Юнит 1. Доверительные интервалы

Пусть $\vec{X} = (X_1, X_2, ..., X_n)$ — выборка объёма n из распределения F_θ с параметром $\theta \in \Theta \subseteq R$. *Определение.* Пусть $0 < \varepsilon < 1$. Интервал со случайными концами

$$(\underline{\theta}, \overline{\theta}) = (\underline{\theta}(\vec{X}, \varepsilon), \overline{\theta}(\vec{X}, \varepsilon))$$

Называется доверительным интервалом для параметра θ уровня доверия $1 - \varepsilon$, если для любого $\theta \in \Theta$ $\quad P\{\underline{\theta} < \theta < \overline{\theta}\} \geqslant 1 - \varepsilon$.

П р и м е р. Пусть $X_1, X_2, ..., X_n$ — выборка объёма n из нормального распределения $N(a, \sigma^2)$, где $a \in R$ —неизвестный параметр, а значение $\sigma > 0$ известно. Требуется при произвольном n построить точный доверительный интервал для параметра a уровня доверия $1 - \varepsilon$.

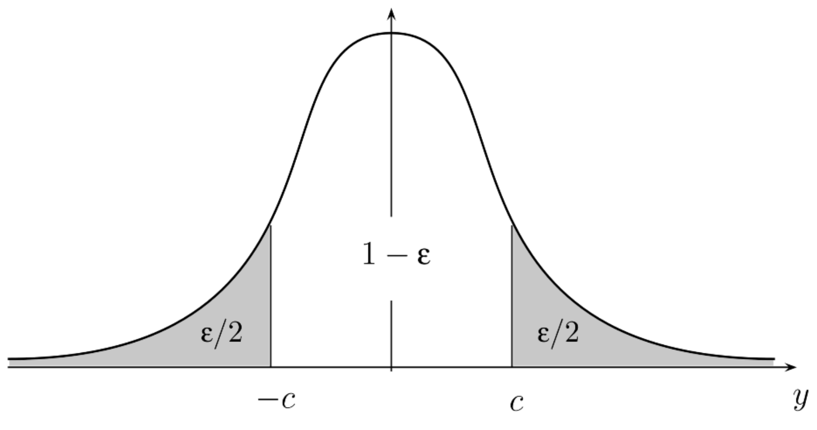
Знаем, что нормальное распределение устойчиво по суммированию. Поэтому распределение суммы элементов выборки при любом её объёме n нормально:

$$n\overline{x} = X_1 + X_2 + ... + X_n \in N(na, n\sigma^2),$$

$$\eta = \frac{n\overline{x} - na}{\sigma\sqrt{n}} \in N(0, 1)$$

По заданному $\varepsilon \in (0, 1)$ найдём число $c > 0$ такое, что $P\{-c < \eta < c\} = 1 - \varepsilon$.

Число c является квантилью уровня $1 - \frac{\varepsilon}{2}$ стандартного нормального распределения.



$$P\{-c < \eta < c\} = 2\Phi(c) = 1 - \varepsilon$$

Определение. Пусть распределение $\mathcal{F}$ с функцией распределения F абсолютно непрерывно. Число τ_δ называется квантилью уровня δ распределения F, если $F(\tau_\delta) = \delta$.

Если функция F строго монотонна, квантиль определяется единственным образом.

По заданному ε в таблице значений функции Лапласа найдём квантили $c = u_{1-\frac{\varepsilon}{2}}$.

Разрешив затем неравенство $-c < \eta < c$ относительно а, получим точный доверительный интервал:

$$1 - \varepsilon = P\{- c < \eta < c\} = P\left\{- u_{1-\frac{\varepsilon}{2}} < \frac{n\bar{x}-na}{\sigma\sqrt{n}} < u_{1-\frac{\varepsilon}{2}}\right\} = P\left\{\bar{x} - u_{1-\frac{\varepsilon}{2}}\frac{\sigma}{\sqrt{n}} < a < \bar{x} + u_{1-\frac{\varepsilon}{2}}\frac{\sigma}{\sqrt{n}}\right\}$$

Итак, получен точный доверительный интервал уровня доверия $1 - \varepsilon$.

Общий принцип построения точных доверительных интервалов.

Чтобы построить точный доверительный интервал, необходимо реализовать следующие шаги.

1. Найти функцию $G(\vec{X}, \theta)$, распределение которой не зависит от параметра θ. Необходимо, чтобы $G(\vec{X}, \theta)$, была обратима по θ при любом фиксированном $\vec{X}$.

2. Найти числа g_1 и g_2 —квантили распределения G, для которых

$$1 - \varepsilon = P\left\{g_1 < G(\vec{X}, \theta) < g_2\right\}$$

3. Разрешив неравенство $g_1 < G(\vec{X}, \theta) < g_2$ относительно θ, получить точный доверительный интервал.

Юнит 2. РАСПРЕДЕЛЕНИЯ, СВЯЗАННЫЕ С НОРМАЛЬНЫМ

Распределение χ^2 Пирсона.

Определение. Распределение суммы k квадратов независимых случайных величин со стандартным нормальным распределением называется распределением χ^2 (хи-квадрат) с k степенями свободы и обозначается H_k.

$$\xi_1, ..., \xi_k \text{ независимы и } \xi_i \in N(0, 1)$$

$\chi^2 = \sum_{i=1}^{k} \xi_i^2 \in H_k \qquad p_{\chi^2}(y) = \{\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} y^{\frac{k}{2}-1} e^{\frac{-y}{2}}$, если $y > 0$ 0, если $y \le 0$

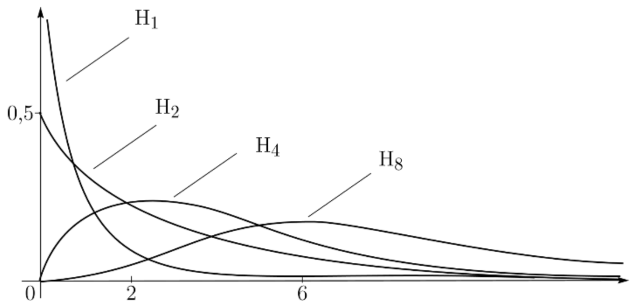


Рис. 8. Плотности χ^2-распределений с различным числом степеней свободы

Свойство 1. Если случайные величины $\chi^2 \in H_k$ и $\psi^2 \in H_m$ независимы, то их сумма $\chi^2 + \psi^2$ имеет распределение H_{k+m}.

Доказательство. Пусть случайные величины $\xi_1, ..., \xi_{k+m}$ независимы и $\xi_i \in N(0, 1)$. Тогда случайная величина χ^2 распределена так же, как $\xi_1^2 + ... + \xi_k^2$, величина ψ^2 распределена так же, как $\xi_{k+1}^2 + ... + \xi_{k+m}^2$, а их сумма — как $\xi_1^2 + ... + \xi_{k+m}^2$,, т.е. имеет распределение H_{k+m}.

Свойство 2. Если величина $\chi^2 \in H_k$, то $M\chi^2 = k$ и $D\chi^2 = 2k$.

Доказательство. Пусть ξ_1, ξ_2, ... независимы и имеют стандартное нормальное распределение. Тогда

$$M\xi_i = 0 \qquad M\xi_i^2 = D\xi_i = 1 \qquad D\xi_i^2 = M\xi_i^4 - \left(M\xi_i^2\right)^2 = 3 - 1 = 2$$

$$M\xi_i^4 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^4 e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^3 e^{-\frac{t^2}{2}} d\frac{t^2}{2} = \frac{-t^3 e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \Big|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} 3t^2 e^{-\frac{t^2}{2}} dt = 3M\xi_i^2 = 3$$

Поэтому

$$M\chi^2 = M\sum_{i=1}^{k}\xi_i^2 = \sum_{i=1}^{k} M\xi_i^2 = k$$

$$D\chi^2 = D\sum_{i=1}^{k}\xi_i^2 = \sum_{i=1}^{k} D\xi_i^2 = 2k$$

Свойство 3. Пусть $\chi_n^2 \in H_n$. Тогда при $n \to\infty$

$$\frac{\chi_n^2}{n}\ p\to 1\ \frac{\chi_n^2-n}{\sqrt{2n}} \Rightarrow N(0, 1).$$

Доказательство. При любом n случайная величина $\chi_n^2 = \sum_{i=1}^{n}\xi_i^2$ где все случайные величины ξ_i независимы и имеют стандартное нормальное распределение. Применяя ЗБЧ и ЦПТ, получаем сходимости

$$\frac{\chi_n^2}{n} = \frac{1}{n}\sum_{i=1}^{n}\xi_i^2\ p\to M\xi_i^2 = 1$$

$$\frac{\chi_n^2-n}{\sqrt{2n}} = \frac{\sum_{i=1}^{n}\xi_i^2-nM\xi_i^2}{\sqrt{nD\xi_i^2}} \Rightarrow N(0, 1)$$

Свойство 4. Если случайные величины $\xi_1, ..., \xi_k$ независимы и $\xi_i \in N\left(a, \sigma^2\right)$, то

$$\chi_k^2 = \sum_{i=1}^{k}\left(\frac{\xi_i-a}{\sigma}\right)^2 \in H_k$$

Распределение Стьюдента.

Определение . Пусть $\xi_0, \xi_1, ..., \xi_k$ независимы и $\xi_i \in N(0, 1)$. Распределение случайной величины

$$t_k = \frac{\xi_0}{\sqrt{\frac{1}{k}\sum_{i=1}^{k}\xi_i^2}}$$

называется распределением Стьюдента с k степенями свободы и обозначается T_k.

