УНИВЕРСИТЕТ ИТМО

Теория вероятностей и математическая статистика

Лекция 6. ***Выборки и их характеристики. Элементы теории оценок***

Генеральная и выборочная совокупности

Задачами математической статистики являются оценивание законов распределения и основных характеристик случайных величин, проверка статистических гипотез, анализ зависимостей между входными и выходными параметрами систем, прогнозирование, планирование эксперимента и т.д.

Эти и другие статистические выводы относительно свойств полной совокупности данных (**генеральной совокупности**) делают на основе некоторой специальным образом сформированной части данных

$$x_1, x_2, \ldots, x_n,$$

называемой **выборкой объема** n.

Свойства выборки

Выборка должна обладать следующими свойствами.

1. Необходимо, чтобы выборка была **репрезентативной**, т. е. достаточно полно, однородно, равномерно и равновероятно по отношению к другим возможным выборкам представляла всю генеральную совокупность.

2. Выборка должна быть **рандомизированной**, т. е. полученной случайным образом в одинаковых условиях в виде последовательности повторных независимых реализаций случайной величины X.

3. В рамках вероятностной математической модели, привлекаемой для анализа данных, выборка должна рассматриваться как реализация n-мерного случайного вектора $(X_1, X_2, \ldots, X_n)$ с взаимно независимыми и одинаково распределенными компонентами.

Генеральная и выборочная совокупности

Пример 7.1. Десять абитуриентов проходят тестирование по математике. Каждый из них может набрать от 0 до 5 баллов включительно. Пусть X_k — количество баллов, набранных k-м ($k = 1, 2, \ldots, 10$) абитуриентом.

Тогда значения 0, 1, 2, 3, 4, 5 — все возможные количества баллов, набранных одним абитуриентом, — образуют генеральную совокупность.

Выборка $X_1, X_2, X_3, \ldots, X_{10}$ -- результат тестирования 10 абитуриентов.

Реализациями выборки могут быть следующие наборы чисел: {5, 3, 0, 1, 4, 2. 5. 4. 1. 5} или {4, 4, 5, 3, 3, 1, 5, 5, 2, 5} или {3, 4, 5. 0. 1. 2. 3, 4, 5, 4} и т. д.

Вариационный ряд

Если элементы выборки расположить в порядке возрастания

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)},$$

то получим **вариационный ряд**, элементы которого называют **порядковыми статистиками**. Наименьшее значение в выборке называют **первой порядковой статистикой** $x_{(1)}$, а наибольшее значение n**-ой порядковой статистикой** $x_{(n)}$. Разность между наибольшим и наименьшим значениями называют **размахом выборки**, обозначают буквой w^* и вычисляют по формуле $w^* = x_{(n)} - x_{(1)}$.

Статистический ряд

Статистическим рядом называют систему пар чисел

$$(z_i, n_i), \quad i = 1, 2, \ldots, k,$$

где z_i различные элементы выборки, расположенные в порядке возрастания, n_i частота элемента в выборке, т. е. число повторений элемента. Обычно статистический ряд представляют в виде **таблицы**, где первая строка содержит элементы z_i, а вторая их частоты. Если в выборке нет одинаковых элементов, то статистический и вариационный ряды совпадают. По вариационному или статистическому ряду строится эмпирическая (выборочная) функция распределения $F_n^*(x)$, которая является оценкой функции распределения $F_X(x)$ случайной величины X, сформировавшей данную выборку.

Статистическое распределение выборки

Пусть изучается некоторая с. в. X. С этой целью над с. в. X производится ряд независимых опытов (наблюдений). В каждом из этих опытов величина X принимает то или иное значение.

Пусть она приняла n_1 раз значение x_1, n_2 раз — значение $x_2, \ldots,$ n_k раз — значение x_k. При этом $n_1+n_2+\ldots+n_k=n$ — объем выборки. Значения $x_1, x_2, \ldots, x_k$ называются *вариантами* с. в. X.

Вся совокупность значений с. в. X представляет собой первичный статистический материал, который подлежит дальнейшей обработке, прежде всего — упорядочению.

Операция расположения значений случайной величины (признака) по неубыванию называется *ранжированием* статистических данных. Полученная таким образом последовательность $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ значений с. в. X (где $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ и $x_{(1)} = \min\limits_{1\leqslant i\leqslant n} X_i, \ldots, x_{(n)} = \max\limits_{1\leqslant i\leqslant n} X_i$) называется *вариационным рядом*.

Статистическое распределение выборки

Числа n_i, показывающие, сколько раз встречаются варианты x_i в ряде наблюдений, называются *частотами*, а отношение их к объему выборки — *частостями* или *относительными частотами* (p_i^*), т. е.

$$p_i^* = \frac{n_i}{n},$$

где $n = \sum_{i=1}^{k} n_i$.

Перечень вариантов и соответствующих им частот или частостей называется *статистическим распределением выборки* или *статистическим рядом*.

Записывается статистическое распределение в виде таблицы. Первая строка содержит варианты, а вторая — их частоты n_i (или частости p_i^*).

Интервальный статистический ряд

В случае, когда число значений признака (с. в. X) велико или признак является непрерывным (т. е. когда с. в. X может принять любое значение в некотором интервале), составляют *интервальный статистический ряд*. В первую строку таблицы статистического распределения вписывают частичные промежутки $[x_0, x_1)$, $[x_1, x_2)$, ..., $[x_{k-1}, x_k)$, которые берут обычно одинаковыми по длине: $h = x_1 - x_0 = x_2 - x_1 = \ldots$. Для определения величины интервала (h) можно использовать формулу Стерджеса:

$$h = \frac{x_{\max} - x_{\min}}{1 + \log_2 n},$$

где $x_{\max} - x_{\min}$ — разность между наибольшим и наименьшим значениями признака, $m = 1 + \log_2 n$ — число интервалов ($\log_2 n \approx 3{,}322 \lg n$).

За начало первого интервала рекомендуется брать величину $x_{\text{нач}} = x_{\min} - \frac{h}{2}$. Во второй строчке статистического ряда вписывают количество наблюдений n_i ($i = \overline{1, k}$), попавших в каждый интервал.

Эмпирическая функция распределения

Эмпирической (статистической) функцией распределения называется функция $F_n^*(x)$, определяющая для каждого значения x частость события $\{X < x\}$:

$$F_n^*(x) = p^*\{X < x\}.$$

Для нахождения значений эмпирической функции удобно $F_n^*(x)$ записать в виде

$$F_n^*(x) = \frac{n_x}{n},$$

где n — объем выборки, n_x — число наблюдений, меньших x ($x \in \mathbb{R}$).

Очевидно, что $F_n^*(x)$ удовлетворяет тем же условиям, что и истинная функция распределения $F(x)$ (см. п. 2.3).

