Элементы прикладной математической статистики. 14.02.2011 год

Литература: Чистяков. Курс теории вероятностей (гл. 9)
Гнеденко. Курс теории вероятностей.
Ватутин, Ивченко, Медведев, Чистяков. Теор. вер. и мат. стат. в задачах.

$F_\xi(x) = P\{\xi < x\}$; зачастую нам ничего не известно о распределении сл. вел.

$$F_\xi(x) = \Phi(x; a; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{(x-a)^2}{2\sigma^2}} dx$$

Одна из задач мат. статистики — восстановить закон распределения сл. вел. по рез-там эксперимента.

Случайной выборкой объёма n из распределения $F_\xi(x)$ назовём n-мерный сл. вектор $\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$, компоненты которого одинаково распределены по закону

$$F_{\xi_i}(x) = F_\xi(x) = F(x), \quad i = 1, \dots, n.$$

Чаще всего предполагается что сл. вел. независимы. Тогда мы называем выборку независимой.

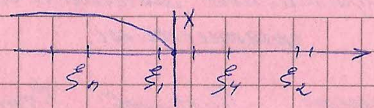
$$(x_1, x_2, \dots, x_n)$$

$P\{x < x_1\}$ — особенность обозначений

Одной из основных характеристик выборки является эмпирическая ф.р.

$\hat{F}_n(x) = \frac{\mu_n(x)}{n}$, где $x \in R$, $\mu_n(x)$ — число эл-тов выборки, строго меньших x.

$\xi_1, \xi_2, \ldots, \xi_n$



n экземпляров сл. вел. ξ.

$\mu_n(x)$ — число наступлений события $\{\xi<x\}$ в n испытаниях.

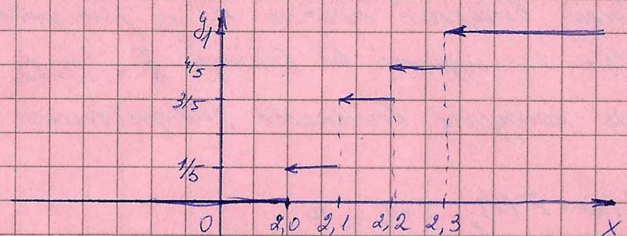
Событие $\{\xi < x\}$ происходит с вер-тью $F_\xi(x) = F(x) = p$, x — число успехов в схеме Бернулли с параметрами $(n, F(x))$.

$M\mu_n(x) = n\cdot F(x)$; $\hat{F}_n(x)$ — сл. величина.

$M\hat{F}_n(x) = F(x)$

Пример. Пусть измеряется некоторая постоянная величина (взвешивание вещ-ва в некотором объёме).

2,1 2,1 2,0 2,2 2,3



$$\eta = \begin{pmatrix} 2{,}0 & 2{,}1 & 2{,}2 & 2{,}3 \\ \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

Вариационный ряд

$x_1, x_2, \ldots, x_n$

$\xi_1, \xi_2, \ldots, \xi_n$ $\quad F_\xi(x) = F(x)$

$\xi_{(1)}, \xi_{(2)}, \ldots, \xi_{(n)}$ расположим в порядке неубывания

вариационный ряд

$\xi_{(1)} \le \xi_{(2)} \le \ldots \le \xi_{(n)}$

$\xi_{(1)} = \min\limits_{1\le i\le n} \xi_i$; $\xi_{(n)} = \max\limits_{1\le i\le n} \xi_i$

$\xi_{(k)}$ — k-тая порядковая статистика

$\xi_{(n)} - \xi_{(1)}$ — размах выборки

Медиана выборки: $\xi_{(m)}$, если $n = 2m+1$

$\xi_{(m)} = \frac{\xi_{(n)} + \xi_{(n)}}{2}$, если n чётное.

Вернёмся к примеру.

Размах: $2{,}3 - 2{,}0 = 0{,}3$

Медиана: 2,1

Пусть

$\xi_1, \xi_2, \ldots, \xi_n$

Всякая измеримая функция от выборки называется статистикой, или оценкой.