Распределение Стьюдента совпадает с распределением случайной величины

$$t_k = \frac{\xi_0}{\sqrt{\frac{1}{k}\chi_k^2}}$$

где $\xi_0 \in N(0, 1)$ и $\chi_k^2 \in H_k$ независимы.

$$p_{t_k}(y) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\bullet\Gamma\left(\frac{k}{2}\right)}\left(1 + \frac{y^2}{k}\right)^{-\frac{k+1}{2}}$$

Свойство1. Распределение Стьюдента симметрично: если случайная величина t_k имеет распределение Стьюдента T_k с k степенями свободы, то и $-\ t_k$ имеет такое же распределение.

Свойство 2. Распределение Стьюдента T_n слабо сходится к стандартному нормальному распределению при $n \rightarrow \infty$.

Доказательство. По свойству распределения $\chi^2 \in H_n$, $\frac{\chi_n^2}{n}\ p \rightarrow 1$ при $n \rightarrow \infty$.

По свойствам слабой сходимости получаем

$$t_k = \frac{\xi_0}{\sqrt{\frac{1}{k}\chi_k^{\ 2}}} \Rightarrow \xi_0 \in N(0, 1)$$

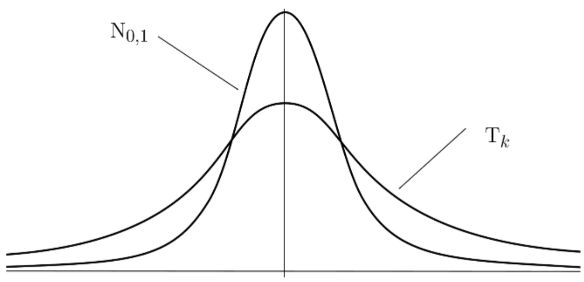


Рис. 9. Плотности распределений T_k и $N_{0,1}$

Распределение Фишера

Следующее распределение тоже тесно связано с нормальным распределением , но понадобится не при построении доверительных интервалов, а в задачах проверки гипотез Там же мы поймём, почему его называют распределением дисперсионного отношения .

Определение Пусть χ_k^2 имеет распределение H_k, а ψ_n^2—распределение H_n , причём эти случайные величины независимы. Распределение случайной величины

$$f_{k,n} = \frac{\frac{\chi_k^2}{k}}{\frac{\psi_n^2}{n}} = \frac{n}{k} \bullet \frac{\chi_k^2}{\psi_n^2}$$

называется распределением Фишера с k и n степенями свободы и обозначается $F_{k,n}$.

Свойства распределения Фишера (или Фишера—Снедекора)

Свойство1 Если случайная величина $f_{k,n}$ имеет распределение Фишера $F_{k,n}$, то $\frac{1}{f_{k,n}}$ имеет распределение Фишера $F_{n,k}$

Заметим, что функции распределения $F_{k,n}$ и $F_{n,k}$ различаются, но связаны соотношением:

$$\forall x > 0,\ F_{n,k}(x) = P\left\{f_{n,k} < x\right\} = P\left\{\frac{1}{f_{n,k}} > \frac{1}{x}\right\} = 1 - F_{k,n}\left(\frac{1}{x}\right)$$

Распределение Фишера табулировано при многих k и n, причём свойство 1 позволяет приводить таблицы распределений только например при $k \geq n$.

Свойство2. Распределение Фишера $F_{k,n}$ слабо сходится к вырожденному в точке $c = 1$ распределению при любом стремлении k и n к бесконечности.

Доказательство

Пусть $\xi_0, \xi_1, \ldots$ и $\eta_0, \eta_1, \ldots$ —две независимые последовательности, составленные из независимых случайных величин со стандартным нормальным распределением.

Тогда по ЗБЧ при $k \to \infty,\ n \to \infty$

$$\frac{1}{k}\sum_{i=1}^{k} \xi_i^2 \ p\to\ M\xi_i^2 = 1$$

$$\frac{1}{n}\sum_{i=1}^{n} \eta_i^2 \ p\to\ M\eta_i^2 = 1$$

$$f_{k,n} = \frac{\frac{\chi_k^2}{k}}{\frac{\psi_n^2}{n}} = \frac{\frac{1}{k}\sum_{i=1}^{k}\xi_i^2}{\frac{1}{n}\sum_{i=1}^{n}\eta_i^2} \ p\to\ 1,$$

*Свойство*3. Пусть $t_k \in T_k$ - случайная величина, имеющая распределение Стьюдента.

Тогда $t_k^2 \in F_{1,k}$.

Юнит 3. Преобразования нормальных выборок

Пусть $\vec{X} = (X_1, X_2, ..., X_n)$ — выборка объёма n из стандартного нормального распределения.

Там, где нам понадобятся операции матричного умножения, будем считать $\vec{X}$ вектором-столбцом. Пусть C —ортогональная матрица $(n\times n)$, т.е.

$C\cdot C^T = E = (1 \cdots 0 \vdots \ddots \vdots 0 \cdots 1)$, и $\vec{Y} = C\vec{X}$ — вектор с координатами $Y_i = C_{i1}X_1 + ... + C_{in}X_n$. Координаты вектора $\vec{Y}$ имеют нормальные распределения как линейные комбинации независимых нормальных величин.

Утверждение. (без доказательства)

Пусть вектор $\vec{X}$ состоит из независимых случайных величин со стандартным нормальным распределением, C —ортогональная матрица, $\vec{Y} = C\vec{X}$. Тогда и координаты вектора $\vec{Y}$ независимы и имеют стандартное нормальное распределение.

Лемма Фишера. Пусть вектор $\vec{X}$ состоит из независимых случайных величин со стандартным нормальным распределением, C —ортогональная матрица, $\vec{Y} = C\vec{X}$. Тогда при любом $k = 1, ..., n - 1$ случайная величина

$$T(\vec{X}) = \sum_{i=1}^{n} X_i^2 - Y_1^2 - ... - Y_k^2$$

не зависит от $Y_1, ..., Y_k$ и имеет распределение H_{n-k}.

Доказательство. Нормы векторов $\vec{X}$ и $\vec{Y} = C\vec{X}$ совпадают:

$$Y_1^2 + ... + Y_n^2 = \| C\vec{X} \|^2 = (C\vec{X})^T(C\vec{X}) = \left(\vec{X}^T C^T\right)(C\vec{X}) = \left(\vec{X}^T \vec{X}\right) = \| \vec{X} \|^2 = X_1^2 + ... + X_n^2 .$$

Поэтом

$$T(\vec{X}) = \sum_{i=1}^{n} X_i^2 - Y_1^2 - ... - Y_k^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 - ... - Y_k^2 = \sum_{i=k+1}^{n} Y_i^2$$

Случайные величины $Y_1, ..., Y_n$ независимы и имеют стандартное нормальное распределение, поэтому

$$T(\vec{X}) = \sum_{i=k+1}^{n} Y_i^2 \in H_{n-k}$$

и не зависит от $Y_1, ..., Y_k$.

Теорема (основное следствие леммы Фишера).

Пусть $X_1,\ ...,\ X_n$ независимы и $X_i \in N\left(a, \sigma^2\right)$. Тогда:

$$\frac{(n-1)S_0^2}{\sigma^2} = \sum_{i=1}^{n} \frac{\left(X_i - \bar{x}\right)^2}{\sigma^2} \in H_{n-1}$$

и случайные величины $\bar{x}$ и S_0^2 независимы.

Доказательство. Убедимся сначала, что можно рассматривать выборку из стандартного нормального распределения вместо $N\left(a, \sigma^2\right)$:

$$z_i = \frac{X_i - a}{\sigma} \in N(0, 1) \qquad \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i - a}{\sigma} = \frac{\bar{x} - a}{\sigma}$$

$$\frac{(n-1)S_0^2}{\sigma^2} = \sum_{i=1}^{n} \frac{\left(X_i - \bar{x}\right)^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - a - (\bar{x} - a)}{\sigma}\right)^2 = \sum_{i=1}^{n} \left(z_i - \bar{z}\right)^2$$

Итак, можно с самого начала считать, что X_i имеют стандартное нормальное распределение, $a = 0,\ \ \sigma = 1$. Применим лемму Фишера.

$$T\left(\vec{X}\right) = (n - 1)S_0^2 = \sum_{i=1}^{n} \left(X_i - \bar{x}\right)^2 = \sum_{i=1}^{n} X_i^{\ 2} - n(\bar{x})^2 = \sum_{i=1}^{n} X_i^{\ 2} - Y_1^{\ 2}$$

$$Y_1 = \sqrt{n}\bar{x} = \frac{\sqrt{n}}{n}\sum_{i=1}^{n} X_i = \frac{X_1}{\sqrt{n}} + \ ...\ + \frac{X_n}{\sqrt{n}}$$

Чтобы применить лемму Фишера, нужно найти ортогональную матрицу C такую, что Y_1 будет первой координатой вектора $\vec{Y} = C\vec{X}$. Возьмём матрицу C с первой строкой $(1/\sqrt{n},\ ...,\ 1/\sqrt{n})$. Так как длина (норма) этого вектора равна единице, его можно дополнить до ортонормального базиса в R^n, и C можно дополнить до ортогональной матрицы. Тогда величина $Y_1 = \sqrt{n}\bar{x}$ и будет первой координатой вектора $\vec{Y} = C\vec{X}$.

$$T\left(\vec{X}\right) = \sum_{i=1}^{n} X_i^{\ 2} - Y_1^{\ 2} \in H_{n-1}$$

и не зависит от $Y_1 = \sqrt{n}\bar{x}$, т.е. $\bar{x}$ и S_0^2 независимы.

Юнит 4. Доверительные интервалы для параметров нормального распределения

Очередное утверждение позволит нам строить **доверительные интервалы для параметров нормального распределения.**

Теорема Пусть X_1, ..., X_n независимы и $X_i \in N(a, \sigma^2)$. Тогда:

1)

$$\frac{\bar{x} - a}{\sqrt{\frac{\sigma^2}{n}}} \in N(0, 1) \text{ (для } a \text{ при } \sigma \text{ известном),}$$

2)

$$\sum_{i=1}^{n} \frac{(X_i - a)^2}{\sigma^2} = H_n \text{ (для } \sigma^2 \text{ при } a \text{ известном),}$$

3)

$$\frac{(n-1)S_0^2}{\sigma^2} \in H_{n-1} \quad \text{(для σ2 при } a \text{ неизвестном),}$$

4)

$$\frac{\bar{x} - a}{\sqrt{\frac{S_0^2}{n}}} \in T_{n-1} \text{ (для } a \text{ при } \sigma \text{ неизвестном).}$$

Доказательство.

