусторонние критические границы случайной величины X связаны следующими соотношениями:

$$\underline{x}_\alpha = x_{\alpha/2} = \omega_{1-\alpha/2}; \qquad \overline{x}_\alpha = x_{1-\alpha/2} = \omega_{\alpha/2}.$$

Раздел 4. Интервальные статистические оценки



Раздел 4. Интервальные статистические оценки

Пример. Вычислить процентную точку $\boldsymbol{F}_{0,005}(3; 7)$.

Решение.

 # MS Excel:

	A	B	C
1	alpha=	0,005	
2	n=	3	
3	m=	7	
4			
5	0,022505	=F.ОБР(B1;B2;B3)	

 # Python:

```
import scipy.stats as sts
```

```
alpha=0.005
n=3
m=7
f_alpha=sts.f.ppf(alpha, n,m)
f_alpha
```

```
0.022505237799116617
```

Определение интервальной оценки

Раздел 4. Интервальные статистические оценки

Точечные оценки $\hat{\theta}_n$ неизвестного параметра θ хороши в качестве первоначальных результатов обработки наблюдений.

Их недостаток состоит в том, что неизвестна точность оценивания параметра.

Поэтому и возникает задача о приближении параметра θ не одним числом, а целым интервалом

$$\hat{\theta}_n - \delta < \theta < \hat{\theta}_n + \delta$$

где δ — **точность оценки**.

Раздел 4. Интервальные статистические оценки

Разумеется, чем меньше длина интервала

$$\hat{\theta}_n - \delta < \theta < \hat{\theta}_n + \delta,$$

тем точнее оценка параметра.

Однако статистические методы позволяют говорить только о том, что это неравенство выполняется с некоторой вероятностью.

Раздел 4. Интервальные статистические оценки

Пусть $X_1, \ldots, X_n$ — выборка объема n из генеральной совокупности X с функцией распределения $F(x; \theta)$, зависящей от параметра θ, значение которого неизвестно.

Доверительным интервалом (интервальной оценкой) для параметра θ с доверительной вероятностью γ называется такой интервал $(\underline{\theta}, \overline{\theta})$, для которого выполняется условие:

$$P\left(\left\{\underline{\theta} < \theta < \overline{\theta}\right\}\right) = \gamma.$$

Раздел 4. Интервальные статистические оценки

Интервальная оценка $(\underline{\theta}, \overline{\theta})$ представляет собой интервал со случайными границами, который с заданной вероятностью γ накрывает неизвестное истинное значение параметра θ.

Нижняя $\underline{\boldsymbol{\theta}}$ и верхняя $\overline{\boldsymbol{\theta}}$ границы интервальной оценки являются функциями случайной выборки: $\underline{\theta} = \underline{\theta}(X_1, \ldots, X_n)$, $\overline{\theta} = \overline{\theta}(X_1, \ldots, X_n)$.

Для различных реализаций случайной выборки они могут принимать различные значения.

Раздел 4. Интервальные статистические оценки

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

$$P\left(\theta < \overline{\theta}\right) = P\left(\theta > \underline{\theta}\right).$$

Раздел 4. Интервальные статистические оценки

Величина

$$\alpha = 1 - \gamma$$

задает вероятность того, что истинное значение параметра θ окажется вне построенного интервала и называется **уровнем значимости**.

Раздел 4. Интервальные статистические оценки

Метод центральной статистики

Раздел 4. Интервальные статистические оценки

Пусть $g_{\vec{x}_n}(\theta)$ – некоторая функция, зависящая от векторного параметра $\vec{x}_n$,

$\vec{X}_n(X_1, \ldots, X_n)$ – выборка объема n из распределения, зависящего от параметра θ.

Заменяя $\vec{x}_n$ на случайный вектор $\vec{X}_n$, получим случайную величину $Y = g_{\vec{X}_n}(\theta)$.

Если распределение Y не зависит от θ, случайная величина $Y = g_{\vec{X}_n}(\theta)$ называется **центральной статистикой**.

Раздел 4. Интервальные статистические оценки

При построении γ -доверительного интервала **методом центральной статистики** предположим, что $g_{\vec{x}_n}(\theta)$ – непрерывная и возрастающая (убывающая) функция от θ при любом фиксированном $\vec{x}_n$.

Тогда при любом $\vec{x}_n$ для функции $y = g_{\vec{x}_n}(\theta)$ **существует обратная монотонная функция** $g^{-1}_{\vec{x}_n}(y)$.