$\eta_n = \eta_n(\xi_1, \ldots, \xi_n)$

$F_\xi(x) = F(x) = F(x, \theta)$

$\eta_n = \eta_n(\xi_1, \ldots, \xi_n)$ называется несмещённой оценкой парам. θ, если $M\eta_n = \theta$.

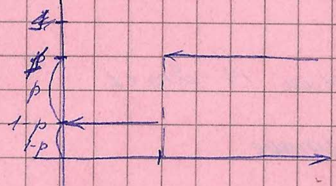
Оценка $\eta_n = \eta_n(\xi_1, \ldots, \xi_n)$ называется состоятельной оценкой параметра θ, если

при $n \to \infty$ $\quad \eta_n \xrightarrow[n\to\infty]{P} \theta$

Пример. Имеется послед-ть независимых испытаний с параметрами (n, p).

$$\mu_n = \xi_1 + \xi_2 + \dots + \xi_n, \qquad \xi_i = \begin{cases} 1, & p \text{ (успех)} \\ 0, & 1-p \text{ (неуспех)} \end{cases}$$

$\xi_1, \dots, \xi_n$ — выборка



$$M\xi_i = p, \quad i = 1, 2, \dots$$

$$D\xi_i = p(1-p) = pq$$

$$\eta_n = \frac{\mu_n}{n} = \frac{\xi_1 + \dots + \xi_n}{n}$$

$$M\eta_n = \frac{1}{n} M\mu_n = \frac{np}{n} = p$$

$$\eta_n = \frac{\mu_n}{n} \xrightarrow[n\to\infty]{P} p$$ по т. Бернулли

$$\zeta_1, \zeta_2, \zeta_3, \dots \zeta_n$$

$$\varepsilon > 0 \quad P\{|\zeta_n - c| \ge \varepsilon\} \xrightarrow[n\to\infty]{} 0, \quad c = p$$

$$\varepsilon > 0 \quad \underbrace{P\left\{\left|\frac{\mu_n}{n} - p\right| \ge \varepsilon\right\}}_{n\to\infty \ \to 0} \le \frac{D\left(\frac{\mu_n}{n}\right)}{\varepsilon^2} = \frac{D\mu_n}{n^2\varepsilon^2} = \frac{n \cdot pq}{n^2\varepsilon^2} = \frac{pq}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0$$

$$\hat{F}_n(x) = \frac{\mu_n(x)}{n} \qquad \{\xi < x\} \qquad F(x) = P\{\xi < x\}$$

$\mu_n(x)$ — бином. распр-ие с параметрами $(n, F(x))$

$$M\hat{F}_n(x) = F(x)$$

$$D\mu_n(x) = \underbrace{F(x)}_{p}\underbrace{(1 - F(x))}_{1-p} \cdot n$$

$$D\hat{F}_n(x) = D\frac{\mu_n(x)}{n} = \frac{F(x)(1 - F(x))}{n} \xrightarrow[n\to\infty]{} 0$$

$$\hat{F}_n(x) \xrightarrow[n\to\infty]{P} F(x)$$

* Гистограмма

Выборки из нормального распределения.

Очевидно, что распределения оценок

$\eta_n = \eta_n(\xi_1, \dots, \xi_n)$ определяются распределением исходной сл. вел., т.е. её ф.р.:

$$F_\xi(x) = P(\xi < x).$$

Отыскание точных распределений оценок зачастую является очень сложной задачей. В таких случаях проще бывает получить предельное распределение этих оценок при $n \to \infty$, и уже по ним приближать точные. Исключением являются выборки из нормального распределения.

$\xi_1, \xi_2, \dots, \xi_n \sim N(a, \sigma^2)$, тогда можно найти точное распределение оценок этих параметров.

Лекция № 3.

14.03.2011г.