1)

$$X_i \in N(a, \sigma^2) \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \in N\left(a, \frac{\sigma^2}{n}\right) \qquad \frac{\bar{x} - a}{\sqrt{\frac{\sigma^2}{n}}} \in N(0, 1) \quad \blacksquare$$

2)

$$\frac{X_i - a}{\sigma} \in N(0, 1) \qquad \sum_{i=1}^{n} \frac{(X_i - a)^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - a}{\sigma}\right)^2 \in H_n \qquad \blacksquare$$

3)

$$\frac{(n-1)S_0^2}{\sigma^2} \in H_{n-1} \text{ следует из леммы Фишера} \qquad \blacksquare$$

4)

$$\xi_0 = \frac{\bar{x} - a}{\sqrt{\frac{\sigma^2}{n}}} \in N(0, 1) \qquad \chi^2_{n-1} = \frac{(n-1)S_0^2}{\sigma^2} \in H_{n-1} \text{ независимы по лемме Фишера}$$

$$t_{n-1} = \frac{\xi_0}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} = \frac{\bar{x} - a}{\sqrt{\frac{\sigma^2}{n}}} \frac{1}{\sqrt{\frac{1}{(n-1)} \frac{(n-1)S_0^2}{\sigma^2}}} = \frac{\bar{x} - a}{\sqrt{\frac{S_0^2}{n}}} \in T_{n-1} \qquad \blacksquare$$

Точные доверительные интервалы для параметров нормального распределения

Пусть X_1, ..., X_n - выборка объёма n из распределения $N\left(a, \sigma^2\right)$. Построим точные доверительные интервалы (ДИ) с уровнем доверия $1 - \varepsilon$ для параметров нормального распределения.

П р и м е р (ДИ для a при известном σ^2). Этот интервал мы построили

$$1 - \varepsilon = P\left\{-u_{1-\frac{\varepsilon}{2}} < \frac{\bar{x} - a}{\sqrt{\frac{\sigma^2}{n}}} < u_{1-\frac{\varepsilon}{2}}\right\} = P\left\{\bar{x} - u_{1-\frac{\varepsilon}{2}}\frac{\sigma}{\sqrt{n}} < a < \bar{x} + u_{1-\frac{\varepsilon}{2}}\frac{\sigma}{\sqrt{n}}\right\} \quad \text{где } \Phi(u_{1-\frac{\varepsilon}{2}}) = \frac{1-\varepsilon}{2}$$

П р и м е р (ДИ для σ^2 при известном a).

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

Пусть $h_{n,\varepsilon/2}$ и $h_{n,1-\varepsilon/2}$ —квантили распределения H_n уровней $\varepsilon/2$ и $1 - \varepsilon/2$ соответственно.

$$1 - \varepsilon = P\left(h_{n,\varepsilon/2} < \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - a\right)^2 < h_{n,1-\varepsilon/2}\right) = P\left(\frac{\sum_{i=1}^{n}\left(X_i - a\right)^2}{h_{n,1-\varepsilon/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}\left(X_i - a\right)^2}{h_{n,\varepsilon/2}}\right)$$

П р и м е р (ДИ для σ^2 при неизвестном a).

$$\frac{(n-1)S_0^2}{\sigma^2} \in H_{n-1}$$

Пусть $h_{n-1,\varepsilon/2}$ и $h_{n-1,1-\varepsilon/2}$ —квантили распределения H_{n-1} уровней $\varepsilon/2$ и $1 - \varepsilon/2$ соответственно.

$$1 - \varepsilon = P\left(h_{n-1,\varepsilon/2} < \frac{(n-1)S_0^2}{\sigma^2} < h_{n-1,1-\varepsilon/2}\right) = P\left(\frac{(n-1)S_0^2}{h_{n-1,1-\varepsilon/2}} < \sigma^2 < \frac{(n-1)S_0^2}{h_{n-1,\varepsilon/2}}\right)$$

П р и м е р (ДИ для a при неизвестном σ^2).

$$\frac{\bar{x} - a}{\sqrt{\frac{S_0^2}{n}}} \in T_{n-1}$$

Пусть $t_{n-1,1-\varepsilon/2}$ —квантиль распределения T_{n-1}уровня $1 - \varepsilon/2$. Распределение Стьюдента симметрично.

$$1 - \varepsilon = P\left\{-t_{n-1,1-\varepsilon/2} < \frac{\bar{x} - a}{\sqrt{\frac{S_0^2}{n}}} < t_{n-1,1-\varepsilon/2}\right\} = P\left\{\bar{x} - t_{n-1,1-\varepsilon/2}\sqrt{\frac{S_0^2}{n}} < a < \bar{x} + t_{n-1,1-\varepsilon/2}\sqrt{\frac{S_0^2}{n}}\right\}$$

Юнит 5. Асимптотический (асимптотически точный) доверительный интервал

Пример 1

Пусть $\vec{X} = (X_1, X_2, ..., X_n)$ выборка объёма n из показательного распределения, $(\lambda > 0)$

Требуется построить асимптотический (асимптотически точный) доверительный интервал для параметра λ уровня доверия $1 - \varepsilon$

Вспомним ЦПТ

$$\frac{\sum_{i=1}^{n} X_i - nMX_i}{\sqrt{nDX_i}} = \sqrt{n} \bullet \frac{\bar{x} - \frac{1}{\lambda}}{\sqrt{\frac{1}{\lambda^2}}} = \sqrt{n} \bullet (\lambda\bar{x} - 1) \Rightarrow \eta \in N(0, 1)$$

Возьмём $c = u_{1-\frac{\varepsilon}{2}}$ —квантиль стандартного нормального распределения. По определению слабой сходимости при $n \to \infty$

$$P\left\{- u_{1-\frac{\varepsilon}{2}} < \sqrt{n} \bullet (\lambda\bar{x} - 1) < u_{1-\frac{\varepsilon}{2}}\right\} \to P\{- c < \eta < c\} = 1 - \varepsilon$$

Разрешив относительно λ неравенство получим асимптотический доверительный интервал.

$$P\left\{\frac{1}{\sqrt{n}x}\left(\sqrt{n} - u_{1-\frac{\varepsilon}{2}}\right) < \lambda < \frac{1}{\sqrt{n}x}\left(\sqrt{n} + u_{1-\frac{\varepsilon}{2}}\right)\right\} \to 1 - \varepsilon \quad \text{при } n \to \infty$$

Пример 2

Пусть $\vec{X} = (X_1, X_2, ..., X_n)$ выборка объёма n из распределения *Пуассона*, $(\lambda > 0)$

Требуется построить асимптотический доверительный интервал для параметра λ уровня доверия $1 - \varepsilon$

Согласно ЦПТ

$$\frac{\sum_{i=1}^{n} X_i - nMX_i}{\sqrt{nDX_i}} = \sqrt{n} \bullet \frac{\bar{x} - \lambda}{\sqrt{\lambda}} \Rightarrow \eta \in N(0, 1)$$

Пусть $c = u_{1-\frac{\varepsilon}{2}}$ —квантиль уровня $1 - \frac{\varepsilon}{2}$ стандартного нормального распределения.

По определению слабой сходимости при $n \to \infty$

$$P\left\{- u_{1-\frac{\varepsilon}{2}} < \sqrt{n} \bullet \frac{\bar{x} - \lambda}{\sqrt{\lambda}} < u_{1-\frac{\varepsilon}{2}}\right\} \to P\{- c < \eta < c\} = 1 - \varepsilon$$

Но разрешить неравенство под знаком вероятности относительно λ непросто, мешает корень в знаменателе. Попробуем от него избавиться. Состоятельной оценкой $\hat{\lambda} = \bar{x}$. Тогда при $n \to \infty$

$$\sqrt{\frac{\lambda}{\bar{x}}}\, p\!\rightarrow 1$$

$$\sqrt{n} \bullet \frac{\bar{x}-\lambda}{\sqrt{\lambda}} \bullet \sqrt{\frac{\lambda}{\bar{x}}} = \sqrt{n} \bullet \frac{\bar{x}-\lambda}{\sqrt{\bar{x}}} \Rightarrow \eta \in N(0, 1)$$

Поэтому при $n \rightarrow \infty$

$$P\left\{- u_{1-\frac{\varepsilon}{2}} < \sqrt{n} \bullet \frac{\bar{x}-\lambda}{\sqrt{\bar{x}}} < u_{1-\frac{\varepsilon}{2}}\right\} \rightarrow P\{- c < \eta < c\} = 1 - \varepsilon$$

Разрешив неравенство под знаком вероятности относительно λ

получим искомый асимптотический доверительный интервал

$$P\left\{\bar{x} - u_{1-\frac{\varepsilon}{2}} \frac{\sqrt{\bar{x}}}{\sqrt{n}} < a < \bar{x} + u_{1-\frac{\varepsilon}{2}} \frac{\sigma\sqrt{\bar{x}}}{\sqrt{n}}\right\} = 1 - \varepsilon$$

Юнит 6. ПРОВЕРКА ГИПОТЕЗ

Пусть дана выборка $\vec{X} = (X_1, X_2, ..., X_n)$ из распределения F.

Определение. **Гипотезой** (H) называется любое предположение относительно вида распределения, параметров распределения или свойств закона распределения наблюдаемой в эксперименте случайной величины: $H = \{F = F_1\}$ или $H = \{F \in F\}$, где $\mathbb{F}$ —некоторое подмножество в множестве всех распределений.

