МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

- Визуализация данных
- Точечные оценки
- Групповые характеристики
- Метод наибольшего правдоподобия
- Метод моментов
- Интервальные оценки
- Алгоритм нахождения доверительных интервалов
- Оценка *a* при известной дисперсии
- Оценка *a* при неизвестной дисперсии
- Оценка среднего квадратического отклонения
- Оценка вероятности события
- Проверка статистических гипотез

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.

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.

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.

1. $\sum_i n_i = n$

2. $\sum_i w_i = 1$

3. $p(X = x_i) \approx w_i$

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).

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.

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

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

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

Выборочная характеристика

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

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	…

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

1. Если $u_i = x_i - c$ для всех i, где c – некоторое число, то

$$\overline{u_в} = \overline{x_в} - c$$

2. Если $u_i = hx_i$ для всех i, где h – некоторое число, то

$$\overline{u_в} = h\overline{x_в}$$

Выборочная дисперсия:

x_i	x_1	x_2	…
n_i	n_1	n_2	…

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

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

1. Если $u_i = x_i - c$ для всех i, где c – некоторое число, то

$$D_в(u) = D_в(x)$$

2. Если $u_i = hx_i$ для всех i, где h – некоторое число, то

$$D_в(u) = h^2 D_в(x)$$

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:

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

1-ая группа: N_1 элементов $\Rightarrow$ $\overline{x_1}$, D_1

2-ая группа: N_2 элементов $\Rightarrow$ $\overline{x_2}$, D_2

…

j-тая группа: N_j элементов $\Rightarrow$ $\overline{x_j}$, D_j

…

$\overline{x_1}, \overline{x_2}, \square$ – **групповые средние**

$D_1, D_2, \ldots$ – **групповые дисперсии**

$$\overline{x_в} = \frac{\sum N_j \overline{x_j}}{n}$$

$$D_{в} = D_{внгр} + D_{межгр}$$

$$D_{внгр} = \frac{\sum N_j D_j}{n}$$ – внутригрупповая дисперсия

$$D_{межгр} = \frac{\sum N_j (\overline{x_j} - \overline{x_в})^2}{n}$$ – межгрупповая дисперсия

Метод максимального (наибольшего) правдоподобия

1. Генеральная совокупность имеет распределение Пуассона

$$p(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$ – вероятность события x

$\Theta = \lambda = ?$

2. Генеральная совокупность имеет нормальное распределение

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

если σ – известно, $\Theta = a = ?$

если a – известно, $\Theta = \sigma = ?$

$x_1, x_2, \ldots, x_n$ – выборка

I. Дискретное распределение

$$L(x_1, x_2, \ldots, x_n, \Theta) = p(x_1) \cdot p(x_2) \cdot \ldots \cdot p(x_n)$$

$p(x_1), \ldots, p(x_n)$ – вероятности значений $x_1, \ldots, x_n$

Пример. Распределение Пуассона

$$p(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} \qquad L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!}$$

II. Непрерывное распределение

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

$f(x)$ – плотность распределения

Пример. Нормальное распределение, σ *– известно*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/2\sigma^2} \qquad L(a) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-a)^2/2\sigma^2}$$

Алгоритм исследования на максимум функции правдоподобия

1. $\ln L(\Theta) = \ln p(x_1) + \boxtimes + \ln p(x_n)$

   $\ln L(\Theta) = \ln f(x_1) + \boxtimes + \ln f(x_n)$

2. $\dfrac{d \ln L(\Theta)}{d\,\Theta}$

3. $\dfrac{d \ln L(\Theta^*)}{d\,\Theta} = 0$

4. $\dfrac{d^2 \ln L(\Theta^*)}{d\,\Theta^2} < 0 \quad \Rightarrow \quad \Theta^*$ – точка максимума

Метод моментов

I. Оценка одного параметра

$$M(X) = \overline{x_в}$$

Пример. Показательное распределение

$$f(x) = \lambda e^{-\lambda x} \qquad \lambda = ?$$

$$M(X) = \frac{1}{\lambda} \quad \Rightarrow \quad \frac{1}{\lambda} = \overline{x_в} \quad \Rightarrow \quad \lambda = \frac{1}{\overline{x_в}}$$

II. Оценка двух параметров

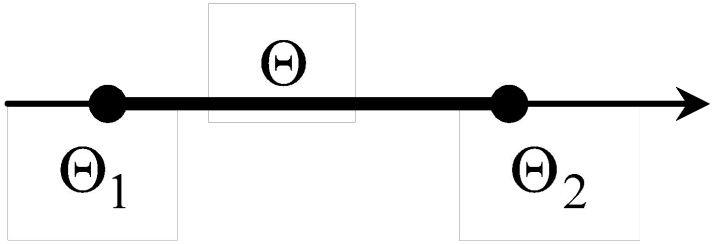
$$M(X) = \overline{x_в} \qquad D(X) = D_в$$

