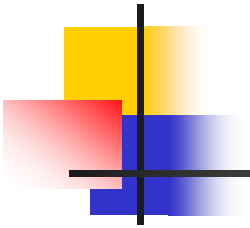
Теория вероятностей и математическая статистика



Математическая статистика

Преподаватель – доцент кафедры ВМ, к.ф.-м.н.,
Шерстнёва Анна Игоревна

Цель любой науки:

описание, объяснение и предсказание явлений действительности на основе установленных законов, что позволяет находить решение в типичных ситуациях.

Для обнаружения общей закономерности, которой подчиняется явление, необходимо многократно его наблюдать в одинаковых условиях.

Задачей математической статистики является создание методов сбора и обработки статистических данных для получения научных и практических выводов.

Выборочный метод

Пусть для получения опытных данных необходимо провести обследование соответствующих объектов.

Примеры.

1. Проверить качество выпускаемых некоторым заводом консервов.

2. Выяснить среднюю заработную плату по России.

Обычно исследуют не всю совокупность объектов, а отбирают из неё некоторое количество объектов и исследуют только их.

В этом и заключается ***выборочный метод***.

Генеральной совокупностью называют совокупность всех объектов, над которыми производят наблюдение.

Выборочной совокупностью (выборкой) называют часть отобранных из генеральной совокупности объектов.

Объёмом совокупности называют количество объектов в ней.

По выборке судят о генеральной совокупности.

Выборка должна правильно представлять генеральную совокупность, то есть быть ***репрезентативной***.

Это обеспечивается случайностью отбора и увеличением объёма выборки.

Способы отбора

1. Отбор, не требующий расчленения генеральной совокупности на части:

а) простой случайный бесповторный отбор,

б) простой случайный повторный отбор.

2. Отбор, при котором генеральная совокупность разбивается на части:

а) типический,

б) механический,

в) серийный.

Комбинированный отбор.

Статистическое распределение выборки

Что такое наблюдаемые данные?

Большой массив беспорядочно расположенных чисел.

Для работы с данными удобно их группировать.

Пример.

0 1 2 2 1 2 0 0 0 0 – выборка

Объём выборки: $n = 10$

x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3

перечень значений

частота встречаемости

относительная частота

x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3

Наблюдаемые значения x_i называют ***вариантами***.

Последовательность вариант, записанных в возрастающем порядке называют ***вариационным рядом***.

Частотой варианты называют число n_i, показывающее сколько раз встречается данная варианта.

Относительной частотой варианты называют отношение частоты к объёму выборки: $w_i = n_i / n$.

Статистическим распределением выборки называется перечень вариант и соответствующих им частот или относительных частот.

x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3

Замечания.

1. $\sum_i n_i = n$ сумма всех частот равна объёму выборки

2. $\sum_i w_i = 1$ сумма всех относительных частот равна 1

3. $p(X = x_i) \approx w_i$ относительная частота варианты даёт приближённое значение вероятности этой варианты

x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3

Если перечень вариант очень велик или одинаковые значения вариант встречаются очень редко, то такая таблица неудобна в использовании.

Поступают следующим образом:

1) разбивают весь интервал, в который попадают варианты, на частичные интервалы;
2) в верхнюю строку записывают полученные интервалы;
3) в нижнюю строку записывают частоту попадания в соответствующий интервал.

27 3,5 21,1 0,8 12,3 18 11 3,4 1,2 5,2 22 17,2 18,1 11,1 0,7 7,9 19
3,2 4,9 25,4 6,1 21,6 22,3 3,4 18,4 3,4 23,2 13,1 6,5 2,4 18,4 14,1
2,1 24,8 17,4 15,1 4,8 19,8 10,4 16,1 3,7 29,4 3,1 28,7 16,4 22,2 1,7
12,4 17 15,3 3,3 14 16,8 10,1 2,4 20 14,1 19 19,8 5,4 2,5 4,1 24,4
0,4 24,7 1,3 13,7 0,1 28 24 17,1 15 3,1 19 0,4 23,1 6,7 4,6 14,8
20,7 16,2 9,4 21,3 13,4 16,1 15,7 11,3 5,1 1,9 2,8 17 2 20,8 3,4
16,7 9,3 15,2 8,7 10,7

1) разбивают весь интервал, в который попадают варианты, на частичные интервалы;

Интервал: числа от 0 до 30.

6 интервалов: (0, 5); (5, 10); (10, 15); (15, 20); (20, 25); (25, 30)

2) в верхнюю строку записывают полученные интервалы;

3) в нижнюю строку записывают частоту попадания в соответствующий интервал.

x_i	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	30	10	15	25	15	5
w_i	0.3	0.1	0.15	0.25	0.15	0.05

Статистическое распределение можно задать также в виде последовательности интервалов и соответсвующих им частот.

На сколько интервалов разбивать выборку?

$k = 1 + 3.332 \lg n$ или $k \le 5 \lg n$

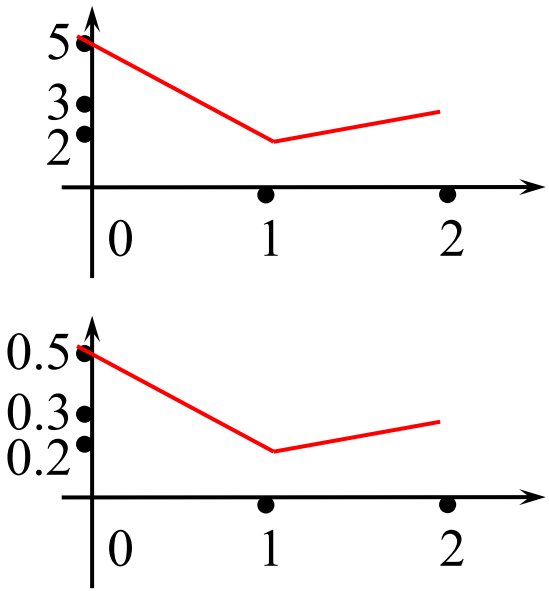
n – объём выборки

Визуализация данных

Полигоном частот называют ломаную, отрезки которой последовательно соединяют точки (x_i, n_i).

Полигоном относительных частот называют ломаную, отрезки которой последовательно соединяют точки (x_i, w_i).

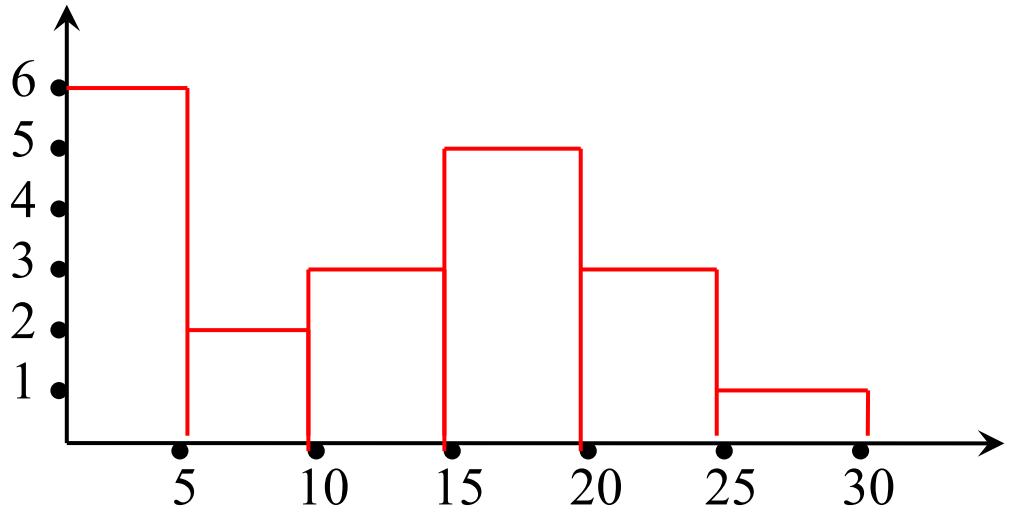
x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3



Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы, а высоты равны отношению частоты попадания в данный интервал к длине интервала.

x_i	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	30	10	15	25	15	5
n_i / h	6	2	3	5	3	1