Проверяемая гипотеза называется основной (или нулевой) и обозначается H_0. Гипотеза, конкурирующая с H_0, называется альтернативной и обозначается H_1.

Гипотеза Н называется простой, если она указывает на единственное распределение: $F = F_1$. Иначе Н называется сложной: $F \in F$.

Определение. **Критерием** $\delta = \delta(X_1, ..., X_n)$ называется измеримое отображение

$$\delta: R^n \to \left\{H_0, ..., H_k\right\}$$

из множества всех возможных значений выборки в множество гипотез. Измеримость понимается в обычном смысле: $\{\omega | \delta(X_1, ..., X_n) = H_i\}$ есть событие при любом $i = 1,..., k$.

Определение. Говорят, что произошла ошибка i-го рода критерия δ, если критерий отверг верную гипотезу H_i. Вероятностью ошибки i-го рода критерия δ называется число

$$\alpha_i(\delta) = P\left\{(\delta(\vec{X}) \neq H_i)/H_i\right\}$$

П р и м е р. Пусть любое изделие некоторого производства оказывается браком с вероятностью p. Контроль продукции допускает ошибки: годное изделие бракует с вероятностью γ, а бракованное пропускает (признаёт годным) с вероятностью ε. Если ввести для проверяемого изделия гипотезы $H_1 = \{\text{изделие годное}\}$ и $H_2 = \{\text{изделие бракованное}\}$, а критерием выбора одной из них считать контроль продукции, то γ —вероятность ошибки первого рода этого критерия, а ε —второго рода:

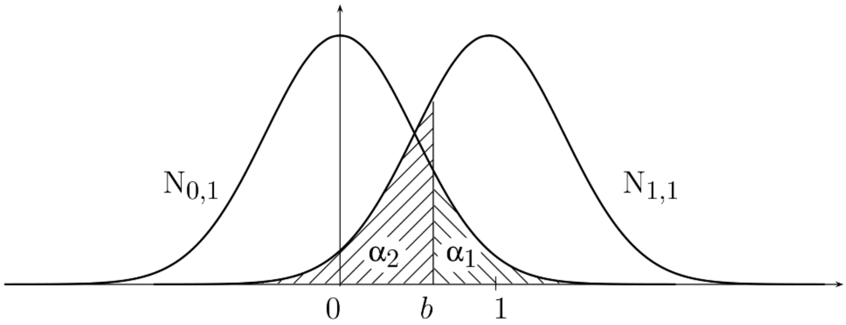
$$\gamma = P\left\{\left(\delta = H_2\right)/H_1\right\} = P(\text{контроль забраковал годное изделие}) = \alpha_1(\delta);$$

$$\varepsilon = P\left\{(\delta = H_1)/H_2\right\} = P(\text{контроль пропустил бракованное изделие}) = \alpha_2(\delta)$$

П р и м е р. Имеется выборка объёма $n = 1$ из нормального распределения $N(a, 1)$ и две простые гипотезы $H_1 = \{a = 0\}$ и $H_2 = \{a = 1\}$. Рассмотрим при некотором $b \in R$ следующий критерий: $\delta(\vec{X}) = \{H_1 \text{ если } X_1 \le b \; H_2 \text{ если } X_1 > b$.

Изобразим на графике соответствующие гипотезам плотности распределений и вероятности ошибок первого и второго рода критерия δ

$\alpha_1(\delta) = P\{(X_1 > b)/H_1\}, \alpha_2(\delta) = P\{(X_1 \le b)/H_2\}.$



Видим, что с ростом числа b вероятность ошибки первого рода α_1 уменьшается, но вероятность ошибки второго рода α_2 растёт

Определение. **Статистикой критерия** называют некоторую числовую функцию $T = \varphi(\vec{x}) = \varphi(x_1, ... x_n)$ выборки $\vec{x}$, обладающую тем свойством, что её закон распределения $F_T(z)$ полностью известен в том случае, когда проверяемая гипотеза H_0 верна.

Определение **Критической областью** G статистического критерия называется область реализаций t статистики $T = \varphi(\vec{x})$, при которых гипотеза H_0отвергается.

Определение. **Уровнем значимости** статистического критерия называется вероятность отвергнуть основную гипотезу H_0, если она верна(вероятность ошибки 0-го рода $\alpha = P\{T \in G | H_0 \text{ верна}\})$.

Проверка статистической гипотезы может быть подразделена на следующие этапы:

1) сформулировать проверяемую гипотезу H_0 и альтернативную к ней гипотезу H_1;

2) выбрать уровень значимости α;

3) выбрать статистику T для проверки гипотезы H_0;

4) найти распределение $F_T(z)$ статистики T, при условии что гипотеза H_0 верна;

5) построить, в зависимости от формулировки гипотезы H1 и уровня значимости α, критическую область G, выписать критерий $\delta(\vec{X})$;

6) получить реализацию выборки $(x_1, ... x_n)$ и вычислить реализацию $T = \varphi(x_1, ... x_n)$ статистики T критерия;

7) принять статистическое решение на уровне значимости α: если $t \in G$(критической области), то отклонить гипотезу H_0 как не согласующуюся с результатами наблюдений, а если $t \notin G$, то принять гипотезу H0 как не противоречащую результатам наблюдений.

Юнит 7. Проверка гипотез о параметрах в одновыборочной гауссовской модели и биномиальных моделях

Рассмотрим процедуру проверки параметрической гипотезы на примере одной из "старинных" статистических задач.

Пример. Рассмотрим следующую статистическую модель. Проводится серия из n испытаний Бернулли. Пусть случайная величина ξ - число «успехов» в n испытаниях, тогда ξ имеет биномиальное распределение Bi(n,p). Как мы уже знаем, неизвестную вероятность p можно оценить частотой «успехов» $\hat{p} = \frac{X}{n}$, и оценка $\hat{p}$ обладает следующими свойствами: несмещенная, т.е. $M[\hat{p}] = p$; состоятельная;

$X = \sum_{i=1}^{n} I(\text{"успех"})$, поэтому $X \Rightarrow N(np,\ np(1-p)$ $\text{ при } n \to \infty$

$$\hat{p} = \frac{X}{n} \Rightarrow N\left(p, \quad \frac{p(1-p)}{n}\right)$$

В практических задачах бывает важно не только оценить вероятность «успеха» *p*, но и проверить гипотезу о том, что p равна некоторой заданной величине p_0.

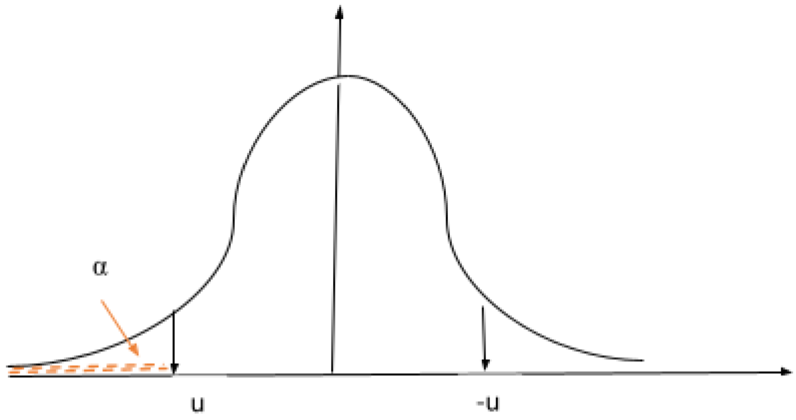
Построим критерий, называемый биномиальным критерием, для проверки гипотезы $H_0 : \{p = p_0\}$ против альтернатив следующего вида: $H_1 : \{p < p_0\}$ $H_2 : \{p > p_0\}$ $H_3 : \{p \neq p_0\}$. Рассмотрим статистику $\frac{X}{n}$, где *X* – количество «успехов» в n испытаниях. Если наблюдений достаточно много, то согласно теореме Муавра-Лапласа, статистика $T = \frac{X}{n}$ будет иметь асимптотически нормальное распределение. То есть при $n > 30$ можно в качестве статистики биномиального критерия выбрать статистику

$$T = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \Rightarrow N(0, \quad 1)$$

Тогда в случае $H_1 : \{p < p_0\}$

$\delta(\vec{X}) = \{H_0 \text{ если } T > u \ H_1 \text{ если } T \leq u$, где *u* квантиль нормального распределения уровня α.

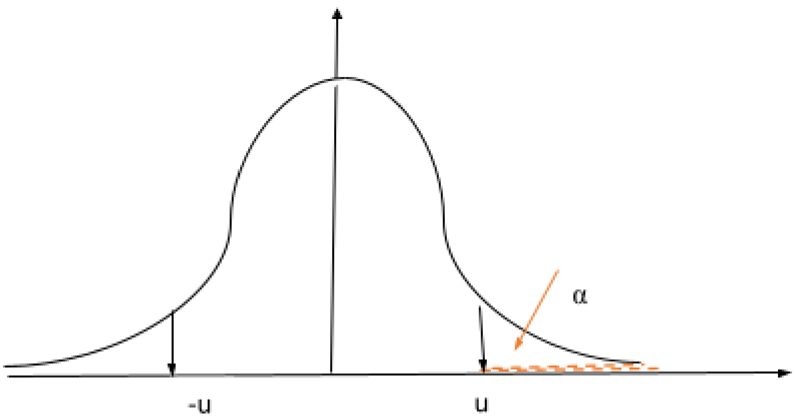
$$\Phi(-u) = 0,5 - \alpha$$



В случае $H_2: \{p > p_0\}$

$\delta(\vec{X}) = \begin{cases} H_0 & \text{если } T < u \\ H_2 & \text{если } T \geq u \end{cases}$, где u квантиль нормального распределения уровня 1-α.