Пример. Нормальное распределение

$$a, \sigma = ? \quad M(X) = a \quad \Rightarrow \quad a = \overline{x_в}$$

$$D(X) = \sigma^2 \quad \Rightarrow \quad \sigma = \sqrt{D(X)} = \sqrt{D_в}$$

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



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

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

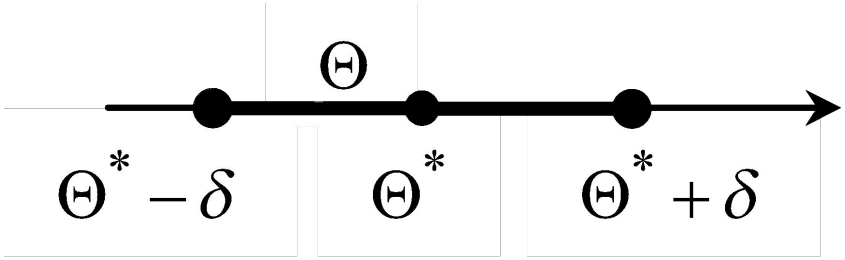
$\Theta_1 = f_1(x_1, x_2, \ldots, x_n)$ $\Theta_2 = f_2(x_1, x_2, \ldots, x_n)$

– 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

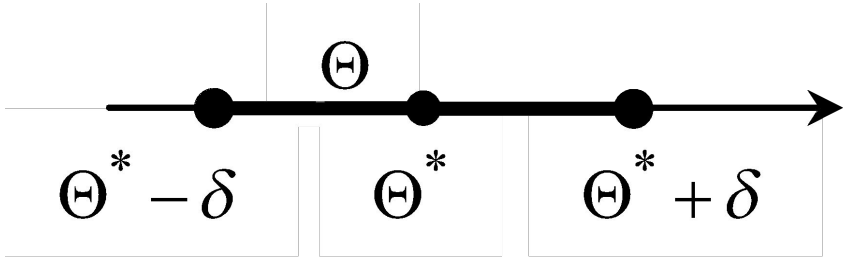
$$p(a < X < b) = F(b) - F(a) = \int_a^b f(x)dx \qquad (*)$$

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

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

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

Алгоритм нахождения доверительных интервалов



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

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

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

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

Уравнения для нахождения δ:

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

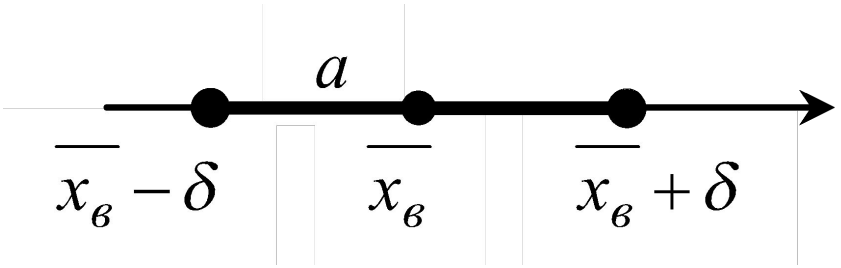
Вопрос: какой вид имеют функции $F(x)$ и $f(x)$?

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

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

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

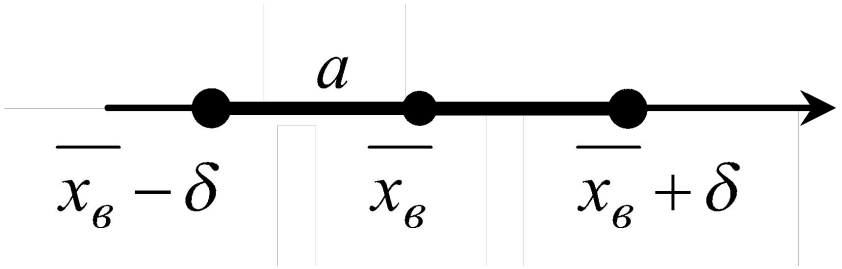
Шаг 2.

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

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

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:

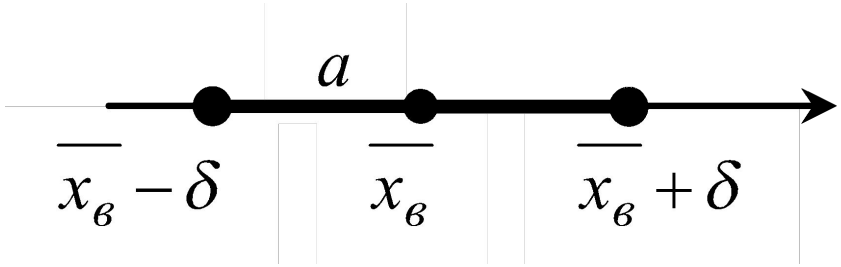
$$\left(\overline{x_в}-\frac{t_\gamma\cdot\sigma}{\sqrt{n}},\ \overline{x_в}+\frac{t_\gamma\cdot\sigma}{\sqrt{n}}\right)$$

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

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

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

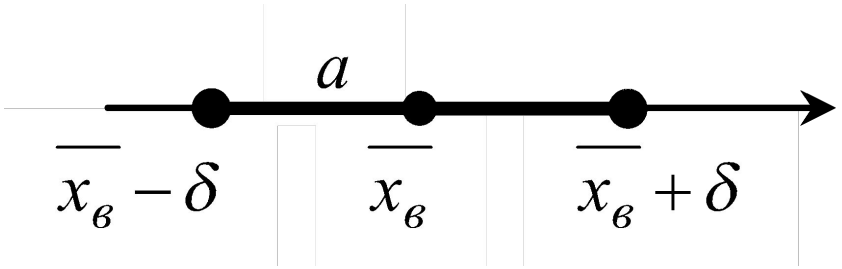
Шаг 2.

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

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

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:

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

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:

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

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:

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

Статистической гипотезой называется любое предположение о виде или параметрах неизвестного закона распределения.

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

Конкурирующей (альтернативной) называют гипотезу, которая противоречит нулевой, обозначают её H_1.

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