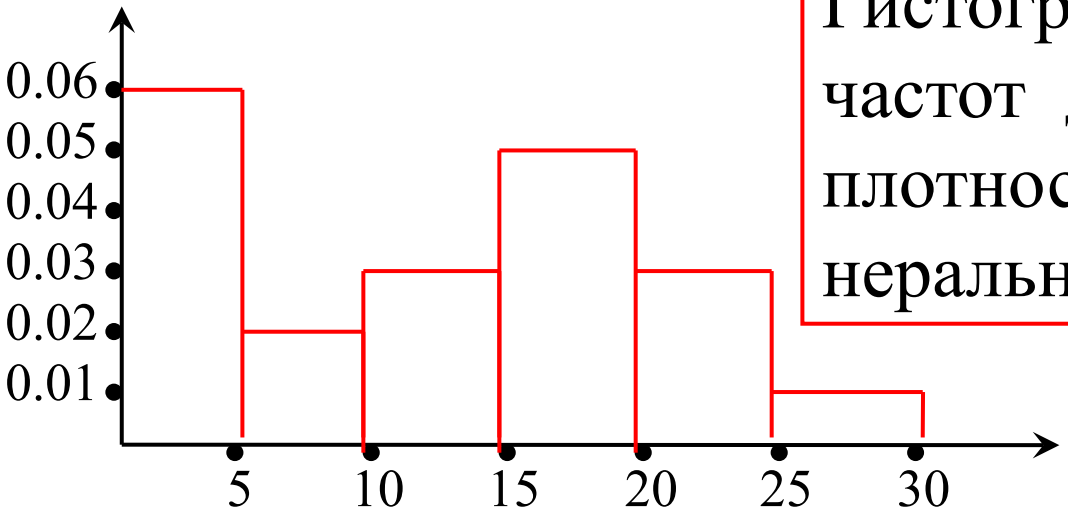
$h = 5$



Аналогично вводится понятие ***гистограммы относительных частот***.

x_i	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	30	10	15	25	15	5
w_i	0.3	0.1	0.15	0.25	0.15	0.05
w_i / h	0.06	0.02	0.03	0.05	0.03	0.01

$h = 5$



Гистограмма относительных частот даёт представление о плотности распределения генеральной совокупности.

Функция распределения

Функция распределения случайной величины X:

$$F(x) = p(X<x)$$

Теоретической функцией распределения называют функцию распределения генеральной совокупности.

Обозначим через n_x – частоту появления вариант, меньших x. Тогда n_x/n – относительная частота появления вариант, меньших x.

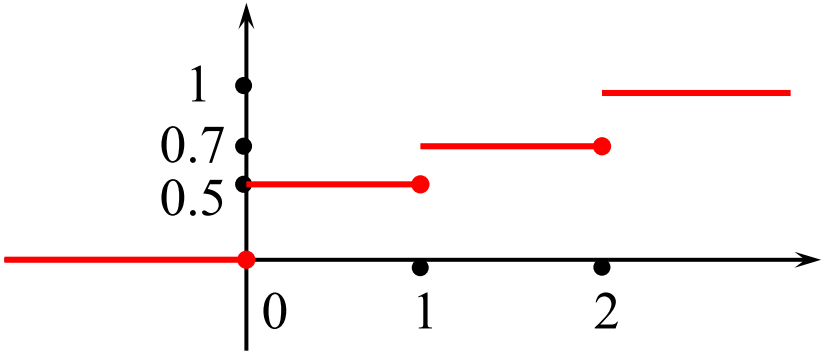
Эмпирической (выборочной) функцией распределения называют функцию

$$F^*(x) = n_x/n.$$

$F^*(x) = n_x/n.$

n_x – частота появления вариант, меньших x

x_i	0	1	2
n_i	5	2	3
w_i	0.5	0.2	0.3



Объём выборки: $n = 10$

$x = -1 \quad n_x = 0$ $\qquad$ $x = 1.5 \quad n_x = 7$

$x = 0 \quad n_x = 0$ $\qquad$ $x = 2 \quad n_x = 7$

$x = 0.5 \quad n_x = 5$ $\qquad$ $x = 3 \quad n_x = 10$

$x = 1 \quad n_x = 5$ $\qquad$ $x = 7 \quad n_x = 10$

Точечные статистические оценки параметров распределения

Записав статистическое распределение выборки и изобразив его графически, можно получить первоначальное представление о закономерностях, имеющих место в генеральной совокупности.

Как оценить числовые характеристики генеральной совокупности?

Пример.

Математическое ожидание – ?

Дисперсия – ?

Параметры распределения – ?

Выборочная характеристика

$$\Theta^* = f(x_1, x_2, ..., x_n),$$

используемая для нахождения приближённого значения неизвестной генеральной характеристики Θ, называется её ***точечной статистической оценкой***.

$$\Theta \approx \Theta^*$$

1. Несмещённость: $M(\Theta^*) = \Theta$

2. Эффективность: Θ^* имеет наименьшую дисперсию среди других оценок Θ.

3. Состоятельность: при увеличении объёма выборки Θ^* стремится по вероятности к Θ, то есть чем больше объём выборки, тем незначительнее отклонение Θ^*от Θ.

Выборочная средняя:

x_i	x_1	x_2	…
n_i	n_1	n_2	…

$$\overline{x_в} = \frac{\sum n_i x_i}{n}$$

оценка ***математического ожидания*** генеральной совокупности

x_i	0	2	3	7
n_i	6	6	2	6

Объём выборки: $n = 20$

$$\overline{x_в} = \frac{0 \cdot 6 + 2 \cdot 6 + 3 \cdot 2 + 7 \cdot 6}{20} = 3$$

Выборочная дисперсия:

x_i	x_1	x_2	…
n_i	n_1	n_2	…

$$D_в = \frac{\sum n_i (x_i - \overline{x_в})^2}{n}$$

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

оценка ***дисперсии***

x_i	0	2	3	7
n_i	6	6	2	6
$(x_i - x_в)^2$	9	1	0	16
$(x_i)^2$	0	4	9	49

$n = 20$

$\overline{x_в} = 3$

$$D_в = \frac{9 \cdot 6 + 1 \cdot 6 + 0 \cdot 2 + 16 \cdot 6}{20} = 7.8$$

$$D_в = \frac{0 \cdot 6 + 4 \cdot 6 + 9 \cdot 2 + 49 \cdot 6}{20} - 3^2 = 7.8$$

Исправленная выборочная дисперсия:

$$s^2 = \frac{n}{n-1} D_в = \frac{n}{n-1} \frac{\sum n_i (x_i - \overline{x_в})^2}{n} = \frac{\sum n_i (x_i - \overline{x_в})^2}{n-1}$$

Выборочное среднее квадратическое отклонение:

$$\sigma_в = \sqrt{D_в}$$

Исправленное выборочное среднее квадратическое отклонение:

$$s = \sqrt{s^2}$$

Мода

Для дискретной случайной величины – наиболее вероятное по сравнению с двумя соседними значение.

Как оценить моду генеральной совокупности?

– по выборке;

– наиболее часто встречающаяся варианта.

Обозначается M_0.

x_i	0	1	2
n_i	5	2	3

$M_0 = 0$

x_i	2	3	7	9	14
n_i	5	8	7	5	8

$M_0 = 3, \quad M_0 = 14$

У случайной величины может быть несколько мод.

Как оценить моду, если выборка задана интервальным рядом?

x_i	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	10	15	25	15	5

Для непрерывной случайной величины мода – это значение, при котором плотность распределения $f(x)$ достигает максимума.

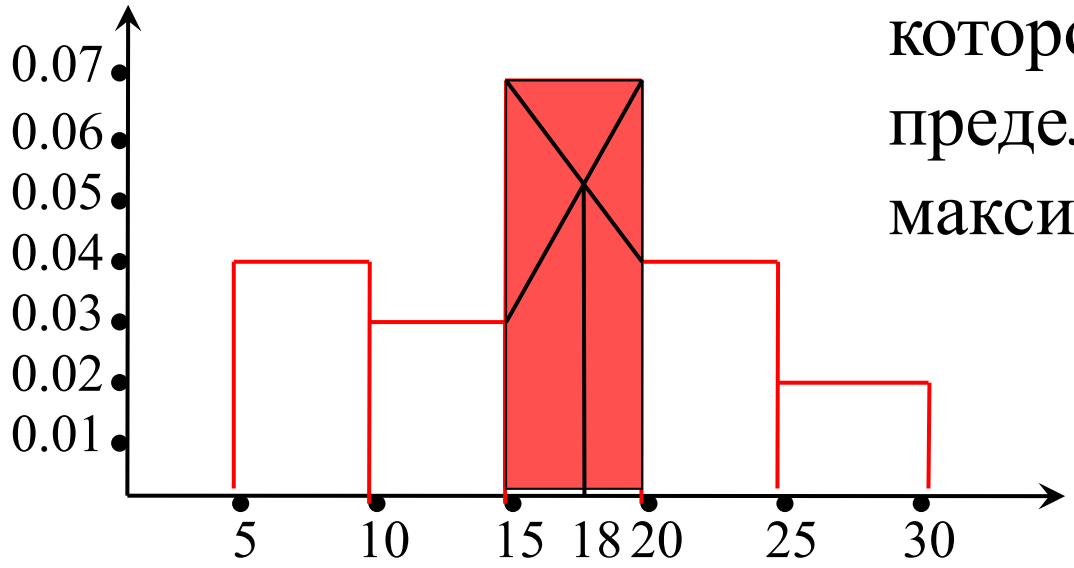
Гистограмма относительных частот даёт представление о плотности распределения генеральной совокупности.