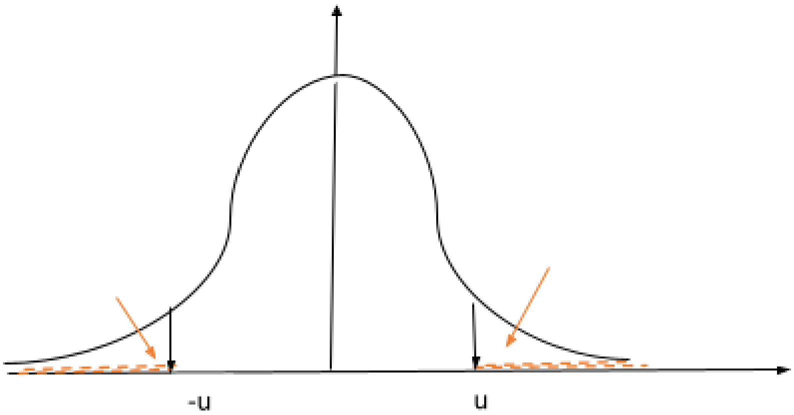
$$\Phi(u) = 0,5 - \alpha$$



В случае $H_3: \{p \neq p_0\}$

$\delta(\vec{X}) = \begin{cases} H_0 & \text{если } |T| < u \\ H_2 & \text{если } |T| \geq u \end{cases}$, где u квантиль нормального распределения уровня $\frac{\alpha}{2}$.

$$\Phi(u) = \frac{1-\alpha}{2}$$



Задача 6.1. (о леди, дегустирующей чай) Согласно принятой в Англии традиции чаепития, в чашку следует сначала наливать молоко, а потом – чай. Считается, что настоящая английская леди умеет отличить «правильный» чай от «неправильного». Чтобы выяснить дегустаторские качества дамы, был проведен следующий эксперимент: в течение 30 дней даме каждый день подавали пару чашек чая, в одну из которых сначала был налит чай, а в другую – молоко. Дама 21 раз верно указала «правильный» чай. Можно ли (на уровне доверия 0.95) считать ее хорошим дегустатором?

Р е ш е н и е. Пусть p – вероятность выбора «правильной» чашки чая. Тогда утверждение о том, что p = 0.5 соответствует ситуации, при которой выбор «правильной» чашки осуществляется случайным образом. Если же p > 0.5 то это означает, что выбор чашки основан на каких-то предпочтениях, т.е. неслучаен. Теперь можем провести процедуру проверки гипотезы.

1. В качестве основной гипотезы следует выбрать простую гипотезу $H_0 : \{p = 0,5\}$, а в качестве альтернативной – сложную гипотезу $H_1 : \{p > 0,5\}$2. Пусть уровень значимости $\alpha = 0.05$.

3. Выберем статистику

$$T = \frac{\frac{X}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \Rightarrow N(0,\ 1)$$

, где $p_0 = 0.5$

4. Альтернативной гипотезе должны соответствовать большие значения статистики *T*, т.е. критическая область расположена справа. Критическая точка $u_{0.95} = 1.65$.

$$\delta(\vec{X}) = \{H_0 \text{ если } T_{\text{набл}} < u = 1.65 \ \ H_1 \text{ если } T_{\text{набл}} \geq u = 1.65$$

5. Вычислим реализацию статистики

$$T_{\text{набл}} = \frac{\frac{X}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{21-30\cdot 0,5}{\sqrt{30\cdot 0,5(1-0,5)}} = 2,91$$

7. Реализация статистики попала в критическую область, следовательно, гипотеза H_0 отвергается. Следовательно, на уровне доверия 0.95 можно считать, что леди – хороший дегустатор.

Юнит 8. Критерии согласия: критерий Пирсона Χ^2

Имеется выборка $\vec{X} = \left(X_1, ..., X_n\right)$ из распределения F.

Проверяется простая гипотеза $H_1 = \left\{F = F_1\right\}$ против сложной альтернативы $H_2 = \left\{F \neq F_1\right\}$.

Критерий χ^2 основывается на группированных данных. Область значений предполагаемого распределения F_1 делят на некоторое число интервалов. После чего строят функцию отклонения ρ по разностям теоретических вероятностей попадания в интервалы группировки и эмпирических частот.

Пусть A_1, ..., A_k — интервалы группировки в области значений случайной величины с распределением F_1.

Обозначим для $j = 1, ..., k$ через ν_i число элементов выборки, попавших в интервал $A_i = [a_i, a_{i+1}[$

$$\nu_i = \sum_{k=1}^{n} I\left\{x_k \in [a_i, a_{i+1}[\right\}$$

, и через $p_i > 0$ — теоретическую вероятность попадания в интервал $A_i = [a_i, a_{i+1}[$

случайной величины ξ с распределением F_1. $p_i = P\left\{\xi \in A_i\right\}$ $\sum_{i=1}^{k} p_i = 1$

Как правило, длины интервалов выбирают так, чтобы $p_1 \approx p_2 \approx ... \approx p_k \approx \frac{1}{k}$

$$\rho(\vec{X}) = \sum_{i=1}^{k} \frac{\left(\nu_i - np_i\right)^2}{np_i}$$

Теорема Пирсона. Если верна гипотеза $H_1 = \left\{F = F_1\right\}$, то при фиксированном k и при

$$n \to \infty \qquad \rho(\vec{X}) = \sum_{i=1}^{k} \frac{\left(\nu_i - np_i\right)^2}{np_i} \Rightarrow H_{k-1}$$

Доказательство (для k=2)

Область значений предполагаемого распределения F_1 делим на два интервала.

$\nu_2 = n - \nu_1 \quad p_2 = 1 - p_1$

$$\rho\left(\vec{X}\right) = \frac{\left(\nu_1 - np_1\right)^2}{np_1} + \frac{\left(\nu_2 - np_2\right)^2}{np_2} = \frac{\left(\nu_1 - np_1\right)^2}{np_1} + \frac{\left(n - \nu_1 - n\left(1 - p_1\right)\right)^2}{n\left(1 - p_1\right)} = \frac{\left(\nu_1 - np_1\right)^2}{np_1} + \frac{\left(\nu_1 - np_1\right)^2}{n\left(1 - p_1\right)} = \frac{\left(\nu_1 - np_1\right)^2}{n}\left(\frac{1}{p_1} + \frac{1}{1 - p_1}\right) = \frac{\left(\nu_1 - n\right.}{np_1\left(1\right.}$$

$$\nu_1 = \sum_{k=1}^{n} I\left\{x_k \in A_1\right\}$$

$$p = p_1 \quad q = p_2 = 1 - p_1$$

$$M\nu_1 = np_1$$

$$D\nu_1 = np_1\left(1 - p_1\right)$$

$$\text{По ЦПТ} \quad \frac{\nu_1 - np_1}{\sqrt{np_1\left(1-p_1\right)}} \Rightarrow \xi \in N(0, 1)$$

$$\rho\left(\vec{X}\right) = \frac{\left(\nu_1 - np_1\right)^2}{np_1\left(1-p_1\right)} \Rightarrow \xi^2 \in H_1 \ \blacksquare$$

а) Если H_1 верна, то X_i имеют распределение F_1. По теореме Пирсона $\rho(\vec{X}) \Rightarrow \eta$, где η имеет распределение χ^2 с k-1 степенями свободы.

б) Если гипотеза H_1 неверна, то X_i имеют какое-то распределение F_2, отличное от F_1.

$$q_i = P\left\{\xi \in A_i\right\} \text{если } \xi \in F_2$$

По ЗБЧ

$$\nu_i = \sum_{k=1}^{n} I\left\{x_k \in A_i\right\} \qquad \frac{\nu_i}{n} \ p\to q_i \neq p_i$$

$$\text{При } n\to\infty \qquad \frac{\left(\nu_i - np_i\right)^2}{np_i} = \frac{n}{p_i}\left(\frac{\nu_i}{n} - p_i\right)^2 \ p\to \infty$$

Можем построить критерий согласия.

Пусть случайная величина $\chi^2 \in H_{k-1}$*. По таблице распределения* χ^2 *с* $k - 1$ *степенями свободы найдем C равное квантили уровня 1 − ε этого распределения. Тогда* $\varepsilon = P\left\{\chi^2 > C\right\}$ *и критерий согласия* χ^2 *выглядит как все критерии согласия:*

$$\delta(\vec{X}) = \{H_1 \text{ если } \rho\left(\vec{X}\right) < C \ \ H_2 \text{ если } \rho\left(\vec{X}\right) \geq C$$

Замечание. На самом деле критерий, который мы построим по функции $\rho\left(\vec{X}\right)$, решает совсем иную задачу. А именно, пусть задан набор вероятностей $p_1,..., p_k$ такой, что $\sum_{i=1}^{k} p_i = 1$.

Критерий χ^2 предназначен для проверки сложной гипотезы

$H_1 = \{F_1$ обладает свойством: $p_i = P\left\{\xi \in A_i\right\}$ для всех $i = 1,..., k\}$

против сложной альтернативы $H_2 = \left\{H_1 \text{ неверна}\right\}$, т.е.

$H_2 = \{$хотя бы для одного из интервалов вероятность $P\left\{\xi \in A_i\right\}$ отличается от $p_i\}$.

На самом деле критерий χ^2 применяют и для решения первоначальной задачи о проверке гипотезы H1 = {F = F1}. Необходимо только помнить, что этот критерий не состоятелен для альтернатив с теми же вероятностями попадания в интервалы разбиения, что и у F_1. Поэтому берут большое число интервалов разбиения — чем больше, тем лучше, чтобы «уменьшить» число альтернатив, неразличимых с предполагаемым распределением.

Но сходимость по распределению $\rho(\vec{X}) \Rightarrow H_{k-1}$ обеспечивается ЦПТ Нельзя, чтобы np_i было маленьким!!!