Поскольку $g^{-1}_{\vec{x}_n}\left(g_{\vec{x}_n}(\theta)\right) = \theta$, для случайной величины Y имеет место аналогичное равенство

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

Раздел 4. Интервальные статистические оценки

Предположим дополнительно, что распределение Y непрерывно.

Используя такую центральную статистику Y, нетрудно построить γ -доверительный интервал для θ.

Действительно, распределение Y не зависит от θ и непрерывно, поэтому даже не зная истинного значения θ, для любого γ можно подобрать интервал (a, b) так, чтобы центральная статистика Y попадала в этот интервал с заранее заданной вероятностью γ,

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

Раздел 4. Интервальные статистические оценки

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

$$\left\{g^{-1}_{\vec{X}_n}(a) < \theta < g^{-1}_{\vec{X}_n}(b)\right\}.$$

Отсюда, с учетом равенства $g^{-1}_{\vec{X}_n}(Y) = \theta$, получаем вероятность двойного неравенства

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

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

Раздел 4. Интервальные статистические оценки

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

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

Раздел 4. Интервальные статистические оценки

Заметим, что метод применим и в тех случаях, когда накладываемые на центральную статистику дополнительные требования выполняются с вероятностью 1.

Так, функция $y = g_{\vec{x}_n}(\theta)$ может и не быть возрастающей при любом $\vec{x}_n$.

Однако, если для $\vec{X}_n(X_1, \ldots, X_n)$ функция $g_{\vec{X}_n}(\theta)$ является возрастающей с вероятностью 1, интервал $\left(g^{-1}_{\vec{X}_n}(a), g^{-1}_{\vec{X}_n}(b)\right)$ все-таки будет γ -доверительным интервалом для θ.

Интервальная оценка математического ожидания нормального распределения

Раздел 4. Интервальные статистические оценки

Теорема. Если распределение генеральной совокупности имеет конечные математическое ожидание и дисперсию, то при $n \to \infty$ основные выборочные характеристики (среднее, дисперсия, эмпирическая функция распределения) являются **нормальными**.

Раздел 4. Интервальные статистические оценки

Пусть $X_1, \ldots, X_n$ выборка объема n из генеральной совокупности X, распределенной по нормальному закону с параметрами m и σ^2.

Задана доверительная вероятность γ (или уровень значимости $\alpha = 1 - \gamma$).

Поскольку распределение $N(m, \sigma^2)$ в отношении выборки $X_1, \ldots, X_n$ играет роль генеральной совокупности, назовем m **генеральным средним**, а σ^2 – **генеральной дисперсией**.

Значение случайной величины X_i будем интерпретировать как значение признака X на i-том элементе выборочной совокупности.

Интервальная оценка математического ожидания при известной дисперсии

Раздел 4. Интервальные статистические оценки

Вследствие того, что

$$E(\bar{X}) = E(X), D(\bar{X}) = \sigma^2 / n,$$

с учетом нормальности $\bar{X}$, имеем $\bar{X} \sim N(m, \sigma^2)$.

Соответственно, случайная величина

$$Z = \frac{\bar{X} - m}{\sigma/\sqrt{n}} \sim N(0; 1).$$

Поскольку распределение Z не зависит от m, Z – **центральная статистика**.

Раздел 4. Интервальные статистические оценки

Кроме того, зависимость $Z = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$ от m является **убывающей**, что позволяет использовать эту центральную статистику для построения доверительной интервальной оценки для генерального среднего, если, конечно, значение σ известно.

Случайная величина $Z = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$ распределена по стандартному нормальному закону $N(0,1)$. Поскольку распределение Z не зависит от m, Z – **центральная статистика**.