Построим гистограмму относительных частот.

x_i	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	20	15	35	20	10
w_i	0.2	0.15	0.35	0.2	0.1
w_i / h	0.04	0.03	0.07	0.04	0.02

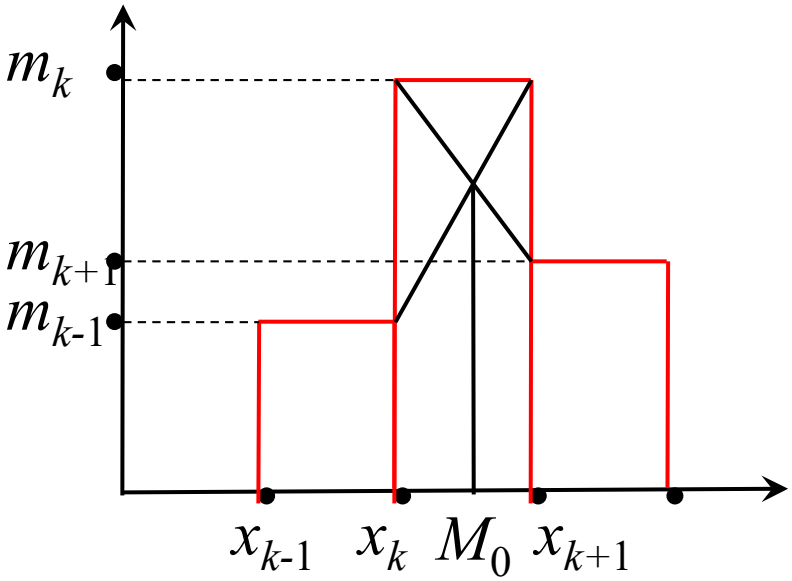
Объём выборки: $n = 100$

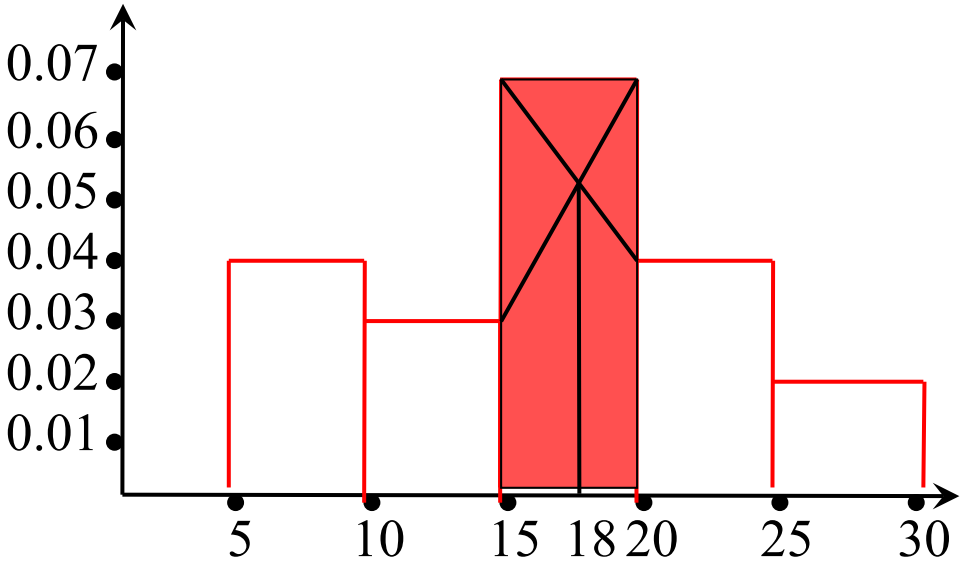
Длина интервала: $h = 5$



Мода – значение, при котором плотность распределения достигает максимума.

$M_0 = 18$





$$M_0 = x_k + \frac{m_k - m_{k-1}}{2m_k - (m_{k-1} + m_{k+1})}(x_{k+1} - x_k)$$

$x_{k-1} = 10 \qquad x_k = 15 \qquad x_{k+1} = 20$

$m_{k-1} = 0.03 \qquad m_k = 0.07 \qquad m_{k+1} = 0.04$

$$M_0 = 15 + \frac{0.07 - 0.03}{2 \cdot 0.07 - (0.03 + 0.04)}(20 - 15) \approx 17.857$$

Медиана

Медиана генеральной совокупности – такое число x, что

$$p(X < x) = p(X > x) = 0.5$$

Как оценить медиану генеральной совокупности?

– такое число M_e, что количество вариант, меньших M_e, равно количеству вариант, больших M_e

0, 0, 1, 2, 2, 2, 4, 5, 5, 5, 5, 6, 6 $\quad M_e = 4$

0, 0, 1, 2, 2, 2, 3, 4, 5, 5, 5, 5, 6, 6 $\quad M_e = ?$

$$M_e = (3+4)/2 = 3.5$$

Если n – нечётное, то $M_e = x_{(n+1)/2}$ (средняя варианта).

Если n – чётное, то $M_e = (x_{n/2} + x_{(n/2)+1})/2$

x_i	(x_1, x_2)	(x_2, x_3)	…
n_i	n_1	n_2	…

n – объём выборки

h – длина интервала

Находим такое число l, что $\sum_{i=1}^{l} n_i \le n/2, \quad \sum_{i=1}^{l+1} n_i > n/2.$

Пусть $f = \sum_{i=1}^{l} n_i.$

$$M_e = x_{l+1} + \frac{n/2 - f}{n_{l+1}} \cdot h$$

x_i	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	20	15	35	20	10

$n/2 = 50$

$h = 5$

$35 < 50$

$70 > 50$

$l = 2 \qquad f = 35$

$x_{l+1} = x_3 = 15 \qquad n_{l+1} = n_3 = 35$

$$M_e = 15 + \frac{50 - 35}{35} \cdot 5 \approx 17.143$$

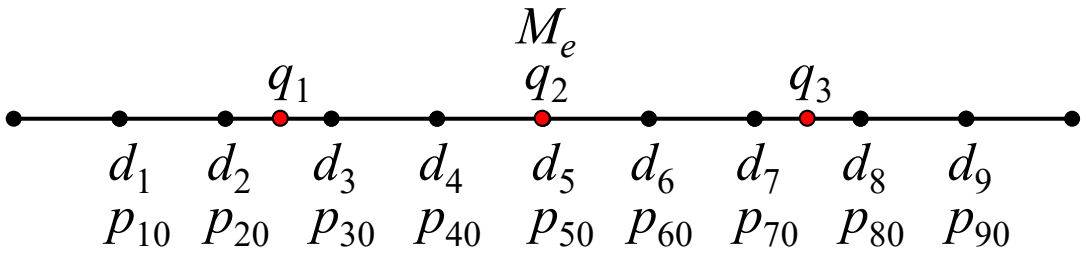
0, 0, 1, 2, 2, 2, 4, 4, 5, 5, 5, 5, 6 $\qquad M_e = 4$

Ряд наблюдений делится на 2 части, равные по количеству вариант.

Разделим ряд наблюдений на 4 равные части.
Получим три числа q_1, q_2, q_3, которые оценивают, соответственно, первый, второй и третий ***квартили***.

На 10 равных частей: $d_1, d_2, \ldots, d_9$ – ***децили***.

На 100 равных частей: $p_1, p_2, \ldots, p_{99}$ – ***процентили***.



Нахождение k-того квартиля q_k, дециля d_k и процентиля p_k

$x_1, x_2, ..., x_n$ – все элементы выборки

1. Находим число m:

для q_k $m = \frac{k}{4} \cdot n$, для d_k $m = \frac{k}{10} \cdot n$, для p_k $m = \frac{k}{100} \cdot n$.

2. Если m – целое число, то

$$q_k = d_k = p_k = \frac{x_m + x_{m+1}}{2}$$

3. Если m – не целое число, то

$$q_k = d_k = p_k = x_j,$$

где j – первое целое число после m.

Нахождение k-того квартиля q_k, дециля d_k и процентиля p_k

x_i	(x_1, x_2)	(x_2, x_3)	…
n_i	n_1	n_2	…

n – объём выборки

h – длина интервала

1. Находим число m:

для q_k $m = \frac{k}{4} \cdot n$, для d_k $m = \frac{k}{10} \cdot n$, для p_k $m = \frac{k}{100} \cdot n$.