$\xi: \ \xi_1, \xi_2, \dots, \xi_n \qquad F(x) = F_\xi(x)$

$F(x) = F(x, \theta) \qquad \Phi(x; a, \sigma) \qquad M\xi = a, \ D\xi = \sigma^2$

$\hat{F}_n(x) = \dfrac{\text{число } \xi_i < x}{n}, \qquad \{\xi_i < x\}$

$F(x) \qquad \hat{F}_n(x) \qquad \begin{pmatrix} \xi_1 = x_1 & \xi_2 = x_2 & \dots & \xi_n = x_n \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix}$

$k = 1$

$M\xi = a \qquad \alpha_k = M\xi^k \qquad a_k = \frac{1}{n}\sum_{i=1}^{n} \xi_i^k, \quad k = 1 \ \ a = \bar{\xi}$

$k = 2$

$\mu_2 = \sigma^2 \qquad \mu_k = M(\xi - M\xi)^k \qquad m_k = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^k, \quad m_2 = S^2$

$\eta_n = \eta(\xi_1, \dots, \xi_n)$

$M\eta_n = \theta \qquad \eta_n \xrightarrow[n \to \infty]{P} \theta$

$D\eta_n \to 0 \quad n \to \infty$

$P\{|\eta_n - \theta| \geq \varepsilon\} \leq \dfrac{D\eta_n}{\varepsilon^2}$

Пользуются асимптотическими оценками

Не для любой оценки можно найти распределение.

$\Phi(x; a, \sigma)$: выборка из норм. распр.

норм. распред.

$\chi_n^2 = \xi_1^2 + \dots + \xi_n^2 \qquad \xi_i \sim N(0;1)$

независимы

$\tau_n = \dfrac{\xi_0}{\sqrt{\frac{1}{n}\chi_n^2}} \qquad \xi_1, \dots, \xi_n; \ \xi_0$

$F_{n_1, n_2} = \dfrac{\frac{1}{n_1}\chi_{n_1}^2}{\frac{1}{n_2}\chi_{n_2}^2}$

Теорема Фишера.

$\xi_1, \xi_2, \dots, \xi_n \sim N(0;1)$

$$\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i, \qquad m_2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$$

1) $\bar{\xi}$ и S^2 — независимы

2) Сл. вел. $\frac{n m^2}{\sigma^2} = \frac{nS^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{\xi_i - \bar{\xi}}{\sigma}\right)^2$ имеет распред. χ^2_{n-1}

Линейные ф-ии от норм. распр. сл. величин снова дают норм. распределение.

$$\sum_{i=1}^{n}\left(\frac{\xi_i - \bar{\xi}}{\sigma}\right)^2 = \eta_1^2 + \dots + \eta_{n-1}^2$$

Теорема 1.

$\xi_1, \xi_2, \dots, \xi_1 \sim N(a, \sigma)$

$$\tau_{n-1} = \frac{\bar{\xi} - a}{s/\sqrt{n}} = \frac{\bar{\xi} - a}{S}\sqrt{n-1} = \frac{\bar{\xi} - a}{\sqrt{m_2}}\sqrt{n-1} = \frac{\eta_0}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}}$$

$$\tau_{n-1} = \frac{\frac{\bar{\xi} - a}{\sigma/\sqrt{n}}}{\sqrt{\frac{nS^2}{\sigma^2(n-1)}}} \qquad \frac{\bar{\xi} - a}{\sigma/\sqrt{n}}$$

$$D\bar{\xi} = \frac{\sigma^2}{n} \qquad \frac{\sigma}{\sqrt{n}} = \sqrt{D\bar{\xi}}$$

$$D\left(\frac{\bar{\xi} - a}{\sigma/\sqrt{n}}\right) = \frac{1}{\sigma^2/n} D\bar{\xi} = 1$$

Теорема 2.