Раздел 4. Интервальные статистические оценки

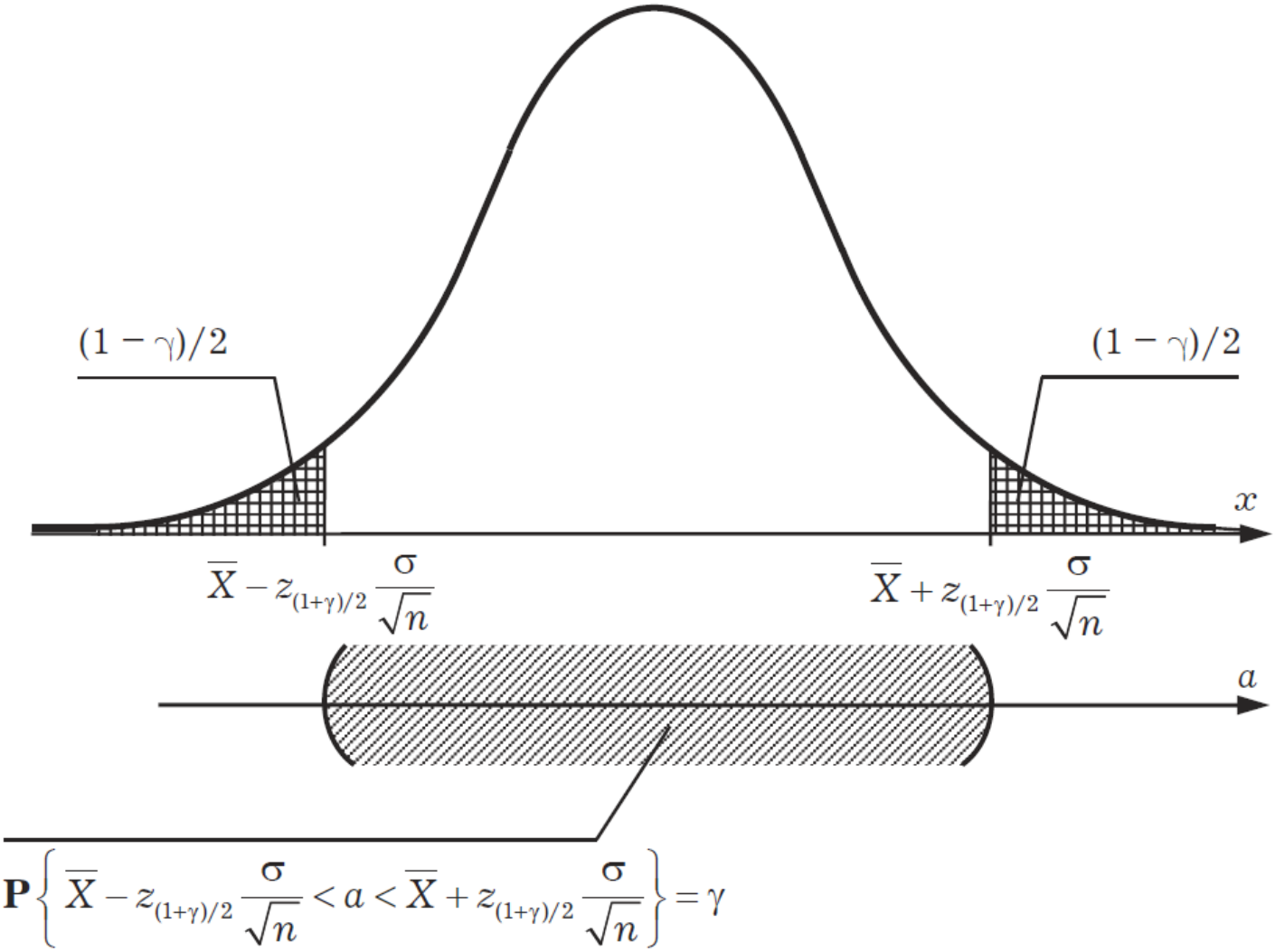
Теорема. Интервал

$$\left(\bar{X} - z_{(1+\gamma)/2} \cdot \frac{\sigma}{\sqrt{n}}; \bar{X} + z_{(1+\gamma)/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

является интервальной оценкой для параметра $m = E(X)$ с доверительной вероятностью γ.

Доказательство.

Раздел 4. Интервальные статистические оценки



Раздел 4. Интервальные статистические оценки

Теорема. Интервал

$$\left(-\infty; \bar{X} - z_{1-\gamma} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

является **правосторонним доверительным интервалом** для параметра $m = E(X)$ с доверительной вероятностью γ.

Доказательство.

Замечание.

$$\left(-\infty; \bar{X} + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Раздел 4. Интервальные статистические оценки

Теорема. Интервал

$$\left(\bar{X} - z_\gamma \cdot \frac{\sigma}{\sqrt{n}}; +\infty\right)$$

является **левосторонним доверительным интервалом** для параметра $m = E(X)$ с доверительной вероятностью γ.

Доказательство.

Замечание.

$$\left(\bar{X} - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}; +\infty\right)$$

Интервальная оценка математического ожидания при неизвестной дисперсии

Раздел 4. Интервальные статистические оценки

Пусть теперь **параметр** σ нормального закона распределения признака X генеральной совокупности **неизвестен**.