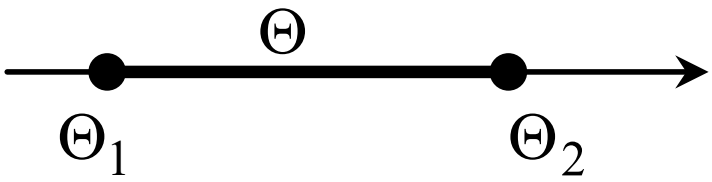
2. Находим такое число l, что $\sum_{i=1}^{l} n_i \le m$, $\sum_{i=1}^{l+1} n_i > m$.

3. Обозначим $f = \sum_{i=1}^{l} n_i$.

$$q_k = d_k = p_k = x_{l+1} + \frac{m - f}{n_{l+1}} \cdot h$$

Интервальные статистические оценки параметров распределения

$\Theta \approx \Theta^*$ – точечная оценка



Интервальной называют оценку, которая определяется двумя числами – концами интервала:

$$\Theta \in (\Theta_1, \Theta_2)$$

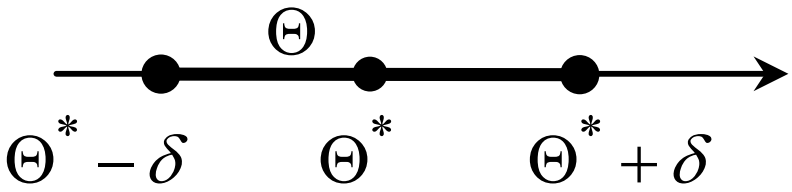
$$\Theta_1 = f_1(x_1, x_2, ..., x_n) \qquad \Theta_2 = f_2(x_1, x_2, ..., x_n)$$

– формулы для нахождения границ интервала по выборочным данным

Интервал (Θ_1, Θ_2), который содержит в себе неизвестный параметр Θ с заданной вероятностью γ, называют ***доверительным интервалом***:

$$p(\Theta_1 < \Theta < \Theta_2) = \gamma$$

При этом вероятность γ называют ***доверительной вероятностью*** или ***надёжностью*** оценки.



$$(\Theta^* - \delta, \Theta^* + \delta)$$

$$p(\Theta^* - \delta < \Theta < \Theta^* + \delta) = p(-\delta < \Theta - \Theta^* < \delta) =$$

$$= p(|\Theta - \Theta^*| < \delta) = \gamma$$

Число δ называют ***точностью*** оценки.

1. Пусть X – непрерывная случайная величина,
$F(x)$ – функция распределения,
$f(x)$ – плотность распределения

$$p(a < X < b) = F(b) - F(a) = \int_a^b f(x)dx \qquad (*)$$

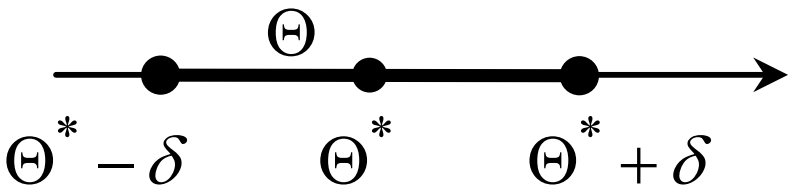
2. Пусть плотность распределения $f(x)$ – чётная функция

$$p(|x| < t) = 2\int_0^t f(x)dx \qquad (**)$$

$$p(|x| > t) = 2(1 - F(t)) \qquad (***)$$

$$p(|x| < t) = 2F(t) - 1 \qquad (****)$$

Алгоритм нахождения доверительных интервалов



$$p(\Theta^* - \delta < \Theta < \Theta^* + \delta) = \gamma$$

$$\Rightarrow \quad p(-\delta < \Theta - \Theta^* < \delta) = \gamma$$

Θ – число, Θ^* – случайная величина $\Rightarrow$

$\Theta - \Theta^*$ – случайная величина $\Rightarrow$ из (*)

$$p(-\delta < \Theta - \Theta^* < \delta) = F(\delta) - F(-\delta) = \int_{-\delta}^{\delta} f(x)dx = \gamma$$

Уравнения для нахождения δ:

$$F(\delta) - F(-\delta) = \gamma \quad \text{или} \quad \int_{-\delta}^{\delta} f(x)dx = \gamma$$

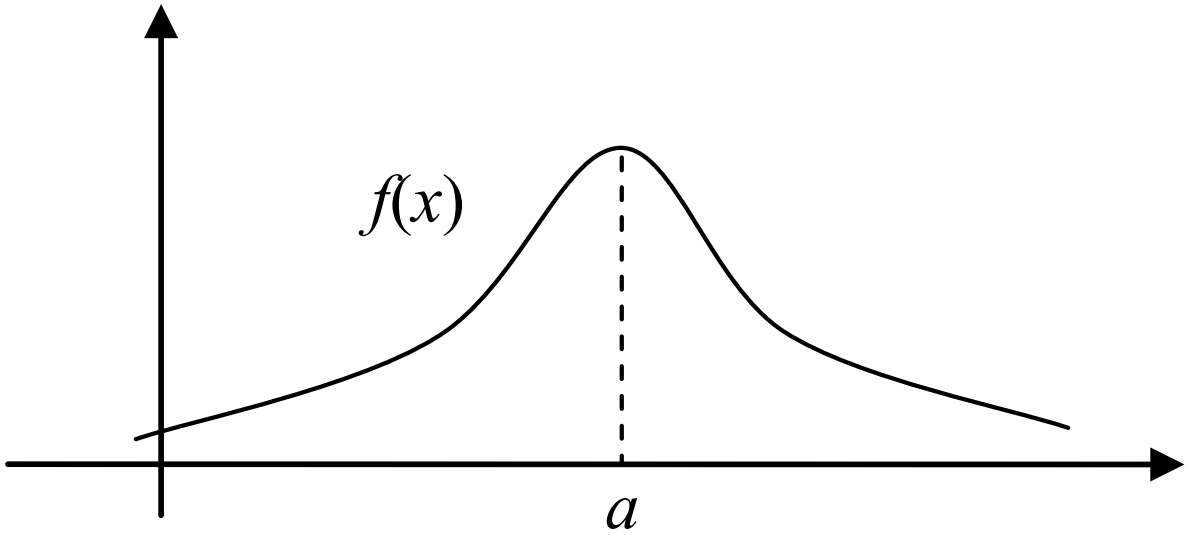
Вопрос: какой вид имеют функции $F(x)$ и $f(x)$, то есть какое распределение имеет $\Theta - \Theta^*$?

Нормальное распределение

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

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{\frac{-(x-a)^2}{2\sigma^2}}dx$$

a, σ – параметры распределения



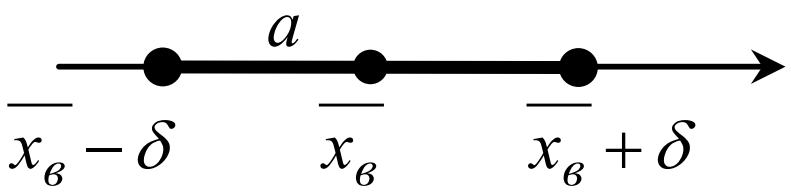
$M(X)=a$ и $D(X)=\sigma^2$

1. Пусть генеральная совокупность имеет нормальное распределение

если σ – известно, $\Theta = a = ?$

1. Задаём надёжность γ.
2. Находим точечную оценку: $a \approx \overline{x_в}$.
3. Находим доверительный интервал $(\overline{x_в} - \delta, \overline{x_в} + \delta)$, то есть такое δ, что

$$p(\overline{x_в} - \delta < a < \overline{x_в} + \delta) = \gamma$$



$\frac{\overline{x_в} - a}{\sigma} \cdot \sqrt{n}$ – случайная величина, имеющая нормальное распределение с нулевым математическим ожиданием и единичной дисперсией

Шаг1. Найдём такое число t_γ, что

$$p\left(\left|\frac{\overline{x_в} - a}{\sigma} \cdot \sqrt{n}\right| < t_\gamma\right) = \gamma$$

$$t_\gamma = F^{-1}\left(\frac{1+\gamma}{2}\right) \quad \text{или} \quad t_\gamma = \Phi^{-1}\left(\frac{\gamma}{2}\right)$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(x-a)^2}{2\sigma^2}} dx$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{t^2}{2}} dt$$

– ***функция Лапласа***.