Маленькие значения np_i в знаменателе приведут к тому, что распределение $\rho(\vec{X})$ будет существенно отличаться от H_{k-1}. Тогда и реальная вероятность $P\left\{\rho(\vec{X}) > C\right\}$ — точный размер полученного критерия — будет сильно отличаться от ε. Поэтому для выборки объема n число интервалов разбиения выбирают так, чтобы обеспечить нужную точность при замене распределения $\rho(\vec{X})$ на H_{k-1}.

Обычно требуют, чтобы $np_1 = ... = np_k$ были не менее 5-6.

Юнит 9. Критерий χ^2 Пирсона для проверки параметрической гипотезы

Критерий χ^2 часто применяют для проверки гипотезы о виде распределения, т.е. о принадлежности распределения выборки некоторому параметрическому семейству.

Имеется выборка $\vec{X} = \left(X_1, ..., X_n\right)$ из неизвестного распределения F.

Проверяется сложная гипотеза $H_1 = \{F \in F(\theta)\}$,где $\theta \in \Theta \subseteq R^l$ — неизвестный параметр (скалярный или векторный), l — его размерность.

Пусть $\mathbb{R}$ разбито на k > l интервалов группировки $A_1, ..., A_k$, и $\nu_i = \sum_{k=1}^{n} I\left\{x_k \in A_i\right\}$.

$$\text{Но вероятность} \quad p_i = P\left\{\frac{\xi \in A_i}{H_1}\right\} = p_i(\theta)$$

теперь зависит от неизвестного параметра θ. Функция отклонения $\rho(\vec{X})$ также зависит от неизвестного параметра θ, и использовать ее в критерии Пирсона нельзя — мы не можем вычислить ее значение: $\rho(\vec{X}, \theta) = \sum_{i=1}^{k} \frac{\left(\nu_i - np_i(\theta)\right)^2}{np_i(\theta)}$

Пусть $\hat{\theta} = \hat{\theta}(\vec{X})$ — значение параметра θ, доставляющее минимум функции $\rho(\vec{X}, \theta)$ при данной выборке $\vec{X}$. Подставив вместо истинных вероятностей p_i их оценки $p_i\left(\hat{\theta}\right)$, получим функцию отклонения

$$\rho(\vec{X}, \hat{\theta}) = \sum_{i=1}^{k} \frac{\left(\nu_i - np_i\left(\hat{\theta}\right)\right)^2}{np_i\left(\hat{\theta}\right)}$$

Теорема Фишера *. (Без доказательства)*

Если верна гипотеза $H_1 = \{F \in F(\theta)\}$, и $dim(\theta) = l$ — размерность параметра (вектора) θ,

$$\text{то при фиксированном } k \text{ и при } n \to \infty \quad \rho(\vec{X}, \hat{\theta}) = \sum_{i=1}^{k} \frac{\left(\nu_i - np_i\left(\hat{\theta}\right)\right)^2}{np_i\left(\hat{\theta}\right)} \Rightarrow H_{k-l-1}$$

, где H_{k-l-1} есть χ^2-распределение с $k - 1 - l$ степенями свободы.

Построим критерий χ^2.

Пусть случайная величина $\chi^2 \in H_{k-1-l}$. По заданному ε найдем C такое, что $\varepsilon = P\left\{\chi^2 > C\right\}$ и критерий согласия χ^2 выглядит как все критерии согласия:

$$\delta(\vec{X}) \ = \{H_1 \ \text{если}\ \rho\left(\vec{X}, \hat{\theta}\right) < \ C \ \ H_2 \ \text{если}\ \rho\left(\vec{X}, \hat{\theta}\right) \geq C$$

Замечание. Оценку $\hat{\theta}$, минимизирующую функцию $\rho(\vec{X}, \theta)$, нельзя заменить на оценку максимального правдоподобия для θ, построенную по выборке $\vec{X} = \left(X_1,..., X_n\right)$. При такой замене Теорема Фишера не верна!

Пусть $\tilde{\theta}$ оценка, построенная по выборке (ММ илиММП). $\rho\left(\vec{X}, \tilde{\theta}\right) \geq \ \rho\left(\vec{X}, \hat{\theta}\right)$

Тогда если $\rho\left(\vec{X}, \tilde{\theta}\right) < \ C$, то и $\rho\left(\vec{X}, \hat{\theta}\right) < \ C$ и гипотеза H_1 принимается.

Если $\rho\left(\vec{X}, \tilde{\theta}\right) \geq C$, стоит вычислить $\rho\left(\vec{X}, \hat{\theta}\right)$

Пример

По данной выборке принять или отвергнуть гипотезу о нормальном распределении выборки с уровнем значимости $\alpha_1 \ = \ 0,05$

$$H_1 = \left\{\xi \ \in \ N(a, \sigma^2\right\}$$

0.4	-1.91	-1.6	5.47	0.72	2.24	-0.55	-0.43	0.93	-1.13
0.09	0.83	3.46	6.64	-1.3	2.96	2.89	1.04	-1.44	-3.45
-1.92	3.14	5.38	0.41	0.25	1.04	-1.42	-0.19	-5.27	1.43
-1.66	1.67	-0.24	3.64	-1.89	1.7	1.44	4.23	1.57	2.06
2.1	5.56	2.84	-0.54	2.3	1.72	0.11	-0.42	-1.18	2.87

Проверяется сложная гипотеза $H_1 \ = \{F \in N(\theta)\}$,где $\theta \ = \left\{a, \sigma^2\right\} \ \in \ \ R^2$ — неизвестный параметр размерности 2.

$X_{min} =- \ 5,27 \ X_{max}$ =6.64

$$\hat{a} = \overline{x} = \ 0.932 \qquad \overline{x^2} \ = \ 6.49 \qquad \hat{\sigma} = \sqrt{S_0^2} \ = \ 2,4$$

$$\Delta a \ = \frac{X_{max} - X_{min}}{k} = \frac{6,64+5,27}{6} \ = \ 1,985 \approx 2$$

$$\nu_i = \sum_{k=1}^{n} I\left\{x_k \in [a_i, a_{i+1}[\right\}$$

$$\rho(\vec{X}, \hat{\theta}) = \sum_{i=1}^{k} \frac{\left(\nu_i - np_i(\hat{\theta})\right)^2}{np_i(\hat{\theta})} \text{ где } p_i = P\left\{\frac{\xi \in A_i}{\xi \in N(\hat{\theta})}\right\} = \Phi\left(\frac{a_{i+1} - \bar{x}}{\sqrt{S_0^2}}\right) - \Phi\left(\frac{a_i - \bar{x}}{\sqrt{S_0^2}}\right) = \Delta\Phi$$

a_i	a_{i+1}	ν_i	$\frac{a_i - \bar{x}}{\sqrt{S_0^2}}$	$\Phi\left(\frac{a_i - \bar{x}}{\sqrt{S_0^2}}\right)$	$p_i = \Delta\Phi$	np_i	$\frac{\left(\nu_i - np_i\right)^2}{np_i}$
$-\infty$	-3,2	2	$-\infty$	-0,5	0,0427	2,135	0,0085
-3,2	-1,2	8	-1,72	-0,4573	0,1440	7,2	0,0889
-1,2	0,8	14	-0,89	-0,3133	0,2894	14,47	0,0153
0,8	2,8	14	-0,06	-0,0239	0,3062	15,31	0,1121
2,8	4,8	8	0,78	0,2823	0,1640	8,2	0,0049
4,8	∞	4	1,61	0,4463	0,0537	2,685	0,6440
∞		50	∞	0,5	$\Sigma = 1$	$\Sigma = 50$	$\Sigma = 0,8937$

$$\rho_{\text{набл.}}\left(\vec{X}, \hat{\theta}\right) = \sum_{i=1}^{k} \frac{\left(\nu_i - np_i(\hat{\theta})\right)^2}{np_i(\hat{\theta})} = 0,8937$$

Пусть случайная величина $\chi^2 \in H_{k-1-l} = H_{6-1-2} = H_3$. По заданному $\alpha_1 = 0,05$ найдем C такое, что $\alpha_1 = 0,05 = P\left\{\chi^2 > C\right\}$ C – квантиль распределения χ^2 с тремя степенями свободы уровня 0,05. C=7,8

$$\delta(\vec{X}) = \{H_1 \text{ если } \rho\left(\vec{X}, \hat{\theta}\right) < C \;\; H_2 \text{ если } \rho\left(\vec{X}, \hat{\theta}\right) \geq C$$

$$\rho_{\text{набл.}}\left(\vec{X}, \hat{\theta}\right) = 0,8937 < C = 7,8$$

$$\delta(\vec{X}) = \{H_1 \text{ если } \rho\left(\vec{X}\right) < C = 7,8 \;\; H_2 \text{ если } \rho\left(\vec{X}\right) \geq C = 7,8$$

Вывод: Выборка не противоречит гипотезе $H_1 = \left\{\xi \in N(a, \sigma^2\right\}$ с уровнем значимости $\alpha_1 = 0,05$.

Юнит 10. Однородность нормальных выборок

10.1 Совпадение дисперсий двух нормальных выборок

Критерий Фишера используют в качестве первого шага в задаче проверки однородности двух независимых нормальных выборок.