При увеличении числа n наблюдений (опытов) относительная частота события $\{X < x\}$ приближается к вероятности этого события

Теорема Гливенко

Теорема (Гливенко). *Эмпирическая функция распределения* $F_n^*(x)$ *при неограниченном увеличении объема выборки сходится по вероятности при любом значении* $x \in \mathbb{R}$ *к теоретической функции распределения* $F_X(x)$ *генеральной совокупности.*

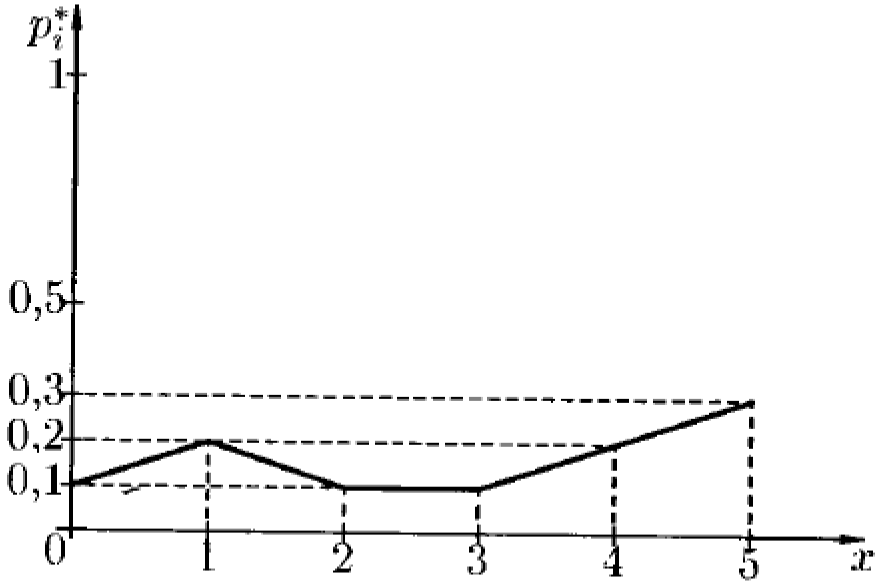
Таким образом, при большом объеме выборки эмпирическая функция распределения $F_n^*(x)$ является достаточно точным приближением для неизвестной заранее теоретической функции распределения $F_X(x)$.

Графическое изображение статистического распределения

Статистическое распределение изображается графически (для наглядности) в виде так называемых полигона и гистограммы. Полигон, как правило, служит для изображения дискретного (т. е. варианты отличаются на постоянную величину) статистического ряда.

Полигоном частот называют ломаную, отрезки которой соединяют точки с координатами $(x_1, n_1), (x_2, n_2), \ldots, (x_k, n_k)$; *полигоном частостей* — с координатами $(x_1, p_1^*), (x_2, p_2^*), \ldots, (x_k, p_k^*)$.

Варианты (x_i) откладываются на оси абсцисс, а частоты и, соответственно, частости — на оси ординат.

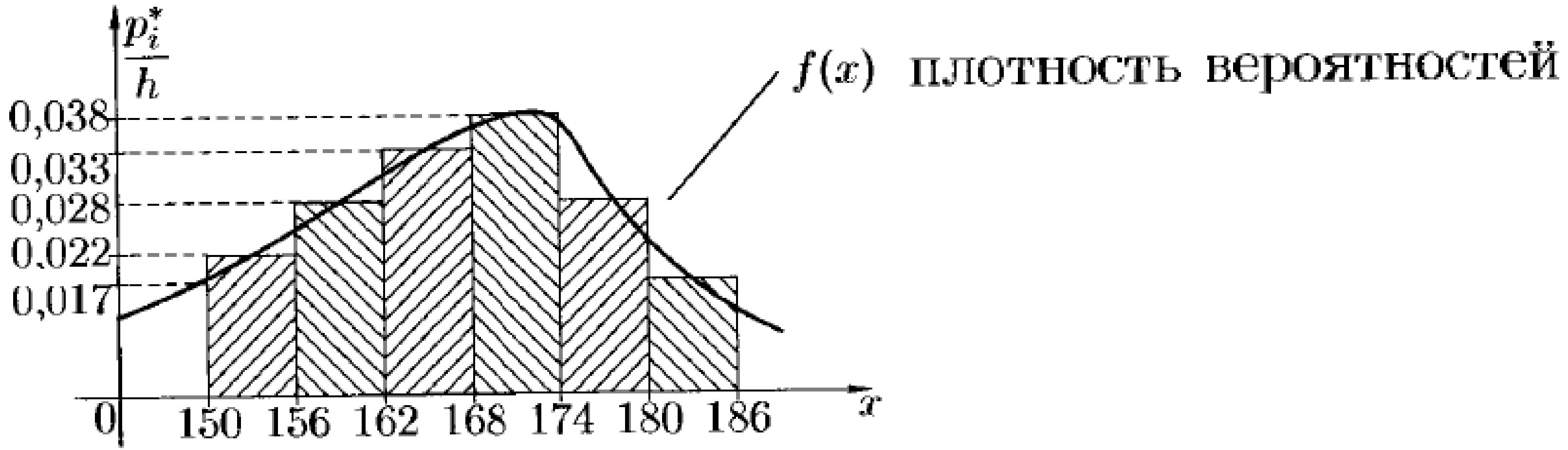


Графическое изображение статистического распределения

Для непрерывно распределенного признака (т. е. варианты могут отличаться один от другого на сколь угодно малую величину) можно построить полигон частот, взяв середины интервалов в качестве значений $x_1, x_2, \ldots, x_k$. Более употребительна так называемая гистограмма.

Гистограммой частот (*частостей*) называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длины h, а высоты равны отношению $\frac{n_i}{h}$ — плотность частоты ($\frac{p_i^*}{h}$ или $\frac{n_i}{n \cdot h}$ — плотности частости).

Очевидно, площадь гистограммы частот равна объему выборки, а площадь гистограммы частостей равна единице.



Числовые характеристики статистического распределения

Пусть статистическое распределение выборки объема n имеет вид:

x_i	x_1	x_2	x_3	$\ldots$	x_k
n_i	n_1	n_2	n_3	$\ldots$	n_k

Выборочным средним $\overline{x}_{\text{в}}$ называется среднее арифметическое всех значений выборки:

$$\overline{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i.$$

Выборочное среднее можно записать и так: $$\overline{x}_{\text{в}} = \sum_{i=1}^{k} x_i \cdot p_i^*,$$

где $p_i^* = \dfrac{n_i}{n}$ — частость. Для обозначения выборочного среднего используют следующие символы: $\overline{x}$, $M^*(X)$, m_x^*.

Отметим, что в случае интервального статистического ряда в качестве x_i берут середины его интервалов, а n_i соответствующие им частоты.

Числовые характеристики статистического распределения

Выборочной дисперсией $D_в$ называется среднее арифметическое квадратов отклонений значений выборки от выборочной средней $\overline{x}_в$, т. е.

$$D_в = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x}_в)^2 \cdot n_i$$

или, что то же самое.

$$D_в = \sum_{i=1}^{k}(x_i - \overline{x}_в)^2 \cdot p_i^*.$$

Можно показать. что $D_в$ может быть подсчитана также по формуле:

$$D_в = \frac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot n_1 - (\overline{x}_в)^2,$$ т. е.

$$D_в = \overline{x^2} - (\overline{x})^2,$$

здесь $\overline{x} = \overline{x}_в$.

Числовые характеристики статистического распределения

Выборочное среднее квадратическое отклонение выборки определяется формулой

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}}.$$

Особенность выборочного с. к. о. ($\sigma_{\text{в}}$) состоит в том, что оно измеряется в тех же единицах, что и изучаемый признак.