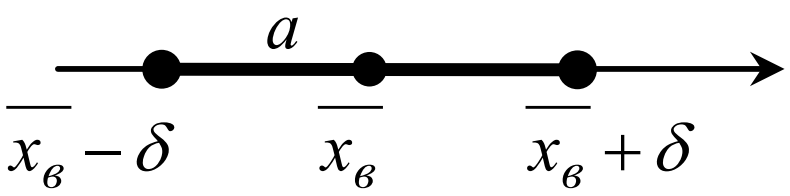
Шаг 2.

$$\left|\frac{\overline{x_в} - a}{\sigma} \cdot \sqrt{n}\right| < t_\gamma \quad \Leftrightarrow \quad \overline{x_в} - \frac{t_\gamma \cdot \sigma}{\sqrt{n}} < a < \overline{x_в} + \frac{t_\gamma \cdot \sigma}{\sqrt{n}}$$

$$p\left(\left|\frac{\overline{x_в}-a}{\sigma}\cdot\sqrt{n}\right|<t_\gamma\right)=p\left(\overline{x_в}-\frac{t_\gamma\cdot\sigma}{\sqrt{n}}<a<\overline{x_в}+\frac{t_\gamma\cdot\sigma}{\sqrt{n}}\right)=\gamma$$

Надо найти такой интервал $(\overline{x_в}-\delta,\ \overline{x_в}+\delta)$, что

$$p(\overline{x_в}-\delta<a<\overline{x_в}+\delta)=\gamma$$



Таким образом, $\delta=\dfrac{t_\gamma\cdot\sigma}{\sqrt{n}}$.

Доверительным интервалом является интервал:

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

Доверительным интервалом является интервал:

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

$$t_\gamma = F^{-1}\left(\frac{1+\gamma}{2}\right) \quad \text{или} \quad t_\gamma = \Phi^{-1}\left(\frac{\gamma}{2}\right)$$

F^{-1} – функция, обратная к функции нормального распределения с нулевым математическим ожиданием и единичной дисперсией

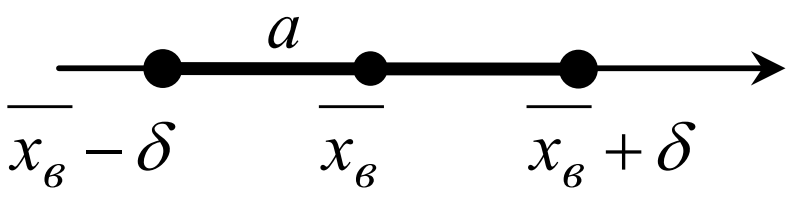
Φ^{-1} – функция, обратная к функции Лапласа

2. Пусть генеральная совокупность имеет нормальное распределение

σ – неизвестно, $\Theta = a = ?$

1. Задаём надёжность γ.
2. Находим точечную оценку: $a \approx \overline{x_в}$.
3. Находим доверительный интервал $(\overline{x_в} - \delta, \overline{x_в} + \delta)$, то есть такое δ, что

$$p(\overline{x_в} - \delta < a < \overline{x_в} + \delta) = \gamma$$



$\frac{\overline{x_в} - a}{s} \cdot \sqrt{n}$ – случайная величина, имеющая распределение Стьюдента с (n-1) степенями свободы

Шаг1. Найдём такое число t_γ, что

$$p\left(\left|\frac{\overline{x_в} - a}{s} \cdot \sqrt{n}\right| < t_\gamma\right) = \gamma$$

$$t_\gamma = F^{-1}\left(\frac{1+\gamma}{2}\right)$$

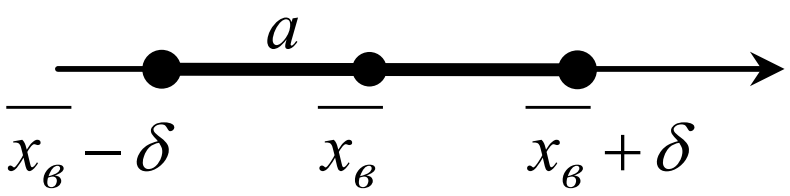
Шаг 2.

$$\left|\frac{\overline{x_в} - a}{s} \cdot \sqrt{n}\right| < t_\gamma \quad \Leftrightarrow \quad \overline{x_в} - \frac{t_\gamma \cdot s}{\sqrt{n}} < a < \overline{x_в} + \frac{t_\gamma \cdot s}{\sqrt{n}}$$

$$p\left(\left|\frac{\overline{x_в}-a}{s}\cdot\sqrt{n}\right|<t_\gamma\right)=p\left(\overline{x_в}-\frac{t_\gamma\cdot s}{\sqrt{n}}<a<\overline{x_в}+\frac{t_\gamma\cdot s}{\sqrt{n}}\right)=\gamma$$

Надо найти такой интервал $(\overline{x_в}-\delta,\ \overline{x_в}+\delta)$, что

$$p(\overline{x_в}-\delta<a<\overline{x_в}+\delta)=\gamma$$



Таким образом, $\delta=\dfrac{t_\gamma\cdot s}{\sqrt{n}}$.

Доверительным интервалом является интервал:

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

Доверительным интервалом является интервал:

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

$$t_\gamma = F^{-1}\left(\frac{1+\gamma}{2} \right)$$

F^{-1} – функция, обратная к функции распределения Стьюдента с $(n - 1)$ степенями свободы

s – исправленное выборочное среднее квадратическое отклонение

3. Пусть генеральная совокупность имеет нормальное распределение

$\sigma = ?$

1. Задаём надёжность γ.

2. Находим точечную оценку: $\sigma \approx s$.

3. Находим доверительный интервал $(s-\delta, s+\delta)$, то есть такое δ, что

$$p(s-\delta < \sigma < s+\delta) = \gamma$$

$\frac{(n-1)\cdot s^2}{\sigma^2}$ – случайная величина, имеющая χ^2- распределение с (n-1) степенями свободы

Шаг1. Найдём такое число q_γ, что

$$p\left(\frac{n-1}{(1+q_\gamma)^2} < \frac{(n-1)\cdot s^2}{\sigma^2} < \frac{n-1}{(1-q_\gamma)^2}\right) = \gamma$$

$$F\left(\frac{n-1}{(1-q_\gamma)^2}\right) - F\left(\frac{n-1}{(1+q_\gamma)^2}\right) = \gamma \quad \Rightarrow \quad q_\gamma$$

Шаг 2.

$$\frac{n-1}{(1+q_\gamma)^2} < \frac{(n-1)\cdot s^2}{\sigma^2} < \frac{n-1}{(1-q_\gamma)^2} \quad \Leftrightarrow \quad s(1-q_\gamma) < \sigma < s(1+q_\gamma)$$

$$p\left(\frac{n-1}{(1+q_\gamma)^2} < \frac{(n-1)\cdot s^2}{\sigma^2} < \frac{n-1}{(1-q_\gamma)^2}\right) = p(s - sq_\gamma < \sigma < s + sq_\gamma) = \gamma$$

Надо найти такой интервал $(s-\delta, s+\delta)$, что

$$p(s-\delta < \sigma < s+\delta) = \gamma$$

Таким образом, $\delta = sq_\gamma$.

Замечание: при $q_\gamma > 1$ имеем $s(1-q_\gamma) < 0$, но $\sigma > 0$

$$\Rightarrow \text{ при } q_\gamma > 1 \quad 0 < \sigma < s(1+q_\gamma)$$

Доверительным интервалом является интервал:

$\left(s(1-q_\gamma), s(1+q_\gamma)\right)$ при $q_\gamma \le 1$

$\left(0, s(1+q_\gamma)\right)$ при $q_\gamma > 1$

Способ 2.

Шаг1. Найдём такие числа χ_1^2 и χ_2^2, что

$$p\left(\chi_1^2 < \frac{(n-1)\cdot s^2}{\sigma^2} < \chi_2^2\right) = \gamma$$

$$\chi_1^2 = F^{-1}\left(\frac{1-\gamma}{2}\right) \qquad \chi_2^2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$$

Шаг 2.

$$\chi_1^2 < \frac{(n-1)\cdot s^2}{\sigma^2} < \chi_2^2 \quad \Leftrightarrow \quad \frac{s}{\chi_2}\cdot\sqrt{n-1} < \sigma < \frac{s}{\chi_1}\cdot\sqrt{n-1}$$

$$p\left(\chi_1^2 < \frac{(n-1)\cdot s^2}{\sigma^2} < \chi_2^2\right) = p\left(\frac{s}{\chi_2}\cdot\sqrt{n-1} < \sigma < \frac{s}{\chi_1}\cdot\sqrt{n-1}\right) = \gamma$$

Доверительным интервалом является интервал:

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

4. Пусть производятся независимые испытания с неизвестной вероятностью p появления события A в каждом испытании.

p – ?