Есть две независимые выборки из нормальных распределений:

$\vec{X} = \left(X_1, ..., X_n\right) \qquad X_i \in N\left(a_1, \sigma_1^2\right)$ и $\vec{Y} = (Y_1, ..., Y_m), \quad Y_i \in N\left(a_2, \sigma_2^2\right)$

средние которых, вообще говоря, неизвестны. Критерий Фишера предназначен для проверки гипотезы $H_1 = \{\sigma_1^2 = \sigma_2^2\}$. Используем несмещенные выборочные дисперсии:

$$S_0^2\left(\vec{X}\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{x}\right)^2$$

$$S_0^2\left(\vec{Y}\right) = \frac{1}{m-1} \sum_{i=1}^{m} \left(Y_i - \overline{y}\right)^2$$

и зададим функцию отклонения $\rho(\vec{X}, \vec{Y})$ как их отношение

$$\rho\left(\vec{X}, \vec{Y}\right) = \frac{S_0^2\left(\vec{X}\right)}{S_0^2\left(\vec{Y}\right)}$$

Удобно, если $\rho > 1$. С этой целью выборкой $\vec{X}$ называют ту из двух выборок, несмещённая дисперсия которой больше. Поэтому предположим, что $S_0^2\left(\vec{X}\right) > S_0^2\left(\vec{Y}\right)$

Теорема. Если гипотеза $H_1 = \{\sigma_1^2 = \sigma_2^2\}$ верна, то случайная величина $\rho(\vec{X}, \vec{Y})$ имеет распределение Фишера $F_{n-1,m-1}$ с $n - 1,\ m - 1$ степенями свободы.

Доказательство.

По лемме Фишера, независимые случайные величины

$$\chi_{n-1}^2 = \frac{(n-1)S_0^2\left(\vec{X}\right)}{\sigma_1^2} \in H_{n-1}$$

$$\chi_{m-1}^2 = \frac{(m-1)S_0^2\left(\vec{Y}\right)}{\sigma_2^2} \in H_{m-1}$$

При $\sigma_1^2 = \sigma_2^2$ отношение

$$\frac{\frac{\chi_{n-1}^2}{(n-1)}}{\frac{\chi_{m-1}^2}{(m-1)}} = \frac{S_0^2\left(\vec{X}\right)}{\sigma_1^2} \bullet \frac{\sigma_2^2}{S_0^2\left(\vec{Y}\right)} = \rho\left(\vec{X}, \vec{Y}\right) \in F_{n-1,m-1} \qquad \blacksquare$$

Если $\sigma_1^2 > \sigma_2^2$, при $n, m \to \infty$.

$$S_0^2(\vec{X}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{x}\right)^2 p\to M\left(S_0^2(\vec{X})\right) = \sigma_1^2 \text{ (ЗБЧ)}$$

$$S_0^2(\vec{Y}) = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{y}\right)^2 p\to M\left(S_0^2(\vec{Y})\right) = \sigma_2^2 \text{ (ЗБЧ)}$$

$$\rho\left(\vec{X}, \vec{Y}\right) = \frac{S_0^2(\vec{X})}{S_0^2(\vec{Y})}\ p\to \frac{\sigma_1^2}{\sigma_2^2} > 1$$

Построим критерий Фишера и убедимся, что он состоятелен.

Возьмём квантиль $f_{1-\varepsilon}$ распределения Фишера $F_{n-1,m-1}$.

Критерием Фишера называют критерий
$\delta(X, Y) = \{H_1,\ \text{если}\ \rho\left(\vec{X}, \vec{Y}\right) < f_{1-\varepsilon}\ H_2, \text{если}\ \rho\left(\vec{X}, \vec{Y}\right) \geq f_{1-\varepsilon}$,

Доказательство состоятельности критерия Фишера.

Покажем, что последовательность квантилей $f_\delta = f_\delta(n, m)$ любого уровня $0 < \delta < 1$ распределения $F_{n,m}$. сходится к *1* при $n, m \to \infty$. Возьмем величину $\psi_{n,m} \in F_{n,m}$

По определению, $P(\psi_{n,m} < f_\delta) = \delta$, $P(\psi_{n,m} > f_\delta) = 1 - \delta$ при всех n, m.

По свойству распределения Фишера, $\psi_{n,m}\ p\to 1$.

Поэтому для любого $\varepsilon > 0$ обе вероятности $P(\psi_{n,m} < 1 - \varepsilon)$ и $P(\psi_{n,m} > 1 + \varepsilon)$ стремятся к нулю при $n, m \to \infty$, становясь рано или поздно меньше как δ, так и $1 - \delta$. Следовательно, при достаточно больших n, m выполнено $1 - \varepsilon < f_\delta < 1 + \varepsilon$.

Пусть гипотеза $H_1 = \{\sigma_1^2 = \sigma_2^2\}$ не верна. Достаточно в качестве альтернативы рассмотреть $H_2 = \left\{\sigma_1^2 > \sigma_2^2\right\}$

$$\rho\left(\vec{X}, \vec{Y}\right) = \frac{S_0^2(\vec{X})}{S_0^2(\vec{Y})}\ p\to \frac{\sigma_1^2}{\sigma_2^2} > 1$$

$$\alpha_2(\delta) = P\left\{\frac{\rho\left(\vec{X},\vec{Y}\right) < f_{1-\varepsilon}}{H_2}\right\} = P\left\{\frac{\rho\left(\vec{X},\vec{Y}\right) - f_{1-\varepsilon} < 0}{H_2}\right\} \to P\left\{\frac{\sigma_1^2}{\sigma_2^2} - 1 < 0\right\} = 0. \quad \blacksquare$$

Сформулировать критерий Фишера в случае, когда средние известны. Какой статистикой стоит воспользоваться теперь?

Пример: По двум независимым выборкам $\vec{X}$ и $\vec{Y}$, объёмы которых $n = 11, \quad m = 14$ извлечённым из нормальных генеральных совокупностей, найдены исправленные выборочные дисперсии $S_0^2(\vec{X}) = 0,76$ $S_0^2(\vec{Y}) = 0,38$ При уровне значимости $\varepsilon = 0,05$ проверить гипотезу $H_1 = \{\sigma_x^2 = \sigma_y^2\}$ при альтернативе $H_2 = \{\sigma_x^2 > \sigma_y^2\}$

$$\rho(\vec{X}, \vec{Y}) = \frac{S_0^2(\vec{X})}{S_0^2(\vec{Y})} = \frac{0,76}{0,38} = 2 = \rho_{\text{набл.}}$$

$$\delta(X, Y) = \{H_1, \text{ если } \rho(\vec{X}, \vec{Y}) < f_{0,05;11-1;14-1} = 2,67\ H_2, \text{ если } \rho(\vec{X}, \vec{Y}) \geq f_{0,05;11-1;14-1} = 2,67$$

Вывод: поскольку $\rho_{\text{набл.}} < 2,76$, нет оснований отвергнуть гипотезу H_1. Считаем, что выборочные дисперсии отличаются незначимо.

10.2 Равенство средних двух нормальных выборок

Особенно часто возникает необходимость проверить равенство средних двух нормальных совокупностей: например, в медицине или биологии для выяснения наличия или отсутствия действия препарата. Эта задача решается с помощью критерия Стьюдента, но только в случае, когда неизвестные дисперсии равны. Для проверки же равенства дисперсий пользуются сначала критерием Фишера.

Критерий Стьюдента.

Есть две независимые выборки из нормальных распределений:

$$\vec{X} = (X_1, ..., X_n) \qquad X_i \in N(a_1, \sigma^2) \text{ и } \vec{Y} = (Y_1, ..., Y_m), \quad Y_i \in N(a_2, \sigma^2)$$

с неизвестными средними и одной и той же неизвестной дисперсией σ^2. Проверяется сложная гипотеза $H_1 = \{a_1 = a_2\}$.. Построим критерий Стьюдента точного размера ε.

Теорема. Случайная величина

$$t_{n+m-2} = \sqrt{\frac{nm}{n+m}} \bullet \frac{(\bar{x}-a_1)-(\bar{y}-a_2)}{\sqrt{\frac{(n-1)S_0^2(\vec{X})+(m-1)S_0^2(\vec{Y})}{n+m-2}}} \in T_{n+m-2}$$

имеет распределение Стьюдента T_{n+m-2}

Доказательство.

Поскольку

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad M(\bar{x}) = a_1 \qquad D(\bar{x}) = \frac{\sigma^2}{n},$$

случайная величина $\left(\overline{x} - a_1\right) \in N\left(0, \frac{\sigma^2}{n}\right)$, а случайная величина $\left(\overline{y} - a_2\right) \in N\left(0, \frac{\sigma^2}{m}\right)$. Тогда их разность распределена тоже нормально с нулевым средним и дисперсией

$$D\left(\left(\overline{x} - a_1\right) - \left(\overline{y} - a_2\right)\right) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \frac{\sigma^2(n+m)}{nm}$$

Нормируем эту разность:

$$\xi_0 = \sqrt{\frac{nm}{(n+m)}}\frac{\left(\overline{x}-a_1\right)-\left(\overline{y}-a_2\right)}{\sigma} \in N(0, 1)$$

Из леммы Фишера следует, что независимые случайные величины

$$\chi^2_{n-1} = \frac{(n-1)S_0^2\left(\vec{X}\right)}{\sigma^2} \in H_{n-1}$$

$$\chi^2_{m-1} = \frac{(m-1)S_0^2\left(\vec{Y}\right)}{\sigma^2} \in H_{m-1}$$

а их сумма

$$\chi^2_{n-1} + \chi^2_{m-1} = \frac{(n-1)S_0^2\left(\vec{X}\right)}{\sigma^2} + \frac{(m-1)S_0^2\left(\vec{Y}\right)}{\sigma^2} \in H_{n-1+m-1}$$

и не зависит от $\overline{x}$ и от $\overline{y}$.