При решении практических задач используется и величина

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{k} (x_i - \overline{x}_{\text{в}})^2 \cdot n_i,$$

т. е.

$$S^2 = \frac{n}{n-1} D_{\text{в}},$$

которая называется *исправленной выборочной дисперсией*

Величина $S = \sqrt{S^2}$ называется *исправленным выборочным средним квадратическим отклонением.*

Для непрерывно распределенного признака формулы для выборочных средних будут такими же, но за значения $x_1, x_2, \ldots, x_k$ надо брать не концы промежутков $[x_0, x_1), [x_1, x_2), \ldots$ а их середины

Числовые характеристики статистического распределения

Размахом вариации называется число $R = x_{(n)} - x_{(1)}$, где $x_{(1)} = \min\limits_{1 \leqslant k \leqslant n} x_k$, $x_{(n)} = \max\limits_{1 \leqslant k \leqslant n} x_k$ или $R = x_{\max} - x_{\min}$, где $x_{\max}$ — наибольший, $x_{\min}$ — наименьший вариант ряда.

Модой M_o^* вариационного ряда называется вариант, имеющий наибольшую частоту.

Медианой M_e^* вариационного ряда называется значение признака (с. в. X), приходящееся на середину ряда.

Если $n = 2k$ (т. е. ряд $x_{(1)}, x_{(2)}, \ldots, x_{(k)}, x_{(k+1)}, \ldots, x_{(2k)}$ имеет четное число членов), то $M_e^* = \dfrac{x_{(k)} + x_{(k+1)}}{2}$; если $n = 2k + 1$, то $M_e^* = x_{(k+1)}$.

Понятие оценки параметров

Пусть изучается случайная величина X с законом распределения, зависящим от одного или нескольких параметров. *Например*, это параметр a в распределении Пуассона $\left(P\{X=m\}=\dfrac{a^m\cdot e^{-a}}{m!}\right)$ или параметры a и σ для нормального закона распределения.

Требуется по выборке $X_1, X_2, \dots, X_n$, полученной в результате n наблюдений (опытов), оценить неизвестный параметр θ.

Напомним, что $X_1, X_2, \dots, X_n$ — случайные величины: X_1 — результат первого наблюдения, X_2 — второго и т. д., причем с. в. X_i, $i = 1, 2, \dots, n$, имеют такое же распределение, что и с. в. X; конкретная выборка $x_1, x_2, \dots, x_n$ — это значения (реализация) независимых с. в. $X_1, X_2, \dots, X_n$.

Статистической оценкой $\tilde{\theta}_n$ (далее просто — *оценкой* $\tilde{\theta}$) параметра θ теоретического распределения называют его приближенное значение, зависящее от данных выбора.

Понятие оценки параметров

Очевидно, что оценка $\tilde{\theta}$ есть значение некоторой функции результатов наблюдений над случайной величиной, т. е.

$$\tilde{\theta} = \tilde{\theta}(X_1, X_2, \ldots, X_n)$$

Функцию результатов наблюдений (т. е. функцию выборки) называют *статистикой*.

Можно сказать, что *оценка* $\tilde{\theta}$ параметра θ есть статистика, которая в определенном смысле близка к истинному значению θ.

Так, $F^*(x)$ есть оценка $F_X(x)$, гистограмма — плотности $f(x)$.

Оценка $\tilde{\theta}$ является случайной величиной, так как является функцией независимых с. в. $X_1, X_2, \ldots, X_n$; если произвести другую выборку, то функция примет, вообще говоря, другое значение.

Если число опытов (наблюдений) невелико, то замена неизвестного параметра θ его оценкой $\tilde{\theta}$, *например* математического ожидания средним арифметическим, приводит к ошибке. Это ошибка в среднем тем больше, чем меньше число опытов.

Свойства статистических оценок

Качество оценки определяют, проверяя, обладает ли она свойствами несмещенности, состоятельности, эффективности.

Оценка $\widetilde{\theta}$ параметра θ называется *несмещенной*, если $M\widetilde{\theta} = \theta$.

Если $M\widetilde{\theta} \neq \theta$, то оценка $\widetilde{\theta}$ называется *смещенной*.

Чтобы оценка $\widetilde{\theta}$ не давала систематической ошибки (ошибки одного знака) в сторону завышения ($M\widetilde{\theta} > \theta$) или занижения ($M\widetilde{\theta} < \theta$), надо потребовать, чтобы «математическое ожидание оценки было равно оцениваемому параметру».

Если $M\widetilde{\theta}_n \to \theta$, то оценка $\widetilde{\theta}_n$ называется *асимптотически несмещенной*.

Требование несмещенности особенно важно при малом числе наблюдений (опытов).

Свойства статистических оценок

Оценка $\widetilde{\theta}_n$ параметра θ называется *состоятельной*, если она сходится по вероятности к оцениваемому параметру:

$$\widetilde{\theta}_n \xrightarrow[n\to\infty]{p} \theta,$$

т. е. для любого $\varepsilon > 0$ выполнено

$$\lim_{n\to\infty} p\left\{|\widetilde{\theta}_n - \theta| < \varepsilon\right\} = 1.$$

Это означает, что с увеличением объема выборки мы все ближе приближаемся к истинному значению параметра θ, т. е. практически достоверно $\widetilde{\theta}_n \approx \theta$.

Свойство состоятельности обязательно для любого правила оценивания (несостоятельные оценки не используются).

Свойства статистических оценок

Состоятельность оценки $\widetilde{\theta}_n$ часто может быть установлена с помощью следующей теоремы.

Теорема Если оценка $\widetilde{\theta}_n$ параметра θ является несмещенной и $D\widetilde{\theta}_n \to 0$ при $n \to \infty$, то $\widetilde{\theta}_n$ — состоятельная оценка.

❑ Запишем неравенство Чебышева для с. в. $\widetilde{\theta}_n$ для любого $\varepsilon > 0$:

$$P\left(|\widetilde{\theta}_n - \theta| < \varepsilon\right) \geqslant 1 - \frac{D\widetilde{\theta}_n}{\varepsilon^2}.$$

Так как по условию $\lim\limits_{n\to\infty} D\widetilde{\theta}_n = 0$, то $\lim\limits_{n\to\infty} P\left(|\widetilde{\theta}_n - \theta| < \varepsilon\right) \geqslant 1$. Но вероятность любого события не превышает 1 и, следовательно,

$$P\left(|\widetilde{\theta}_n - \theta| < \varepsilon\right) = 1,$$

т. е. $\widetilde{\theta}_n$ — состоятельная оценка параметра θ. ■

Свойства статистических оценок

Несмещенная оценка $\widetilde{\theta}_n$ параметра θ называется *эффективной*, если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра θ, т. е. оценка $\widetilde{\theta}_n$ эффективна, если ее дисперсия минимальна.

Эффективную оценку в ряде случаев можно найти, используя *неравенство Рао–Крамера*:

$$D\widetilde{\theta}_n \geqslant \frac{1}{n \cdot I},$$

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

$$I = M\left[\frac{\partial}{\partial\theta}\ln p(X,\theta)\right]^2 = \sum_{i=1}^{m}\left[\frac{p'_\theta(x_i,\theta)}{p(x_i,\theta)}\right]^2 \cdot p(x_i,\theta),$$

где $p(x,\theta) = p\{X = x\}$, а в непрерывном — формулой

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

где $f(x,\theta)$ — плотность распределения н. с. в. X.