Тогда вместо дисперсии будет использоваться ее оценка — **исправленная выборочная дисперсия**

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

Пусть $t_{\alpha/2;n-1}$ — $100 \cdot \alpha/2$ -процентная точка распределения Стьюдента с $n - 1$ степенями свободы.

Раздел 4. Интервальные статистические оценки

Чтобы найти оценки, не использующие σ^2, заменим в центральной статистике Z σ на S – корень из исправленной выборочной и рассмотрим статистику

$$T = \frac{\bar{X} - m}{S/\sqrt{n}}.$$

Полученная таким образом статистика является **центральной**.

Раздел 4. Интервальные статистические оценки

Этот факт является очевидным следствием следующей теоремы

Теорема. Если $X_1, \ldots, X_n$ независимы и распределены по нормальному закону $N(m, \sigma^2)$, то

$$T = \frac{\bar{X} - m}{S/\sqrt{n}} \sim t(n-1).$$

$T = \frac{\bar{X} - m}{S/\sqrt{n}}$ – **центральной статистика**

Раздел 4. Интервальные статистические оценки

Теорема. Интервал

$$\left(\bar{X} - t_{\alpha/2;n-1} \cdot \frac{S}{\sqrt{n}}; \bar{X} + t_{\alpha/2;n-1} \cdot \frac{S}{\sqrt{n}}\right)$$

является интервальной оценкой для параметра $m = E(X)$ с доверительной вероятностью γ.

Доказательство.

Рассмотрим статистику

$$T = \frac{\bar{X} - m}{S/\sqrt{n}} \sim t(n-1).$$

Используя симметричность распределения Стьюдента, получаем:

$$P\left(\left\{-t_{\alpha/2;n-1} < T < t_{\alpha/2;n-1}\right\}\right) = 1 - 2P\left(\left\{T > t_{\alpha/2;n-1}\right\}\right) =$$
$$= 1 - 2 \cdot \frac{\alpha}{2} = 1 - \alpha = \gamma.$$

Доказательство (продолжение).

$$-t_{\alpha/2;n-1} < T < t_{\alpha/2;n-1} \Leftrightarrow -t_{\alpha/2;n-1} < \frac{\bar{X} - m}{\frac{S}{\sqrt{n}}} < t_{\alpha/2;n-1}$$

$$\Leftrightarrow -\frac{S}{\sqrt{n}} t_{\alpha/2;n-1} < \bar{X} - m < \frac{S}{\sqrt{n}} t_{\alpha/2;n-1} \Leftrightarrow$$

$$\bar{X} - t_{\alpha/2;n-1} \cdot \frac{S}{\sqrt{n}} < m < \bar{X} + t_{\alpha/2;n-1} \cdot \frac{S}{\sqrt{n}}$$

Следовательно, полученные неравенства дают интервальную оценку для параметра m с доверительной вероятностью γ. □

Теорема. Интервал $\left(\bar{X} - t_{\alpha/2;n-1} \cdot \frac{S}{\sqrt{n}}; \bar{X} + t_{\alpha/2;n-1} \cdot \frac{S}{\sqrt{n}}\right)$ является интервальной оценкой для параметра $m = E(X)$ с доверительной вероятностью γ.

Раздел 4. Интервальные статистические оценки

Пример. Требуется найти доверительный интервал с доверительной вероятностью $\gamma = 0{,}95$ для математического ожидания нормально распределенной случайной величины, если по выборке объема $n = 30$ найдено выборочное среднее $\bar{x} = 4$. Дисперсия случайной величины:

а) известна и равна $1{,}21$;

б) неизвестна; по выборке найдена исправленная выборочная дисперсия $s^2 = 1{,}21$.

Решение.

Интервальная оценка дисперсии нормального распределения

Раздел 4. Интервальные статистические оценки

Рассмотрим **два случая**, в зависимости от того, известно или нет математическое ожидание $E(X)$.

Интервальная оценка дисперсии при известном математическом ожидании

Раздел 4. Интервальные статистические оценки

При известном $E(X) = m$ в качестве точечной оценки дисперсии используется её эффективная точечная оценка

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 .$$

Пусть $\chi^2_{q;n}$ — $100 \cdot q$-процентная точка χ^2-распределения с n степенями свободы.

Раздел 4. Интервальные статистические оценки

Теорема. Интервал

$$\left(\frac{nS_0^2}{\chi^2_{\alpha/2;n}} ; \frac{nS_0^2}{\chi^2_{1-\alpha/2;n}} \right)$$

является интервальной оценкой для параметра σ^2 с доверительной вероятностью γ.