По определению $\frac{\xi_0}{\sqrt{\frac{\chi^2_n}{n}}} \in T_n$ имеет распределение Стьюдента

$$\frac{\xi_0}{\sqrt{\frac{\chi^2_{n-1}+\chi^2_{m-1}}{n+m-2}}} \in T_{n+m-2}$$

$$\frac{\xi_0}{\sqrt{\frac{\chi^2_{n-1}+\chi^2_{m-1}}{n+m-2}}} = \sqrt{\frac{nm}{(n+m)}}\frac{\left(\overline{x}-a_1\right)-\left(\overline{y}-a_2\right)}{\sigma\bullet\sqrt{\frac{(n-1)S_0^2\left(\vec{X}\right)+(m-1)S_0^2\left(\vec{Y}\right)}{\sigma^2(n+m-2)}}} = \sqrt{\frac{nm}{n+m}} \bullet \frac{\left(\overline{x}-a_1\right)-\left(\overline{y}-a_2\right)}{\sqrt{\frac{(n-1)S_0^2\left(\vec{X}\right)+(m-1)S_0^2\left(\vec{Y}\right)}{n+m-2}}} \in T_{n+m-2} \qquad \blacksquare$$

Введём функцию

$$\rho\left(\vec{X}, \vec{Y}\right) = \sqrt{\frac{nm}{n+m}} \bullet \frac{(\overline{x}-\overline{y})}{\sqrt{\frac{(n-1)S_0^2\left(\vec{X}\right)+(m-1)S_0^2\left(\vec{Y}\right)}{n+m-2}}}$$

1):если H_1 верна, т.е. если $a_1 = a_2$, то величина $\rho\left(\vec{X}, \vec{Y}\right) \in T_{n+m-2}$ имеет распределение Стьюдента.

2): если $a_1 \neq a_2$, величина $|\rho|$ неограниченно возрастает по вероятности с ростом n и m.

$$(\overline{x} - \overline{y})\ p\!\to \left(a_1 - a_2\right) \neq 0 \qquad \text{(ЗБЧ)}$$

$$\frac{(n-1)S_0^2(\vec{X})+(m-1)S_0^2(\vec{Y})}{n+m-2}\ p\to \sigma^2 \qquad (\text{ЗБЧ})$$

$$\left|\rho\left(\vec{X},\vec{Y}\right)\right| = \sqrt{\frac{nm}{n+m}} \bullet \frac{|\bar{x}-\bar{y}|}{\sqrt{\frac{(n-1)S_0^2(\vec{X})+(m-1)S_0^2(\vec{Y})}{n+m-2}}}\ p\to \sqrt{\frac{nm}{n+m}}\frac{\left|a_1-a_2\right|}{\sqrt{\sigma^2}}\to\infty$$

Поэтому остаётся по ε найти $C\ =\ \tau_{1-\varepsilon/2}$ —квантиль распределенияT_{n+m-2}.

Критерий Стьюдента выглядит как все критерии согласия:

$$\delta(X,Y)\ =\{H_1,\ \text{если}\ \left|\rho\left(\vec{X},\vec{Y}\right)\right|\ <\ \tau_{1-\varepsilon/2}\ H_2,\text{если}\ \left|\rho\left(\vec{X},\vec{Y}\right)\right|\geq\tau_{1-\varepsilon/2}\ ,$$

Пример: По двум независимым выборкам $\vec{X}$ и $\vec{Y}$, объёмы которых $n\ =\ 12,\quad m\ =\ 18$ извлечённым из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x}\ =\ 31,2\ \ \bar{y}\ =\ 29,2$ и исправленные выборочные дисперсии $S_0^2(\vec{X})=\ 0,84$ $S_0^2(\vec{Y})=\ 0,40$ При уровне значимости $\varepsilon\ =\ 0,05$ проверить гипотезу $H_1\ =\ \{a_x\ =\ a_y\}$ при альтернативе $H_2\ =\ \{a_x\neq a_y\}$

Сначала проверим гипотезу о равенстве дисперсий с помощью критерия Фишера.

$$\rho_1\left(\vec{X},\vec{Y}\right)=\frac{S_0^2(\vec{X})}{S_0^2(\vec{Y})}=\frac{0,84}{0,40}\ =\ 2,1\ =\rho_{1\text{набл.}}$$

$$\delta(X,Y)\ =\{H_1,\ \text{если}\ \rho_1\left(\vec{X},\vec{Y}\right)\ <\ f_{0,05;12-1;18-1}\ =\ 2,41\ H_2,\text{если}\ \rho_1\left(\vec{X},\vec{Y}\right)\geq f_{0,05;11-1;14-1}\ =\ 2,41$$

Поскольку $\rho_{1\text{набл.}}\ <\ 2,41$,можем считать, что дисперсии нормальных генеральных совокупностей равны.

Теперь можем воспользоваться критерием Стьюдента для проверки равенства математических ожиданий.

$$\rho\left(\vec{X},\vec{Y}\right)=\sqrt{\frac{nm}{n+m}}\bullet\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{(n-1)S_0^2(\vec{X})+(m-1)S_0^2(\vec{Y})}{n+m-2}}}=\sqrt{\frac{12\cdot18}{12+18}}\bullet\frac{(31,2-29,2)}{\sqrt{\frac{11\cdot0,84+17\cdot0,40}{12+18-2}}}\ =\ 7,09$$

$$\delta(X,Y)=\{H_1,\ \text{если}\ \left|\rho\left(\vec{X},\vec{Y}\right)\right|\ <\ \tau_{0,025,18+12-2}=\ 2,05\ H_2,\text{если}\ \left|\rho\left(\vec{X},\vec{Y}\right)\right|\geq\tau_{0,025,18+12-2}=\ 2,05$$

Вывод: поскольку $\rho_{\text{набл.}}\ =\ 7,09\ >\ 2,05$, отвергаем гипотезу H_1 о равенстве математических ожиданий. Считаем, что выборочные средние различаются значимо.

Юнит 11. Критерии, основанные на доверительных интервалах

Имеется выборка $\vec{X} = (X_1, ..., X_n)$ из семейства распределений F_θ, где $\theta \in \Theta$.

Проверяется простая гипотеза $H_1 = \{\theta = \theta_0\}$ против сложной альтернативы $H_2 = \{\theta \neq \theta_0\}$. Пусть имеется доверительный интервал

$$P\{\underline{\theta} < \theta < \overline{\theta}\} = 1 - \varepsilon$$

Взяв произвольное θ′, для выборки из распределения $F_{\theta'}$ имеем $P\{\underline{\theta} < \theta' < \overline{\theta}\} = 1 - \varepsilon$

Тогда критерий

$$\delta(X, Y) = \{H_1, \text{ если } \underline{\theta} < \theta_0 < \overline{\theta}\ H_2, \text{если } \theta_0 \notin (\underline{\theta}, \overline{\theta}),$$

имеет точный размер ε:

$$\alpha_1(\delta) = P\left\{\frac{\delta = H_2}{H_1}\right\} = P\left\{\frac{\theta_0 \notin (\underline{\theta}, \overline{\theta})}{H_1}\right\} = 1 - P\left\{\frac{\underline{\theta} < \theta_0 < \overline{\theta}}{H_1}\right\} = \varepsilon.$$

Если доверительный интервал строится с помощью функции $G(\vec{X}; \theta)$, то эта же функция годится и в качестве «функции отклонения» $\rho(\vec{X})$ для построения критерия согласия.

Проверка гипотезы о среднем нормального распределения с известной дисперсией.

Имеется выборка $\vec{X} = (X_1, ..., X_n)$ из нормального распределения $X_i \in N(a, \sigma^2)$ с известной дисперсией σ^2. Проверяется простая гипотеза $H_1 = \{a = a_0\}$ против сложной альтернативы $H_2 = \{a \neq a_0\}$. Построим критерий размера ε с помощью функции

$$\rho(\vec{X}) = \frac{\overline{x} - a_0}{\frac{\sigma}{\sqrt{n}}}$$

1): если H_1 верна, то $\rho(\vec{X}) \in N(0, 1)$. По ε выберем $C = u_{\varepsilon/2}$ —квантиль стандартного нормального распределения уровня $\left(1 - \frac{\varepsilon}{2}\right)$

$$\Phi(u_{\varepsilon/2}) = \frac{1-\varepsilon}{2}$$

Критерий выглядит как все критерии согласия:

$$\delta(\vec{X}) = \{H_1, \text{ если } \left|\rho(\vec{X})\right| < u_{\varepsilon/2}\ H_2, \text{если } \left|\rho(\vec{X})\right| \geq u_{\varepsilon/2}$$

2): если $a \neq a_0$, то

$$\left|\rho\left(\vec{X}\right)\right| = \frac{\left|\bar{x} - a_0\right|}{\frac{\sigma}{\sqrt{n}}} \; p \rightarrow \infty$$

Критерий имеет точный размер ε и является состоятельным.

Проверка гипотезы о среднем нормального распределения с неизвестной дисперсией.

Имеется выборка $\vec{X} = \left(X_1, ..., X_n\right)$ из нормального распределения $X_i \in N\left(a, \sigma^2\right)$ с неизвестной дисперсией σ^2. Проверяется простая гипотеза $H_1 = \{a = a_0\}$ против сложной альтернативы $H_2 = \{a \neq a_0\}$.

Критерий, который мы построим, называют одновыборочным критерием Стьюдента. Введём функцию отклонения

$$\rho\left(\vec{X}\right) = \frac{\bar{x} - a_0}{\sqrt{\frac{S_0^2}{n}}}$$

По следствию леммы Фишера выполнено условие:

1) если $a = a_0$, то $\rho\left(\vec{X}\right)$ имеет распределение Стьюдента T_{n-1}.

2): если $a \neq a_0$, то

$$\left|\rho\left(\vec{X}\right)\right| = \frac{\left|\bar{x} - a_0\right|}{\sqrt{\frac{S_0^2}{n}}} \; p \rightarrow \infty$$

Критерий строится так же, но в качестве C следует брать квантиль распределения Стьюдента, а не стандартного нормального распределения. Поэтому остаётся по ε найти $C = \tau_{1-\varepsilon/2}$ —квантиль распределения T_{n-1}.

Критерий Стьюдента выглядит как все критерии согласия:

$$\delta(X, Y) = \{H_1, \text{ если } \left|\rho\left(\vec{X}, \vec{Y}\right)\right| < \tau_{1-\varepsilon/2} \; H_2, \text{если } \left|\rho\left(\vec{X}, \vec{Y}\right)\right| \geq \tau_{1-\varepsilon/2} \, ,$$