Свойства статистических оценок

Эффективность оценки определяется отношением

$$\operatorname{eff} \tilde{\theta}_n = \frac{D\tilde{\theta}_n^{\text{э}}}{D\tilde{\theta}_n},$$

где $\tilde{\theta}_n^{\text{э}}$ — эффективная оценка. Чем ближе $\operatorname{eff} \tilde{\theta}_n$ к 1, тем эффективнее оценка $\tilde{\theta}_n$. Если $\operatorname{eff} \tilde{\theta}_n \to 1$ при $n \to \infty$, то оценка называется *асимптотически эффективной*.

Отметим, что на практике не всегда удается удовлетворить всем перечисленным выше требованиям (несмещенность, состоятельность, эффективность), и поэтому приходится довольствоваться оценками, не обладающими сразу всеми тремя свойствами. Все же три свойства, как правило, выделяют оценку однозначно.

Точечные оценки математического ожидания и дисперсии

Пусть изучается с. в. X с математическим ожиданием $a = MX$ и дисперсией DX; оба параметра неизвестны.

Статистика, используемая в качестве приближенного значения неизвестного параметра генеральной совокупности, называется ее *точечной оценкой*. То есть точечная оценка характеристики генеральной совокупности — это число, определяемое по выборке.

Пусть $x_1, x_2, \ldots, x_n$ — выборка, полученная в результате проведения n независимых наблюдений за с. в. X. Чтобы подчеркнуть случайный характер величин $x_1, x_2, \ldots, x_n$, перепишем их в виде $X_1, X_2, \ldots, X_n$, т. е. под X_i будем понимать значение с. в. X в i-м опыте. Случайные величины $X_1, X_2, \ldots, X_n$ можно рассматривать как n независимых «экземпляров» величины X. Поэтому $MX_1 = MX_2 = \ldots = MX_n = MX = a$, $DX_1 = DX_2 = \ldots = DX_n = DX$.

Точечные оценки математического ожидания и дисперсии

Теорема Пусть $X_1, X_2, \ldots, X_n$ — выборка из генеральной совокупности и $MX_i = MX = a$, $DX_i = DX$ $(i = \overline{1, n})$. Тогда выборочное среднее $\overline{X}_в = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ — несмещенная и состоятельная оценка математического ожидания MX.

❑ Найдем м. о. оценки $\overline{X}_в$:

$$M\overline{X}_в = M\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}M\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} MX_i = \frac{1}{n}\cdot n \cdot a.$$

Отсюда по определению получаем, что $\overline{X}_в$ — несмещенная оценка MX.

Точечные оценки математического ожидания и дисперсии

Далее, согласно теореме Чебышева , для любого $\varepsilon > 0$ имеет

место равенство

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} MX_i\right| < \varepsilon\right\} = 1,$$

которое, согласно условию теоремы, можно переписать так:

$$\lim_{n\to\infty} P\left\{|\overline{X}_{\text{в}} - MX| < \varepsilon\right\} = 1$$

или, что то же самое, $\lim\limits_{n\to\infty} p\left\{|\overline{\theta} - \theta| < \varepsilon\right\} = 1$. Согласно определению получаем, что $\overline{X}_{\text{в}}$ — состоятельная оценка MX. ■

Точечные оценки математического ожидания и дисперсии

Можно показать, что при нормальном распределении с. в. X эта оценка, т. е. $\overline{X}_в$, будет и эффективной. На практике во всех случаях в качестве оценки математического ожидания используется среднее арифметическое, т. е. $\overline{X}_в$.

В статистике оценку математического ожидания принято обозначать через $\overline{X}$ или $\overline{X}_в$, а не $\widetilde{X}$.

Можно показать, что

$$MD_в = \frac{n-1}{n} DX.$$

Из равенства следует, что $MD_в \neq DX$, т. е. *выборочная дисперсия является смещенной оценкой дисперсии* DX. Поэтому выборочную дисперсию исправляют, умножив ее на $\frac{n}{n-1}$, получая формулу

$$S^2 = \frac{n}{n-1} D_в$$

S^2 называется *исправленной выборочной дисперсией*

Точечные оценки математического ожидания и дисперсии

Теорема Пусть $X_1, X_2, \ldots, X_n$ — выборка из генеральной совокупности и $MX_i = MX = a$, $DX_i = DX$ $(i = \overline{1,n})$. Тогда исправленная выборочная дисперсия $S^2 = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2 = \dfrac{n}{n-1}\cdot D_{\text{в}}$ — несмещенная состоятельная оценка дисперсии DX.

☐ Примем без доказательства состоятельность оценки S^2. Докажем ее несмещенность.

Имеем

$$MS^2 = M\left(\frac{n}{n-1}D_{\text{в}}\right) = \frac{n}{n-1}\cdot MD_{\text{в}} = \frac{n}{n-1}\cdot\frac{n-1}{n}DX = DX,$$

т. е. $MS^2 = DX$. Отсюда по определению получаем, что S^2 — несмещенная оценка DX. ■

Точечные оценки математического ожидания и дисперсии

Отметим, что при больших значениях n разница между $D_{\text{в}}$ и S^2 очень мала и они практически равны, поэтому оценку S^2 используют для оценки дисперсии при малых выборках, обычно при $n \leqslant 30$.

Имеют место следующие теоремы.

Теорема Относительная частота $\dfrac{n_A}{n}$ появления события A в n независимых испытаниях является несмещенной состоятельной и эффективной оценкой неизвестной вероятности $p = P(A)$ этого события (p — вероятность наступления события A в каждом испытании).

Теорема Эмпирическая функция распределения выборки $F^*(x)$ является несмещенной состоятельной оценкой функции распределения $F(x)$ случайной величины X.

Методы нахождения точечных оценок

Метод моментов

Метод моментов для нахождения точечных оценок неизвестных параметров заданного распределения состоит *в приравнивании теоретических моментов распределения соответствующим эмпирическим моментам*, найденных по выборке.

Так, если распределение зависит от одного параметра θ (*например*, задан вид плотности распределения $f(x, \theta)$), то для нахождения его оценки надо решить относительно θ одно уравнение:

$$MX = \overline{X}_{\text{в}}$$

$\left(MX = \int\limits_{-\infty}^{\infty} x \cdot f(x,\theta)\, dx = \varphi(\theta)\right.$ есть функция от θ).

Если распределение зависит от двух параметров (например, вид плотности распределения $f(x, \theta_1, \theta_2)$) — надо решить относительно θ_1 и θ_2 систему уравнений:

$$\begin{cases} MX = \overline{X}_{\text{в}}, \\ DX = D_{\text{в}}. \end{cases}$$

Метод моментов

И, наконец, если надо оценить n параметров $\theta_1, \theta_2, \ldots, \theta_n$ — надо решить одну из систем вида:

$$\begin{cases} MX = \frac{1}{n}\sum_{i=1}^{n} X_i, \\ MX^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \\ \ldots \\ MX^k = \frac{1}{n}\sum_{i=1}^{n} X_i^k; \end{cases} \quad \text{или} \quad \begin{cases} MX = \overline{X}, \\ DX = D_{\text{в}}, \\ \ldots \\ M(X - MX)^k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_{\text{в}})^k. \end{cases}$$

Метод моментов является наиболее простым методом оценки параметров. Он был предложен в 1894 г. Пирсоном. Оценки метода моментов обычно состоятельны, однако их эффективность часто значительно меньше единицы.

Метод моментов

Пример Найти оценки параметров нормального распределения с. в. X методом моментов.