1. Задаём надёжность γ.

2. Находим точечную оценку: $p \approx w = \frac{m}{n}$

   m – число появлений события A при n испытаниях.

$$M(w) = p \qquad \sigma(w) = \sqrt{p(1-p)/n}$$

3. Находим доверительный интервал (p_1, p_2), то есть такие числа p_1 и p_2, что

$$p(p_1 < p < p_2) = \gamma$$

w – случайная величина, имеющая нормальное распределение, причём $a = p$ и $\sigma = \sqrt{p(1-p)/n}$

$\dfrac{w-p}{\sqrt{p(1-p)/n}}$ – случайная величина, имеющая нормальное распределение с нулевым математическим ожиданием и единичной дисперсией

Шаг 1. Найдём такое t, что

$$p\left(\left|\frac{w-p}{\sqrt{p(1-p)/n}}\right| < t\right) = \gamma$$

$$t = F^{-1}\left(\frac{1+\gamma}{2}\right) \quad \text{или} \quad t = \Phi^{-1}\left(\frac{\gamma}{2}\right)$$

Шаг 2.

$$\left|\frac{w-p}{\sqrt{p(1-p)/n}}\right| < t \quad \Leftrightarrow \quad (1+t^2/n)\cdot p^2 - (2w+t^2/n)\cdot p + w^2 < 0$$

$$\Leftrightarrow p_1 < p < p_2, \text{ где } \quad p_1 = \frac{n}{t^2+n}\left(w + \frac{t^2}{2n} - t\sqrt{\frac{w(1-w)}{n} + \left(\frac{t}{2n}\right)^2}\right)$$

$$p_2 = \frac{n}{t^2+n}\left(w + \frac{t^2}{2n} + t\sqrt{\frac{w(1-w)}{n} + \left(\frac{t}{2n}\right)^2}\right)$$

$$p\left(\left|\frac{w-p}{\sqrt{p(1-p)/n}}\right| < t\right) = p(p_1 < p < p_2) = \gamma$$

Доверительным интервалом является интервал:

$$(p_1, p_2)$$

$$p_1 = \frac{n}{t^2+n}\left(w + \frac{t^2}{2n} - t\sqrt{\frac{w(1-w)}{n} + \left(\frac{t}{2n}\right)^2} \right)$$

$$p_2 = \frac{n}{t^2+n}\left(w + \frac{t^2}{2n} + t\sqrt{\frac{w(1-w)}{n} + \left(\frac{t}{2n}\right)^2} \right)$$

При больших значениях n (порядка сотен)

$$\frac{t^2}{2n} \to 0 \qquad \left(\frac{t}{2n}\right)^2 \to 0 \qquad \frac{n}{t^2+n} \to 1 \qquad \Rightarrow$$

$$p_1 = w - t\sqrt{\frac{w(1-w)}{n}} \qquad \text{и} \qquad p_2 = w + t\sqrt{\frac{w(1-w)}{n}}$$

Доверительным интервалом является интервал:

$$(w-\delta, w+\delta), \text{ где } \delta = t\sqrt{\frac{w(1-w)}{n}}$$

Доверительным интервалом является интервал:

$$(w-\delta, w+\delta), \text{ где } \delta = t\sqrt{\frac{w(1-w)}{n}}$$

$$t = F^{-1}\left(\frac{1+\gamma}{2}\right) \quad \text{или} \quad t = \Phi^{-1}\left(\frac{\gamma}{2}\right)$$

F^{-1} – функция, обратная к функции нормального распределения с нулевым математическим ожиданием и единичной дисперсией

Φ^{-1} – функция, обратная к функции Лапласа

Проверка статистических гипотез

Пусть для получения опытных данных необходимо провести обследование соответствующих объектов (***генеральную совокупность***).

Обычно исследуют не всю совокупность объектов, а отбирают из неё некоторое количество объектов и исследуют только их (***выборочную совокупность*** или другими словами ***выборку***).

По выборке судят о генеральной совокупности,
следовательно, любое высказывание о генеральной совокупности является гипотезой.

Статистическая гипотеза

– это любое предположение о виде или параметрах неизвестного закона распределения.

Примеры.

1. Генеральная совокупность распределена по закону Пуассона.
2. Математическое ожидание генеральной совокупности равно 100.
3. Дисперсии двух генеральных совокупностей равны.
4. ~~На Марсе есть жизнь.~~

Проверяемую гипотезу называют ***нулевой (основной)***, обозначают её H_0.

Выдвинутая гипотеза может быть принята или отвергнута.

Наряду с выдвинутой гипотезой H_0 рассматривают и противоречащую ей гипотезу, которую называют ***конкурирующей (альтернативной)*** и обозначают H_1.

В зависимости от выборочных данных принимается либо основная гипотеза, либо конкурирующая.

Задача: проверить, верна ли нулевая гипотеза H_0 при альтернативной гипотезе H_1?

Пример.

Пусть известно, что генеральная совокупность распределена по показательному закону.

$$f(x)=\begin{cases}0, & x<0\\ \lambda e^{-\lambda x}, & x\geq 0\end{cases}$$

λ – параметр распределения

λ – неизвестен

H_0: $\lambda = 10$

H_1: $\lambda = 20$

H_1: $\lambda = 5$

H_1: $\lambda > 10$

H_1: $\lambda \neq 10$

Гипотеза H_0	Принимается	Отвергается
Верна	Правильное решение	Ошибка 1-го рода
Неверна	Ошибка 2-го рода	Правильное решение

Обозначим через α – вероятность допустить ошибку 1-го рода, через β – 2-го рода.

Вероятность α допустить ошибку 1-го рода, то есть отвергнуть верную гипотезу H_0, называют ***уровнем значимости***.

Общая схема проверки статистических гипотез

1 этап

Задаём уровень значимости α .

α – вероятность ошибки 1-го рода (ошибочно отвергнуть верную гипотезу)

В качестве α обычно берётся малое значение:

0.05, 0.01, 0.005, 0.001.

2 этап

Строим случайную величину K, называемую ***статистическим критерием***, для которой выполняются следующие условия:

1) она является функцией от выборочных данных: $K=K(x_1,x_2,\ldots,x_n)$;

2) её значения позволяют судить о «расхождении выборки с гипотезой H_0», то есть о том, надо принимать или отвергать гипотезу H_0;

3) распределение этой величины известно.

3 этап

Вычисляем значения критерия, подставляя в него выборочные данные. Это число называют ***наблюдаемым значением критерия*** и обозначают $K_{набл}$.

4 этап

Находим ***критическую область*** данного критерия, то есть совокупность значений критерия, при которых нулевую гипотезу отвергают.

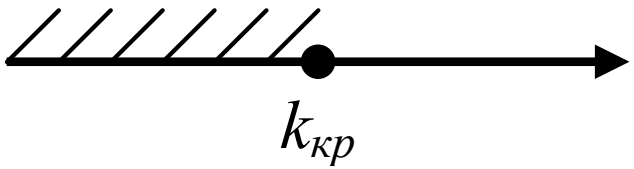
Все остальные значения критерия образуют область, называемую ***областью принятия гипотезы***.

Если наблюдаемое значение критерия попадает в критическую область, то гипотезу отвергаем, если в область принятия гипотезы, то принимаем.

Точки, которые отделяют критическую область от области принятия гипотезы, называют ***критическими точками***.

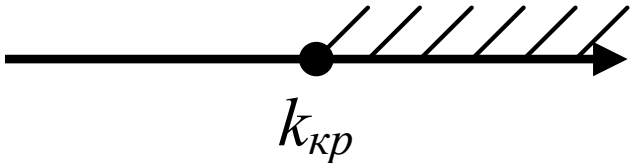
Чаще всего встречаются следующие виды критических областей:

а) левосторонняя



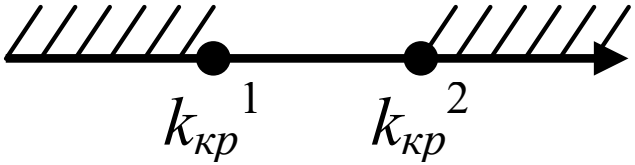
$K < k_{кр}$

б) правосторонняя



$K > k_{кр}$

в) двусторонняя



$K < k_{кр}^{1}$

$K > k_{кр}^{2}$

Критическую область W целесообразно находить согласно следующим требованиям:

1. $p(K \in W) = \alpha$
2. вероятность β ошибки 2-го рода – минимальная, то есть вероятность $(1-\beta)$ – максимальная

Вероятность $(1-\beta)$ не допустить ошибку 2-го рода, то есть отвергнуть гипотезу H_0, когда она неверна, называется ***мощностью*** критерия.