Доказательство.

Интервальная оценка дисперсии при неизвестном математическом ожидании

Раздел 4. Интервальные статистические оценки

Предположим теперь, что математическое ожидание $E(X)$ **неизвестно**.

В этом случае вместо оценки S_0^2 будет использоваться исправленная выборочная дисперсия S^2.

Раздел 4. Интервальные статистические оценки

Теорема. Интервал

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

является интервальной оценкой для параметра σ^2 с доверительной вероятностью γ.

Доказательство.

Приближенная интервальная оценка математического ожидания

Раздел 4. Интервальные статистические оценки

В случае, когда распределение случайной величины X **отлично от нормального**, в общем случае не удается точно найти вероятность, с которой интервал $(\underline{\theta};\overline{\theta})$ накрывает параметр θ распределения, однако предел вероятности

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

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

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

интервал $(\underline{\theta};\overline{\theta})$ называется **приближенным доверительным интервалом с доверительной вероятностью γ**.

Раздел 4. Интервальные статистические оценки

Пусть $X_1, \ldots, X_n$ — выборка из некоторого распределения с математическим ожиданием m и дисперсией σ^2.

Раздел 4. Интервальные статистические оценки

Разберем **первый вариант построения интервальной оценки для среднего m**:

дисперсия σ^2 известна.

Рассмотрим статистику

$$T = \frac{\bar{X} - m}{\sigma/\sqrt{n}} = \frac{X_1 - m}{\sigma/\sqrt{n}} + \cdots + \frac{X_n - m}{\sigma/\sqrt{n}}.$$

На основании **центральной предельной теоремы**, статистика T при больших объемах n случайной выборки имеет закон распределения, близкий к стандартному нормальному закону.

Раздел 4. Интервальные статистические оценки

Следовательно, интервальная оценка

$$\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

является **приближенной интервальной оценкой** для параметра m с доверительной вероятностью γ.

Раздел 4. Интервальные статистические оценки

Рассмотрим **второй случай**:

дисперсия σ^2 случайной величины X существует, но неизвестна.

Применяем еще одно приближение: вместо величины σ используем ее оценку $\tilde{\sigma}$.

Величина $\tilde{\sigma}$ при выполнении определенных условий сходится по вероятности к σ.

Достаточным условием такой сходимости является, например, существование четвертого начального момента ν_4.

Раздел 4. Интервальные статистические оценки

Приближенной интервальной оценкой с доверительной вероятностью, приблизительно равной γ, является интервал

$$\left(\bar{X} - z_{\alpha/2} \cdot \frac{\tilde{\sigma}}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \cdot \frac{\tilde{\sigma}}{\sqrt{n}}\right)$$

Раздел 4. Интервальные статистические оценки

Пример. В некотором городе население составляет 1 млн человек. Для случайно отобранных 625 жителей средний возраст $\bar{x}$ составил 33 года, исправленное выборочное стандартное отклонение составляет $\tilde{\sigma} = 15$ лет. Требуется найти приближенный доверительный интервал для среднего возраста жителей с доверительной вероятностью $\gamma = 0{,}95$.

Решение.

Приближенная интервальная оценка вероятности события

Раздел 4. Интервальные статистические оценки

Пусть проводится серия из n испытаний по схеме Бернулли с постоянной, но **неизвестной вероятностью успеха p** в каждом испытании.

И пусть X_i, $i = 1, \ldots, n$, — исход i-го испытания, т. е. $X_i = 1$, если в i-м испытании произошел успех, и $X_i = 0$ в противном случае.

По данным случайной выборки $X_1, \ldots, X_n$ требуется найти **приближенную интервальную оценку величины p с заданной доверительной вероятностью γ**.

Раздел 4. Интервальные статистические оценки

Пусть k — суммарное число успехов в n испытаниях. Для k справедливо представление: $k = X_1 + \ldots + X_n$.

Для построения доверительного интервала используется следующая **статистика**:

$$T = \frac{k - np}{\sqrt{np(1-p)}}.$$

Раздел 4. Интервальные статистические оценки

Теорема. Пусть $\hat{p} = k/n$ — точечная оценка вероятности p. Приближенной интервальной оценкой для вероятности p с доверительной вероятностью γ является интервал

$$\left(\hat{p} - z\alpha_{/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z\alpha_{/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

Доказательство.