❍ Требуется по выборке $x_1, x_2, \ldots, x_n$ найти точечные оценки неизвестных параметров $a = MX = \theta_1$ и $\sigma^2 = DX = \theta_2$.

По методу моментов приравниваем их, соответственно, к выборочному среднему и выборочной дисперсии ($\alpha_1 = MX$ — начальный момент I порядка, $\mu_2 = DX$ — центральный момент II порядка). Получаем

$$\begin{cases} MX = \overline{x}_\text{в}, \\ DX = D_\text{в}, \end{cases}$$

т. е.

$$\begin{cases} a = \overline{x}_\text{в}, \\ \sigma^2 = D_\text{в}. \end{cases}$$

Итак, искомые оценки параметров нормального распределения: $\tilde{\theta}_1 = x_\text{в}$ и $\tilde{\theta}_2 = \sqrt{D_\text{в}}$. ●

Метод максимального правдоподобия

Пусть $x_1, x_2, \ldots, x_n$ — выборка, полученная в результате проведения n независимых наблюдений за с. в. X. И пусть вид закона распределения величины X, например, вид плотности $f(x, \theta)$, известен, но неизвестен параметр θ, которым определяется этот закон. Требуется по выборке оценить параметр θ.

В основе метода максимального правдоподобия (ММП), предложенного Р. Фишером, лежит понятие функции правдоподобия.

Метод максимального правдоподобия

Функцией правдоподобия, построенной по выборке $x_1, x_2, \ldots, x_n$, называется функция аргумента θ вида

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

или

$$L(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta),$$

где $f(x, \theta)$ — плотность распределения с. в. X в случае, если X — непрерывная. Если X — дискретная с. в., то функция правдоподобия имеет вид

$$L(x, \theta) = p(x_1, \theta) \cdot p(x_2, \theta) \cdot \ldots \cdot p(x_n, \theta) = \prod_{i=1}^{n} p(x_i, \theta),$$

где $p(x_i, \theta) = p\{X = x_i, \theta\}$.

Метод максимального правдоподобия

Из определения следует, что чем больше значение функции $L(x, \theta)$, тем более вероятно (правдоподобнее) появление (при фиксированном θ) в результате наблюдений чисел $x_1, x_2, \ldots, x_n$.

За точечную оценку параметра θ, согласно ММП, берут такое его значение θ, при котором функция правдоподобия достигает максимума.

Эта оценка, называемая *оценкой максимального правдоподобия*, является решением уравнения

$$\frac{dL(x, \theta)}{d\theta} = 0.$$

Так как функции $L(x, \theta)$ и $\ln L(x, \theta)$ достигают максимума при одном и том же значении θ, то вместо отыскания максимума функции $L(x, \theta)$ ищут (что проще) максимум функции $\ln L(x, \theta)$.

Метод максимального правдоподобия

Таким образом, для нахождения оценки максимального правдоподобия надо:

1. решить уравнение правдоподобия

$$\frac{d(\ln L(x,\theta))}{d\theta} = 0;$$

2. отобрать то решение, которое обращает функцию $\ln L(x,\theta)$ в максимум (удобно использовать вторую производную: если

$$\left.\frac{d^2 \ln L(x,\theta)}{d\theta}\right|_{\theta=\widetilde{\theta}} < 0,$$

то $\theta = \widetilde{\theta}$ — точка максимума).

Если оценке подлежат несколько параметров $\theta_1, \theta_2, \ldots, \theta_n$ распределения, то оценки $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_n$ определяются решением системы уравнений правдоподобия:

$$\begin{cases} \dfrac{\partial(\ln L)}{\partial\theta_1} = 0, \\ \ldots, \\ \dfrac{\partial(\ln L)}{\partial\theta_n} = 0. \end{cases}$$

Метод моментов для нахождения точечных оценок

Задача Найти оценку параметра a распределения Пуассона методом максимального правдоподобия.

- В данном случае $p\{X = m\} = \dfrac{a^m \cdot e^{-a}}{m!}$. Поэтому

$$p(x_i, \theta) = p\{X = x_i, \theta\} = \frac{\theta^{x_i} \cdot e^{-\theta}}{x_i!}$$

при $x_i \in \mathbb{N}$. Составляем функцию правдоподобия (для дискретной с. в. X):

$$L(x, \theta) = \frac{\theta^{x_1} \cdot e^{-\theta}}{x_1!} \cdot \frac{\theta^{x_2} \cdot e^{-\theta}}{x_2!} \cdot \ldots \cdot \frac{\theta^{x_n} \cdot e^{-\theta}}{x_n!} = e^{-\theta n} \cdot \theta^{\sum\limits_{i=1}^{n} x_i} \cdot \frac{1}{x_1! \cdot \ldots \cdot x_n!}.$$

Тогда

$$\ln L(x, \theta) = -n \cdot \theta + \sum_{i=1}^{n} x_i \cdot \ln \theta - \ln(x_1! \cdot x_2! \cdot \ldots \cdot x_n!)$$

и

$$\frac{d \ln L(x, \theta)}{d\theta} = -n + \frac{1}{\theta} \cdot \sum_{i=1}^{n} x_i.$$

Метод моментов для нахождения точечных оценок

Задача Найти оценку параметра a распределения Пуассона методом максимального правдоподобия.

Уравнение правдоподобия имеет вид:

$$\left(-n + \frac{1}{\theta} \cdot \sum_{i=1}^{n} x_i\right)\Bigg|_{\theta=\tilde{\theta}} = 0.$$

Отсюда находим

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}_{\text{в}}.$$

А так как

$$\frac{d^2 \ln L(x,\theta)}{d\theta^2}\Bigg|_{\theta=\tilde{\theta}} = -\frac{1}{\theta^2}\sum_{i=1}^{n} x_i < 0,$$

то оценка $\tilde{\theta} = \overline{x}_{\text{в}}$ является оценкой максимального правдоподобия.
Итак, $\tilde{\theta} = \tilde{a} = \overline{x}_{\text{в}}$.

Метод наименьших квадратов

Метод нахождения оценки $\tilde{\theta}$ неизвестного параметра θ, основанный на *минимизации суммы квадратов отклонений выборочных данных* от определяемой (искомой) оценки θ, называется *методом наименьших квадратов* (коротко: МНК).

Другими словами, в МНК требуется найти такое значение $\tilde{\theta}$, которое минимизировало бы сумму

$$F(\theta) = \sum_{i=1}^{n} (X_i - \theta)^2 \to \min .$$

Отметим, что МНК является наиболее простым методом нахождения оценок параметра θ.

Метод наименьших квадратов

Пример Найти оценку параметра a распределения Пуассона методом наименьших квадратов.

- Найдем точку минимума функции $F(\theta)=\sum_{i=1}^{n}(X_i-\theta)^2$:

$$F'(\theta)=\left(F=\sum_{i=1}^{n}(X_i-\theta)^2\right)'_\theta=\sum_{i=1}^{n}2(X_i-\theta)\cdot(-1);$$

из уравнения $F'(\theta)=0$ находим критическую точку: $-2\sum_{i=1}^{n}(X_i-\theta)=0$,

т. е. $\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\theta=0$, т. е. $\sum_{i=1}^{n}X_i=n\theta$, $\theta_{\text{кр}}=\frac{1}{n}\sum_{i=1}^{n}X_i$.