1. $p(K \in W) = \alpha$
2. мощность критерия – максимальная

Схема проверки статистических гипотез

1. Задаём уровень значимости.
    - зависит от «тяжести последствий» ошибок 1-го и 2-го рода для каждой конкретной задачи
2. Строим статистический критерий.
    - для каждой гипотезы имеет свой вид
    - описаны в литературе

3. Вычисляем наблюдаемое значение критерия.

- подставляем в формулу выборочные данные

4. Находим критическую область и проверяем, попадает ли в неё наблюдаемое значение критерия.

- критическая область зависит от вида конкурирующей гипотезы
- критические точки находятся по специальным таблицам или с помощью компьютера

Критерий Стьюдента

Известно, что генеральная совокупность распределена по нормальному закону, но его параметры неизвестны.

a, σ – параметры распределения $\quad M(X) = a$

Проверить гипотезу: $H_0 : a = a_0$

a_0 – некоторое число

Критерий:

$$T = \frac{\overline{x_в} - a_0}{s} \cdot \sqrt{n}$$

$\overline{x_в}$ – выборочная средняя

n – объём выборки

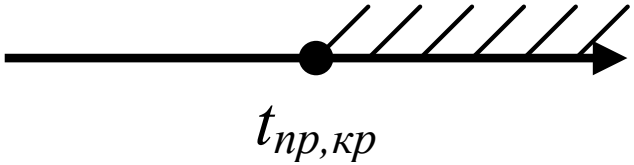
s – исправленное среднее квадратическое отклонение

T имеет распределение Стьюдента с $(n\text{-}1)$ степенями свободы

Критическая область строится в зависимости от вида конкурирующей гипотезы.

1. $H_1 : a > a_0$

Критическая область W – правосторонняя:



Из требования 1 для критической области:

$$p(T \in W) = \alpha \quad \Rightarrow \quad p(T > t_{пр,кр}) = \alpha$$

$$p(T < t_{пр,кр}) = 1 - p(T > t_{пр,кр}) = 1 - \alpha$$

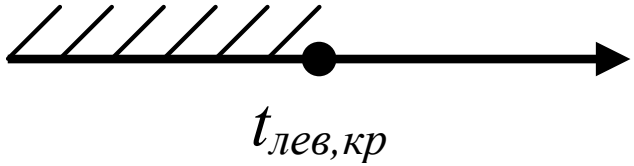
$p(T < t_{пр,кр}) = F(t_{пр,кр})$, $F(x)$ – функция распределения T

$$F(t_{пр,кр}) = 1 - \alpha \quad \Rightarrow \quad t_{пр,кр} = F^{-1}(1-\alpha)$$

$F(x)$ – функция распределения Стьюдента с $(n\text{-}1)$ степенями свободы

2. $H_1 : a < a_0$

Критическая область W – левосторонняя:



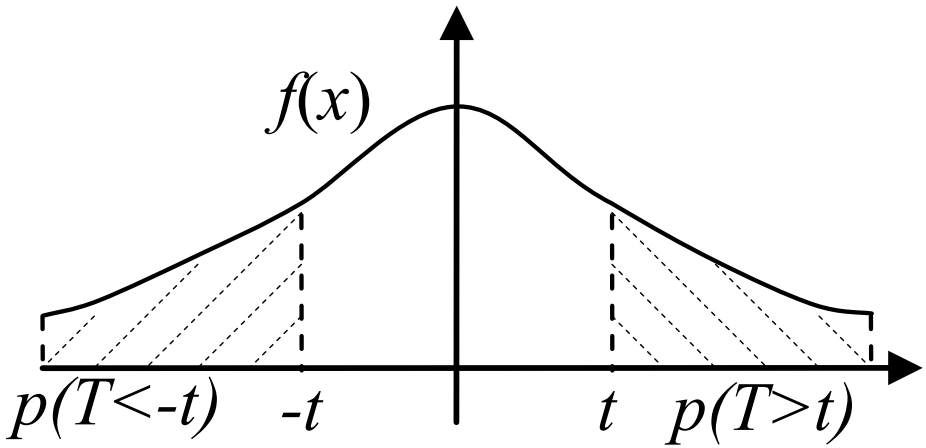
Из требования 1 для критической области:

$$p(T \in W) = \alpha \quad \Rightarrow \quad p(T < t_{лев,кр}) = \alpha$$

$p(T < t_{лев,кр}) = F(t_{лев,кр}) = \alpha$, $F(x)$ – функция распределения T

$\Rightarrow$ $t_{лев,кр} = F^{-1}(\alpha)$, $F(x)$ – функция распределения Стьюдента с (n-1) степенями свободы

Плотность распределения Стьюдента – чётная функция



$$\Rightarrow \quad p(T > t) = p(T < -t)$$

Критическая точка $t_{пр,кр}$ находится из требования:

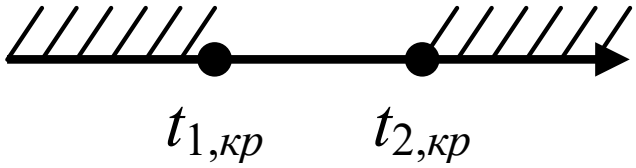
$$p(T > t_{пр,кр}) = \alpha \quad \Rightarrow \quad p(T < -t_{пр,кр}) = \alpha \quad \Rightarrow$$

$-t_{пр,кр}$ является критической точкой для левосторонней области:

$$t_{лев,кр} = -t_{пр,кр}$$

3. $H_1 : a \neq a_0$

Критическая область W – двусторонняя:



Пусть $p(T < t_{1,кр}) = p(T > t_{2,кр}) = \alpha / 2$

В силу чётности плотности распределения Стьюдента:

$$t_{1,кр} = -t_{2,кр}$$

Аналогично пунктам **1** и **2** получаем:

$$t_{2,кр} = F^{-1}(1-\alpha/2), \quad t_{1,кр} = -t_{2,кр}$$

или

$$t_{1,кр} = F^{-1}(\alpha/2), \quad t_{2,кр} = -t_{1,кр}$$

Пример.

Проектный контролируемый размер изделий, изготовляемых станком-автоматом, $a = 35$ *мм*. Измерения 20 случайно отобранных изделий дали следующие результаты:

x_i	34.8	34.9	35.0	35.1	35.3
n_i	2	3	4	6	5

Требуется при уровне значимости 0.05 проверить нулевую гипотезу H_0: $a = 35$ при конкурирующей гипотезе H_1: $a \neq 35$.

x_i	34.8	34.9	35.0	35.1	35.3
n_i	2	3	4	6	5

$n = 20$

$\alpha = 0.05$

$H_0: a = 35$ $H_1: a \neq 35$ $a_0 = 35$

$$T = \frac{\overline{x_в} - a_0}{s} \cdot \sqrt{n}$$

$\overline{x_в}$ – выборочная средняя

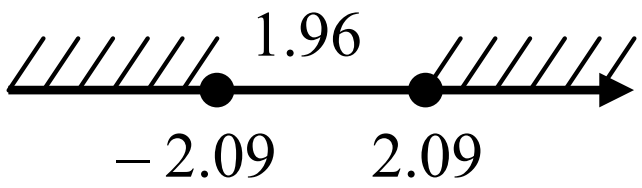
s – исправленное среднее квадратическое отклонение

$\overline{x_в} = 35.07$

$s = 0.16$

$$T_{набл} = \frac{35.07 - 35}{0.16} \cdot \sqrt{20} \approx 1.96$$

Критическая область двусторонняя:



Принимаем нулевую гипотезу, то есть станок обеспечивает проектный размер изделий.

Критерии, с помощью которых проверяется гипотеза о теоретическом законе распределения, называются ***критериями согласия***.

H_0: генеральная совокупность имеет некоторое определённое распределение

(высказано предположение о ***виде*** и ***параметрах*** распределения)

1. Генеральная совокупность имеет биномиальное распределение с m=10 и p=0.4.

2. Генеральная совокупность распределена нормально с математическим ожиданием, равным 5 и дисперсией, равной 4.

Критерий Пирсона (χ^2-критерий)

Найдём ***теоретические частоты*** вариант.

1. Распределение дискретное $\Rightarrow$ $p(x)$.

x_i	x_1	x_2	…	x_{l-1}	x_l
p_i	$p_1=p(x_1)$	$p_2=p(x_2)$	…	$p_{l-1}=p(x_{l-1})$	$p_l=1-p_1-p_2-\ldots-p_{l-1}$

Теоретическая частота появления варианты x_i – это np_i.