(Используем **предельную теорему Муавра – Лапласа).**

Интегральная теорема Муавра-Лапласа. Пусть k — число успехов в n испытаниях, проводимых по схеме Бернулли с вероятностью успеха p в одном испытании, $0 \leq k_1 < k_2 \leq n$ ($k \sim Bin(n; p)$). Тогда

$$\mathrm{P}(\{k_1 \leq k \leq k_2\}) = \mathrm{P}\left(\left\{\frac{k_1 - np}{\sqrt{np(1-p)}} \leq \frac{k-np}{\sqrt{np(1-p)}} \leq \frac{k_2-np}{\sqrt{np(1-p)}}\right\}\right) \underset{n\to\infty}{\to}$$

$$\Phi\left(\frac{k_2-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1-np}{\sqrt{np(1-p)}}\right).$$

Раздел 4. Интервальные статистические оценки

Пример. Для случайно отобранных 625 жителей оказалось 125 учащихся. Требуется найти приближенный доверительный интервал для доли p учащихся среди всех жителей города с доверительной $\gamma = 0{,}95$.

Решение.

Интервал предсказания

Раздел 4. Интервальные статистические оценки

Пусть $X_1, \ldots, X_{n+1}$ — выборка из некоторого распределения $N(\mu, \sigma^2)$.

Будем считать, что X_i – результат наблюдения X в испытании, проводимом в момент времени $i = 1, \ldots, n+1$.

Тогда имеет смысл **следующая задача**:

Построить по данным $X_1, \ldots, X_n$ **интервал предсказания** (L, H), накрывающий X_{n+1} с доверительной вероятностью γ.

Раздел 4. Интервальные статистические оценки

Используя стандартные статистики

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{и} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

построим статистику

$$T = \frac{X_{n+1} - \bar{X}}{s\sqrt{1+\frac{1}{n}}} \sim t(n-1)$$

получим симметричный по вероятности интервал предсказания

$$\bar{X} - t_{\frac{\alpha}{2}}(n-1)s\sqrt{1+\frac{1}{n}} < X_{n+1} < \bar{X} + t_{\frac{\alpha}{2}}(n-1)s\sqrt{1+\frac{1}{n}}$$

Раздел 4. Интервальные статистические оценки

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

$$X_{n+1} > \bar{X} - t_{\alpha}(n-1)s\sqrt{1+\frac{1}{n}}$$

$$X_{n+1} < \bar{X} + t_{\alpha}(n-1)s\sqrt{1+\frac{1}{n}}$$

Раздел 4. Интервальные статистические оценки

Пример. Пусть A и B – цены некоторых активов, причем разность $A - B \sim N(\mu, \sigma^2)$, с неизвестными μ и σ^2. В результате 5 торгов разность $A - B$ приняла значения:

$$8, 3, -2, 7, 6.$$

Сколько нужно зарезервировать денег M, чтобы, продав на очередных торгах актив B, их хватило для покупки актива A с надежностью $0{,}95$?

Решение.

Интервальная оценка коэффициента корреляции

Раздел 4. Интервальные статистические оценки

Интервальная оценка коэффициента корреляции:

$$th\left(arth\,\widehat{\boldsymbol{\rho}}_n - \frac{1}{\sqrt{n-3}} z_{\frac{1+\boldsymbol{\gamma}}{2}}\right) < \boldsymbol{\rho} < th\left(arth\,\widehat{\boldsymbol{\rho}}_n + \frac{1}{\sqrt{n-3}} z_{\frac{1+\boldsymbol{\gamma}}{2}}\right)$$

где функция $th\,x = \frac{e^{\boldsymbol{x}} - e^{-\boldsymbol{x}}}{e^{\boldsymbol{x}} + e^{-\boldsymbol{x}}}$ (гиперболический тангенс),

$arth\,\boldsymbol{x}$ – обратный гиперболический арктангенс.

Раздел 4. Интервальные статистические оценки

Пример. По выборке из 55 наблюдений двумерной нормальной случайной величины получен выборочный коэффициент корреляции 0,6. Построить доверительный интервал для коэффициента корреляции с надежностью 94%.

Решение.

Интервальная оценка коэффициента корреляции:

$$th\left(arth\,\widehat{\boldsymbol{\rho}}_n - \frac{1}{\sqrt{n-3}} z_{\frac{1+\boldsymbol{\gamma}}{2}}\right) < \boldsymbol{\rho} < th\left(arth\,\widehat{\boldsymbol{\rho}}_n + \frac{1}{\sqrt{n-3}} z_{\frac{1+\boldsymbol{\gamma}}{2}}\right)$$

Теория вероятностей и математическая статистика

Конец лекции