Метод наименьших квадратов

Пример Найти оценку параметра a распределения Пуассона методом наименьших квадратов.

А так как $$F''(\theta_{\text{кр}}) = \left(-2\sum_{i=1}^{n}(X_i - \theta)\right)'_\theta = -2\sum_{i=1}^{n}(-1) = 2n > 0$$

при любом значении θ, то $\theta_{\text{кр}} = \frac{1}{n}\sum_{i=1}^{n} X_i$ — точка минимума функции $F(\theta)$. Таким образом, оценкой параметра a в распределении Пуассона $P(m;a) = \frac{a^m \cdot e^{-a}}{m!}$, $m = 0, 1, 2, \ldots$ согласно МНК, является $\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Можно доказать, что:

$$M(\tilde{\theta}) = \theta = a, \quad D(\tilde{\theta}) = \frac{\theta}{n}.$$

●

Интервальное оценивание параметров

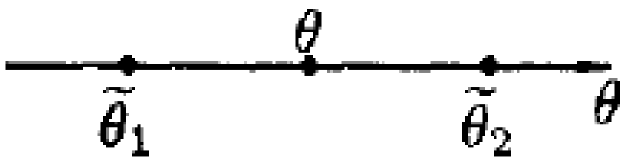
Точечные оценки неизвестного параметра θ хороши в качестве первоначальных результатов обработки наблюдений. Их недостаток в том, что неизвестно, с какой точностью они дают оцениваемый параметр.

Для выборок небольшого объема вопрос о точности оценок очень существенен, так как между θ и $\widetilde{\theta}$ может быть большое расхождение в этом случае. Кроме того, при решении практических задач часто требуется определить и надежность этих оценок. Тогда и возникает задача о приближении параметра θ не одним числом, а целым интервалом $(\widetilde{\theta}_1, \widetilde{\theta}_2)$.

Оценка неизвестного параметра называется *интервальной*, если она определяется двумя числами — концами интервала.

Задачу интервального оценивания можно сформулировать так: по данным выборки построить числовой интервал $(\widetilde{\theta}_1, \widetilde{\theta}_2)$, относительно которого с заранее выбранной вероятностью γ можно сказать, что внутри этого интервала находится точное значение оцениваемого параметра

Интервальное оценивание параметров



Интервал $(\widetilde{\theta}_1, \widetilde{\theta}_2)$, накрывающий с вероятностью γ истинное значение параметра θ, называется *доверительным интервалом*, а вероятность γ — *надежностью оценки* или *доверительной вероятностью*.

Очень часто (но не всегда) доверительный интервал выбирается симметричным относительно несмещенной точечной оценки $\widetilde{\theta}$, т. е. выбирается интервал вида $(\widetilde{\theta} - \varepsilon, \widetilde{\theta} + \varepsilon)$ такой, что

$$p\left\{\theta \in (\widetilde{\theta} - \varepsilon, \widetilde{\theta} + \varepsilon)\right\} = p\left\{|\theta - \widetilde{\theta}| < \varepsilon\right\} = \gamma.$$

Число $\varepsilon > 0$ характеризует точность оценки: чем меньше разность $|\theta - \widetilde{\theta}|$, тем точнее оценка.

Интервальное оценивание параметров

Величина γ выбирается заранее, ее выбор зависит от конкретно решаемой задачи. Так, степень доверия авиапассажира к надежности самолета, очевидно, должна быть выше степени доверия покупателя к надежности телевизора, лампочки, игрушки... Надежность γ принято выбирать равной 0,9; 0,95; 0,99 или 0,999. Тогда практически достоверно нахождение параметра θ в доверительном интервале $(\widetilde{\theta} - \varepsilon, \widetilde{\theta} + \varepsilon)$.

Доверительные интервалы для параметров нормального распределения

Построим доверительные интервалы для параметров нормального распределения, т. е. когда выборка производится из генеральной совокупности, имеющей нормальное распределение с параметрами a и σ^2.

Доверительный интервал для математического ожидания при известной дисперсии

Пусть с. в. $X \sim N(a, \sigma)$; σ — известна, доверительная вероятность (надежность) γ — задана.

Пусть $x_1, x_2, \ldots, x_n$ — выборка, полученная в результате проведения n независимых наблюдений за с. в. X. Чтобы подчеркнуть случайный характер величин $x_1, x_2, \ldots, x_n$, перепишем их в виде $X_1, X_2, \ldots, X_n$, т. е. под X_i будем понимать значение с. в. X в i-м опыте. Случайные величины $X_1, X_2, \ldots, X_n$ — независимы, закон распределения любой из них совпадает с законом распределения с. в. X (т. е. $X_i \sim N(a, \sigma)$). А это значит, что $MX_1 = MX_2 = \ldots = MX_n = MX = a$, $DX_1 = DX_2 = \ldots = DX_n = DX$.

Выборочное среднее

$$\overline{X}_{\text{в}} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

также будет распределено по нормальному закону (примем без доказательства)

Доверительный интервал для математического ожидания при известной дисперсии

Параметры распределения $\overline{X}$ таковы: $M(\overline{X}) = a$, $D(\overline{X}) = \dfrac{\sigma^2}{n}$. Действительно,

$$M(\overline{X}) = M\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\cdot\sum_{i=1}^{n} MX_i = \frac{1}{n}\cdot\sum_{i=1}^{n} MX = MX = a,$$

$$D(\overline{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{1}{n^2}\sum_{i=1}^{n} DX = \frac{1}{n}\cdot DX = \frac{\sigma^2}{n}.$$

Таким образом, $\overline{X} \sim N\left(a, \dfrac{\sigma}{\sqrt{n}}\right)$.

Доверительный интервал для математического ожидания при известной дисперсии

Следовательно, пользуясь формулой

$$p\{|X-a|<l\}=2\Phi_0\left(\frac{l}{\sigma}\right)=2\Phi\left(\frac{l}{\sigma}\right)-1$$ можно записать

$$\gamma=p\{|\overline{X}-a|<\varepsilon\}=2\Phi_0\left(\frac{\varepsilon\cdot\sqrt{n}}{\sigma}\right)=2\Phi_0(t),$$

где $t=\frac{\varepsilon\cdot\sqrt{n}}{\sigma}$. Из последнего равенства находим

$$\varepsilon=\frac{t\cdot\sigma}{\sqrt{n}},$$

поэтому $\gamma=p\left\{|\overline{X}-a|<\frac{t\cdot\sigma}{\sqrt{n}}\right\}=2\Phi_0(t)$ или

$$p\left\{\overline{X}-t\cdot\frac{\sigma}{\sqrt{n}}<a<\overline{X}+t\cdot\frac{\sigma}{\sqrt{n}}\right\}=2\Phi_0(t)=\gamma.$$

Доверительный интервал для математического ожидания при известной дисперсии

В соответствии с определением доверительного интервала получаем, что доверительный интервал для $a = MX$ есть

$$\left(\overline{X} - t \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + t \cdot \frac{\sigma}{\sqrt{n}}\right),$$

где t определяется из уравнения

$$\Phi_0(t) = \frac{\gamma}{2}$$

(или $\Phi(t) = \frac{1+\gamma}{2}$); при заданном γ по таблице функции Лапласа находим аргумент t.