2. Распределение непрерывное $\Rightarrow$ $F(x)$.

x_i	(x_1, x_2)	(x_2, x_3)	…	(x_{l-1}, x_l)	(x_l, x_{l+1})
p_i	$p_1=p(X<x_2)$ $=F(x_2)$	$p_2=p(x_2<X<x_3)$ $=F(x_3)- F(x_2)$	…	$p_{l-1}=p(x_{l-1}<X<x_l)$ $=F(x_l)- F(x_{l-1})$	$p_l=1-p_1-p_2-$ $\ldots-p_{l-1}$

Теоретическая частота попадания в интервал (x_i, x_{i+1}) – это np_i.

Критерий: $$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}$$

n_i – эмпирические частоты

np_i – теоретические частоты

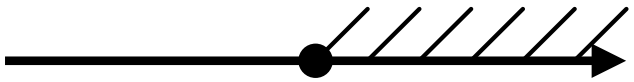
При $n \to \infty$ случайная величина χ^2 имеет распределение Пирсона с k степенями свободы, где

$$k = l - 1 - r,$$

l – число вариант (интервалов),

r – число параметров предполагаемого распределения, оцениваемых по выборке

Критическая область W – правосторонняя:



$\chi^2_{кр}$

Из требования 1 для критической области:

$$p(\chi^2 \in W) = \alpha \quad \Rightarrow \quad p(\chi^2 > \chi^2_{кр}) = \alpha$$

$$p(\chi^2 < \chi^2_{кр}) = 1 - p(\chi^2 > \chi^2_{кр}) = 1 - \alpha$$

$$p(\chi^2 < \chi^2_{кр}) = F(\chi^2_{кр}),\ F(x) - \text{функция распределения } \chi^2$$

$$F(\chi^2_{кр}) = 1 - \alpha \quad \Rightarrow \quad \chi^2_{кр} = F^{-1}(1-\alpha)$$

$F(x)$ – функция распределения Пирсона с $k=l-1-r$ степенями свободы, l – число вариант (интервалов), r – число параметров, оцениваемых по выборке.

Критерий Колмогорова

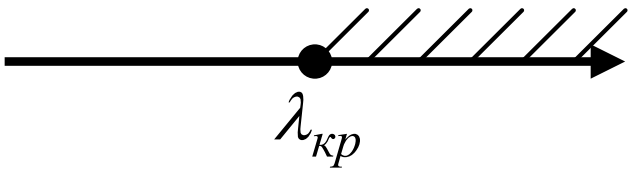
$F(x)$ – теоретическая функция распределения
$F_n(x)$ – эмпирическая функция распределения

Обозначим $D = \max\left|F_n(x) - F(x)\right|$

– ***статистика*** критерия Колмогорова

Критерий: $\lambda = D\sqrt{n}$

Критическая область W – правосторонняя:



Из требования 1 для критической области:

$$p(\lambda \in W) = p(\lambda > \lambda_{кр}) = p(D\sqrt{n} > \lambda_{кр}) = \alpha$$

Можно доказать, что при $n \to \infty$

$$p(D\sqrt{n} > \lambda_{кр}) \to p(\lambda_{кр}) = 1 - \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 \lambda_{кр}^2}$$

$$\Rightarrow \quad 1 - \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 \lambda_{кр}^2} = \alpha \quad \Rightarrow \quad \lambda_{кр}$$

α	0.4	0.3	0.2	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
$\lambda_{кр}$	0.89	0.97	1.07	1.22	1.36	1.48	1.63	1.73	1.95	2.03

Критерий Фишера

Две генеральные совокупности X и Y распределены нормально.

Проверить гипотезу: $H_0 : D(X) = D(Y)$

Обозначим n_X – объём выборки из совокупности X,
n_Y – объём выборки из совокупности Y,
$s^2{}_X$ и $s^2{}_Y$ – исправленные выборочные дисперсии.

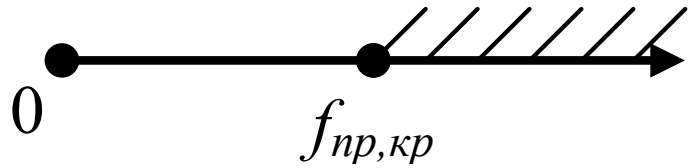
Критерий: $F = \dfrac{s_X^2}{s_Y^2}$

F имеет распределение Фишера с $(n_X - 1)$ и $(n_Y - 1)$ степенями свободы

Критическая область строится в зависимости от вида конкурирующей гипотезы.

1. $H_1 : D(X) > D(Y)$

Критическая область W – правосторонняя:



Так как $s^2_X > 0$ и $s^2_Y > 0$, то $F > 0 \Rightarrow$ положительная часть

Из требования 1 для критической области:

$$p(F \in W) = \alpha \quad \Rightarrow \quad p(F > f_{пр,кр}) = \alpha$$

$$p(F < f_{пр,кр}) = 1 - p(F > f_{пр,кр}) = 1 - \alpha$$

$p(F < f_{пр,кр}) = F(f_{пр,кр})$, $F(x)$ – функция распределения F

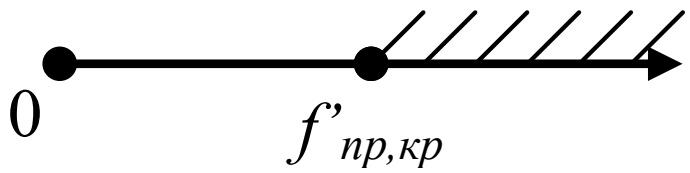
$$F(f_{пр,кр}) = 1 - \alpha \quad \Rightarrow \quad f_{пр,кр} = F^{-1}(1-\alpha)$$

$F(x)$ – функция распределения Фишера с $(n_X - 1)$ и $(n_Y - 1)$ степенями свободы

2. $H_1 : D(X) < D(Y)$

Обозначим $F' = \frac{1}{F} = \frac{s_Y^2}{s_X^2}$, F' имеет распределение Фишера с $(n_Y - 1)$ и $(n_X - 1)$ степенями свободы

$H_1 : D(Y) > D(X) \Rightarrow$ предыдущий случай:



$f'_{пр,кр} = F^{-1}(1-\alpha)$, где $F(x)$ – функция распределения F'

$$p(F' > f'_{пр,кр}) = \alpha$$

$$p(F' > f'_{пр,кр}) = p(1/F > f'_{пр,кр}) = p(F < 1/f'_{пр,кр})$$

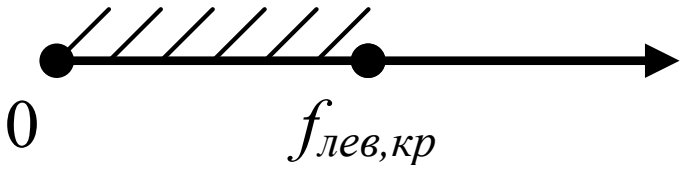
$$\Rightarrow \quad p(F < 1/f'_{пр,кр}) = \alpha$$

$$p(F < 1/f'_{пр,кр}) = \alpha$$

Обозначим $f'_{лев,кр} = 1/f'_{пр,кр}$, тогда

$$p(F < f_{лев,кр}) = \alpha$$

Таким образом, критическая область для критерия F имеет вид:

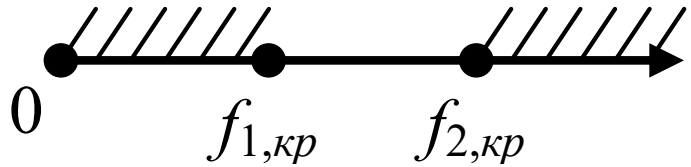


$$f_{лев,кр} = \frac{1}{f'_{пр,кр}} = \frac{1}{F^{-1}(1-\alpha)},$$

где $F(x)$ – функция распределения Фишера с $(n_Y - 1)$ и $(n_X - 1)$ степенями свободы

3. $H_1 : D(X) \neq D(Y)$

Критическая область W – двусторонняя:



Пусть $p(F < f_{1,кр}) = p(F > f_{2,кр}) = \alpha / 2$

Аналогично пунктам **1** и **2** получаем:

$$f_{2,кр} = F_1^{-1}(1-\alpha/2)$$

где $F_1(x)$ – функция распределения Фишера с $(n_X - 1)$ и $(n_Y - 1)$ степенями свободы

$$f_{1,кр} = \frac{1}{f'_{2,кр}} = \frac{1}{F_2^{-1}(1-\alpha/2)}$$

где $F_2(x)$ – функция распределения Фишера с $(n_Y - 1)$ и $(n_X - 1)$ степенями свободы