$\xi_1, \xi_2, \dots, \xi_{n_1}$ $\quad$ $n_1 + n_2$ сл. вел. нез. н.р. (a, σ)

$\xi_1, \xi_2, \dots, \xi_{n_2}$

$$m_2 = S^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(\xi_i - \bar{\xi})^2$$

$$m_2' = S^{2'} = \frac{1}{n_2}\sum_{i=1}^{n_2}(\xi_i' - \bar{\xi}')^2$$

$$\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i \qquad M\bar{\xi} = a \qquad M\xi = a$$

$$m_2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2 \qquad D\bar{\xi} = \frac{\sigma^2}{n} \to 0$$

$$s^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$$

$$Ms^2 = \sigma^2$$

$$DS^2 = \frac{\mu_4 - \mu_2^2}{n} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n}\right), \quad \mu_2 = \sigma^2$$

состоятельная оценка

$Ds^2 \to 0$

$\alpha_k = \frac{1}{n}\sum_{i=1}^{n}\xi_i^k$ асимптотически нормально $\sim (\alpha_k, \sqrt{D\alpha_k})$

$$\sqrt{D\alpha_k} = \frac{\alpha_{2k} - \alpha_k^2}{n}$$

$$m_k = \frac{1}{n}\sum_{i=1}^{n}\xi_i - \bar{\xi} \qquad (\mu_k, \sqrt{Dm_k})$$

$\xi_1, \dots, \xi_n$ $\quad$ $\eta_n = \eta_n(\xi_1, \dots, \xi_n)$ оценки или статистики θ

Точечные оценки

1) Метод моментов

2) Метод наибольшего правдоподобия

$\theta = (\theta_1, \dots, \theta_s)$ находим теорет. моменты α_k

$\alpha_k(\theta_1, \dots, \theta_s) = \alpha_k$, система из s уравнений, $k = 1, \dots, s$ $\Big\}$ метод моментов

Интервальная оценка

Точечные оценки дают прибл. значение неизвестного параметра. Если известен з-н распр-ия, а он не зависит от неизвестного параметра, то можно получать интервальные оценки неизвестного параметра.

$\xi_1, \dots, \xi_n \,\|\, \xi \sim F(x, \theta)$

$h_1(\xi_1, \xi_2, \dots, \xi_n)$ $\quad h_2(\xi_1, \xi_2, \dots, \xi_n)$ $\quad h_1() \le h_2()$

$\theta \quad P\{h_1() < \theta < h_2()\} \ge 1-\varepsilon$ $\quad \varepsilon = 2\alpha$

$[h_1, h_2]$ накрывает неизвестный параметр с заданной вер-тью => имеет место интервальная оценка.

$h_1 < \theta < h_2$

$1-\varepsilon = 0{,}999$

Можно построить доверительный интервал, когда существует оценка, зависящая от этого параметра, такая, что её распределение от этого парам. не зависит

$\eta = \eta(\xi_1, \dots, \xi_n, \theta)$

$F_\eta(y) = P\{\eta < y\}$ от θ не зависит

$\xi_i \quad F(x, \theta) \quad i = 1, \dots, n$

$\xi_1, \xi_2, \dots, \xi_n$ н.р. (a, σ) $\quad \Phi(a, \sigma, x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}} dt$

1) a - неизв., σ - неизв.

2) a - неизв., σ - изв.

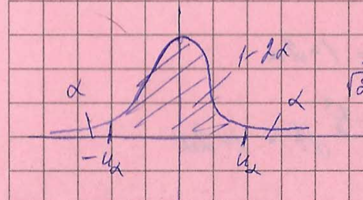
Найти доверит. интервал для a.