Распределение функций нормальных случайных величин

Если каждому возможному значению с. в. X по определенному правилу соответствует одно возможное значение с. в. Y, то Y называют *функцией случайного аргумента* X, записывают $Y = \varphi(X)$.

Если каждой паре возможных значений с. в. X и Y по определенному правилу соответствует одно возможное значение с. в. Z, то Z называют *функцией двух случайных аргументов* X *и* Y, записывают $Z = \varphi(X, Y)$.

Рассмотрим распределение некоторых с. в., представляющих функции нормальных величин, используемые в математической статистике.

Распределение χ^2 (хи-квадрат или Пирсона)

Распределением χ_n^2 *с* n *степенями свободы* называется распределение суммы квадратов n независимых стандартных случайных величин, т. е.

$$\chi_n^2 = \sum_{i=1}^{n} X_i^2, \quad \text{где} \quad X_i \sim N(0,1), \quad i = 1, 2, \ldots, n.$$

Плотность вероятности с. в. χ^2 зависит только от числа n, т. е. числа слагаемых. Если $n = 1$, то $\chi^2 = X^2$, где $X \sim N(0,1)$,

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

Плотность распределения с. в. $Y = X^2$ равна

$$g(y) = \frac{1}{2\sqrt{2\pi y}} e^{-\frac{y}{2}}, \quad y > 0.$$

Распределение χ^2 (хи-квадрат или Пирсона)

Плотность распределения χ_n^2 имеет вид

$$f_{\chi_n^2}(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & \text{при } x > 0, \\ 0, & \text{при } x \leqslant 0, \end{cases}$$

где $\Gamma(p) = \int\limits_0^\infty t^{p-1} e^{-t}\, dt$ — гамма-функция Эйлера ($\Gamma(p) = (p-1)!$ для целых положительных p). С возрастанием числа степеней свободы n распределение χ^2 приближается к нормальному закону распределения (при $n > 30$ распределение χ^2 практически не отличается от нормального);

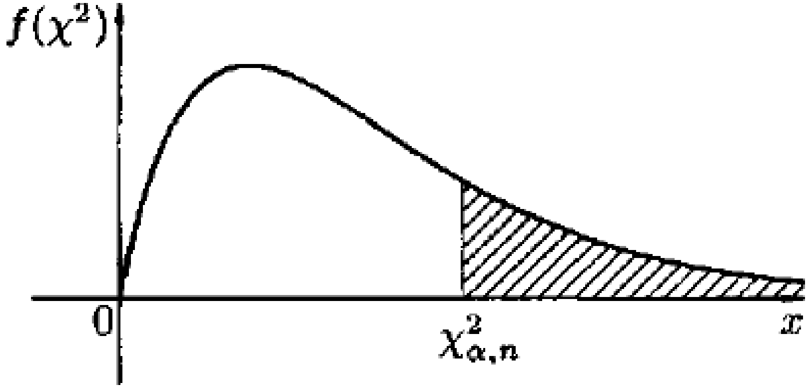
$$M\chi_n^2 = n, \quad D\chi_n^2 = 2n.$$

На практике, как правило, используют не плотность вероятности, а квантили распределения χ_n^2.

Распределение χ^2 (хи-квадрат или Пирсона)

Квантилью распределения χ_n^2, отвечающей уровню значимости α, называется такое значение $\chi_n^2 = \chi_{\alpha,n}^2$, при котором

$$P\{\chi_n^2 > \chi_{\alpha,n}^2\} = \int\limits_{\chi_{\alpha,n}^2}^{\infty} f_{\chi_n^2}(x)\,dx = \alpha.$$



С геометрической точки зрения нахождение квантили $\chi_{\alpha,n}^2$ заключается в выборе такого значения $\chi_n^2 = \chi_{\alpha,n}^2$, чтобы площадь заштрихованной на рис. фигуры была равна α.

Значения квантилей приводятся в специальных таблицах-приложениях.

Для стандартного нормального распределения квантили уровня α обозначаются через $\pm u_\alpha$, причем u_α является решением уравнения $\Phi(u_\alpha) = \dfrac{1-\alpha}{2}$.

Распределение Стьюдента

Распределением Стьюдента (или *t-распределением*) с n степенями свободы называется распределение с. в.

$$T_n = \frac{Z}{\sqrt{\frac{1}{n}\chi_n^2}},$$

где $Z \sim N(0,1)$ — стандартная нормальная величина, независимая от χ_n^2-распределения.

Плотность вероятности распределения Стьюдента имеет вид

$$f_{T_n}(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}; \quad t \in (-\infty; \infty).$$

При $n \to \infty$ распределение Стьюдента приближается (уже при $n > 30$ почти совпадает) к нормальному;

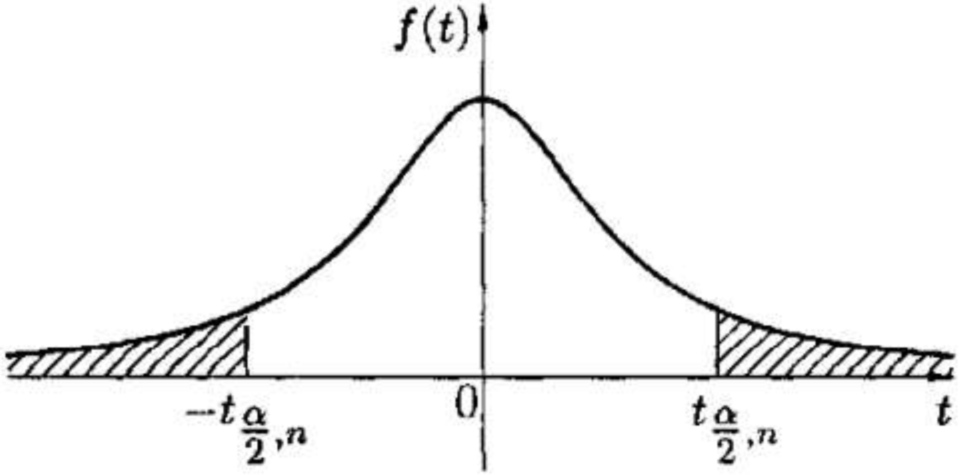
$$MT_n = 0, \qquad DT_n = \frac{n}{n-2}, \quad n > 2.$$

Распределение Стьюдента

На практике используют квантили t-распределения: такое значение $t = t_{\frac{\alpha}{2},n}$, что

$$P\{|t| > t_{\frac{\alpha}{2},n}\} = 2\int\limits_{t_{\frac{\alpha}{2},n}}^{\infty} f(t)\,dt = \alpha.$$

С геометрической точки зрения нахождение квантилей заключается в выборе такого значения $t = t_{\frac{\alpha}{2},n}$, чтобы площадь заштрихованной на рис. фигуры была равна α.



Распределение Фишера-Снедекора

Распределением Фишера–Снедекора (или *F-распределением*) с m и n *степенями свободы* называется распределение с. в.

$$F = \frac{\frac{1}{m}\chi_m^2}{\frac{1}{n}\chi_n^2},$$

где χ_m^2 и χ_n^2 — независимые с. в., имеющие χ^2-распределение соответственно с m и n степенями свободы.

При $n \to \infty$ F-распределение стремится к нормальному закону.

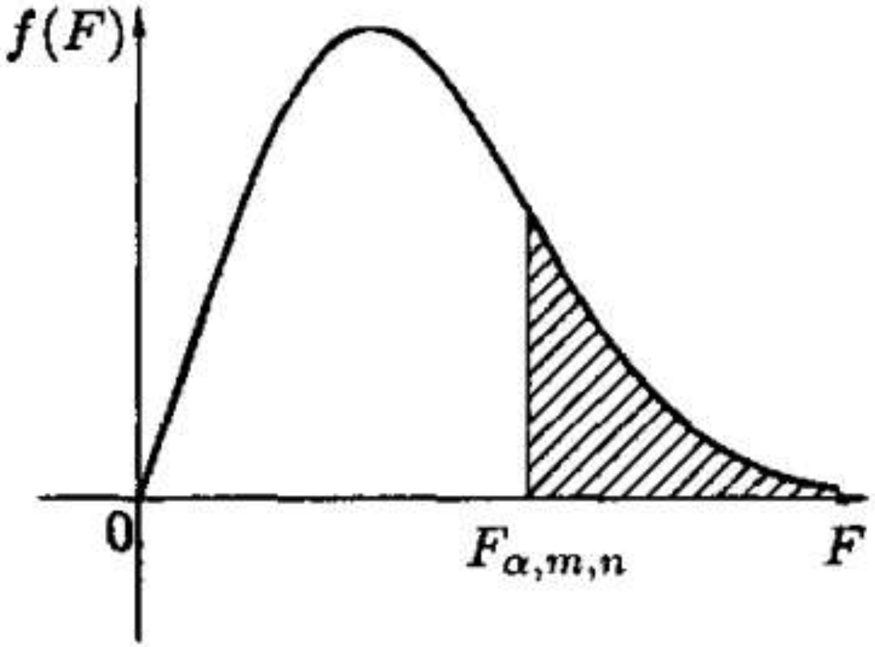
$$MF = \frac{n}{n-2}, \quad n > 2, \qquad DF = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \quad n > 4.$$

На практике обычно используют квантили распределения: такое значение $F = F_{\alpha,m,n}$, что

$$P\{F > F_{\alpha,m,n}\} = \int\limits_{F_{\alpha,m,n}}^{\infty} f(F)\,dF = \alpha.$$

Распределение Фишера-Снедекора

С геометрической точки зрения нахождение квантили заключается в выборе такого значения $F = F_{\alpha,m,n}$, чтобы площадь заштрихованной на рис. фигуры была равна α.



Доверительный интервал для математического ожидания при неизвестной дисперсии

Пусть с. в. $X \sim N(a, \sigma)$, σ — неизвестна, γ — задана. Найдем такое число ε, чтобы выполнялось соотношение $p\{\overline{X} - \varepsilon < a < \overline{X} + \varepsilon\} = \gamma$ или

$$p\{|\overline{X} - a| < \varepsilon\} = \gamma. \qquad (*)$$

Введем случайную величину

$$T = \frac{\overline{X} - a}{\frac{S}{\sqrt{n}}},$$

где S — исправленное среднее квадратическое отклонение с. в. X, вычисленное по выборке:

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

Доверительный интервал для математического ожидания при неизвестной дисперсии

Доказывается, что с. в. T имеет распределение Стьюдента с $n-1$ степенью свободы. Плотность этого распределения имеет вид:

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

где $\Gamma(p) = \int\limits_0^\infty u^{p-1} \cdot e^{-u} du$ — гамма-функция; $f_T(t, n-1)$ — четная функция.

Перейдем в левой части равенства (*) от с. в. $\overline{X}$ к с. в. T:

$$p\left\{\frac{|\overline{X}-a|}{\frac{S}{\sqrt{n}}} < \frac{\varepsilon}{\frac{S}{\sqrt{n}}}\right\} = \gamma$$

или $p\left\{|T| < \frac{\varepsilon\sqrt{n}}{S}\right\} = \gamma$ или $p\{|T| < t_\gamma\} = \gamma$, где $t_\gamma = \frac{\varepsilon \cdot \sqrt{n}}{S}$. (**)

Доверительный интервал для математического ожидания при неизвестной дисперсии

Величина t_γ находится из условия

$$p\{|T| < t_\gamma\} = \int\limits_{-t_\gamma}^{t_\gamma} f_T(t, n-1)\,dt = 2 \cdot \int\limits_{0}^{t_\gamma} f_T(t, n-1)\,dt = \gamma,$$

т. е. из равенства

$$2 \cdot \int\limits_{0}^{t_\gamma} f_T(t, n-1)\,dt = \gamma.$$

Пользуясь таблицей квантилей распределения Стьюдента находим значение t_γ в зависимости от доверительной вероятности γ и числа степеней свободы $n-1$ (t_γ — квантиль уровня $1-\gamma$).

Определив значение t_γ из равенства (**) , находим значение ε: $\varepsilon = t_\gamma \cdot \dfrac{S}{\sqrt{n}}$.

Следовательно, равенство (*) принимает вид

$$p\left\{\overline{X} - t_\gamma \cdot \frac{S}{\sqrt{n}} < a < \overline{X} + t_\gamma \cdot \frac{S}{\sqrt{n}}\right\} = \gamma.$$

Доверительный интервал для математического ожидания при неизвестной дисперсии

А это значит, что интервал

$$\left(\overline{X} - t_\gamma \cdot \frac{S}{\sqrt{n}}, \overline{X} + t_\gamma \cdot \frac{S}{\sqrt{n}}\right)$$

покрывает $a = MX$ с вероятностью γ, т. е. является доверительным интервалом для неизвестного математического ожидания с. в. X.

Доверительный интервал для среднего квадратического отклонения нормального распределения

Пусть с. в. $X \sim N(a, \sigma)$, σ — неизвестно, γ — задано. Можно показать, что *если* $MX = a$ *известно*, то доверительный интервал для среднего квадратического отклонения σ имеет вид:

$$\left(\frac{\sqrt{n} \cdot S_0}{\chi_2}, \frac{\sqrt{n} \cdot S_0}{\chi_1} \right),$$

где n — объем выборки, $S_0^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - a)^2$, а

$$\chi_1^2 = \chi^2_{\frac{1+\gamma}{2};n}; \quad \chi_2^2 = \chi^2_{\frac{1-\gamma}{2};n}$$

являются квантилями χ^2-распределения с n степенями свободы (см. п. 4.3), определяемые по таблице квантилей $\chi^2_{\alpha,n}$ распределения χ^2_n (см. приложение 3

Приложение 3. Квантили $\chi^2_{\alpha,k}$ распределения χ^2_k (k — число степеней свободы)

k	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,26
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,3	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17.7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Доверительный интервал для среднего квадратического отклонения нормального распределения

Если $a = MX$ *неизвестно*, то доверительный интервал для неизвестного σ имеет вид:

$$\left(\frac{\sqrt{n-1}\cdot S}{\chi_2}, \frac{\sqrt{n-1}\cdot S}{\chi_1}\right),$$

где n — объем выборки, $S^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(X_i - \overline{X})^2$ — исправленное среднее квадратическое отклонение, квантили

$$\chi_1^2 = \chi^2_{\frac{1+\gamma}{2};n-1}, \quad \chi_2^2 = \chi^2_{\frac{1-\gamma}{2};n-1}$$

определяются по таблице $\chi^2_{\alpha,k}$ при $k = n-1$ и $\alpha = \frac{1+\gamma}{2}$ и $\alpha = \frac{1-\gamma}{2}$ соответственно.