$\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i$

$\eta = \eta(\xi_1, \dots, \xi_n, a) = \frac{\bar{\xi} - a}{\sigma/\sqrt{n}}$

$\frac{\bar{\xi} - a}{\sigma/\sqrt{n}} \sim N(0, 1)$

$P\left\{\left|\frac{\bar{\xi} - a}{\sigma/\sqrt{n}}\right| < u\right\} = P\left\{-u < \frac{\bar{\xi} - a}{\sigma/\sqrt{n}} < u\right\} = \frac{1}{\sqrt{2\pi}} \int_{-u}^{u} e^{-\frac{t^2}{2}} dt = 1 - \varepsilon = 1 - 2\alpha$



$\frac{1}{\sqrt{2\pi}} \int_{u_\alpha}^{+\infty} e^{-\frac{t^2}{2}} dt = \alpha$

$P\left\{\underbrace{\bar{\xi} - u_\alpha \frac{\sigma}{\sqrt{n}}}_{h_1()} < a < \underbrace{\bar{\xi} + u_\alpha \frac{\sigma}{\sqrt{n}}}_{h_2()}\right\}$

$\left[\bar{\xi} - u_\alpha \frac{\sigma}{\sqrt{n}};\ \bar{\xi} + u_\alpha \frac{\sigma}{\sqrt{n}}\right]$

Лекция № 4. 28.03.2011 год.

$\xi \quad F_\xi(x) = F(x) \qquad (\xi_1, \xi_2, \dots, \xi_n) - \xi$

ξ: n-мерная сл. величина;
набор чисел

$F(x) = F(x, \Theta) \qquad \Phi(x) = \Phi(x, a, \sigma)$

? > Как по выборке оценивать эти параметры?

$\eta_n = \eta_n(\xi_1, \xi_2, \dots, \xi_n, \Theta)$

Оценка – любая функция выборки, являющаяся снова сл. величиной

$M\eta_n = \Theta$

$\eta_n \xrightarrow[n\to\infty]{P} \Theta$ состоятельная оценка.

$D\eta_n \xrightarrow[n\to\infty]{} 0$

$$P\{h_1(\xi_1, \dots, \xi_n) < \Theta < h_2(\xi_1, \dots, \xi_n)\} = 1 - 2\alpha$$

$[h_1, h_2]$ доверительный интервал

$\eta_n = \eta_n(\xi_1, \dots, \xi_n, \Theta)$; $F_{\eta_n}(x) = P(\eta_n < x)$

$\xi_i \sim N(a, \sigma^2)$

$$\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i \ ; \ S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$$

точечные оценки $M\xi$ и $D\xi$.

$$\eta_n = \frac{\bar{\xi} - a}{\sigma/\sqrt{n}} \sim N(0,1), \quad M\bar{\xi} = a, \quad D\bar{\xi} = \frac{\sigma^2}{n}$$

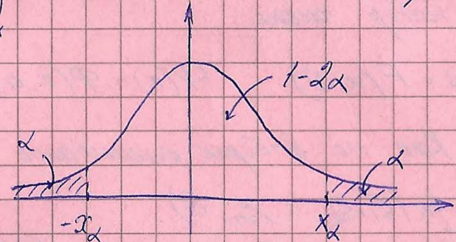
центрированная и
нормированная сл. величина

С помощью η_n можно строить доверит. интервалы.

1) Доверительный интервал для a при известном σ.

$$P\left\{\left|\frac{\bar\xi - a}{\sigma/\sqrt{n}}\right| < x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-x}^{x} e^{-\frac{t^2}{2}}\,dt = 1-2\alpha,$$ α — корень уравнения:

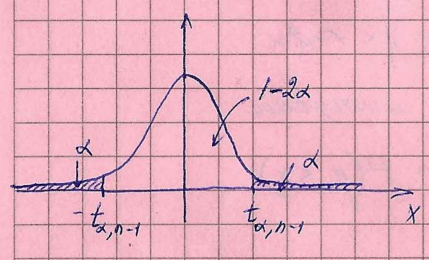
$$\frac{1}{\sqrt{2\pi}}\int_{x_\alpha}^{+\infty} e^{-\frac{t^2}{2}}\,dt = \alpha$$

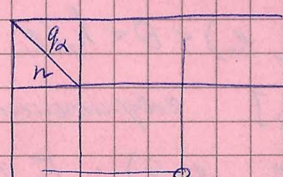


$$P\Big\{\underbrace{\bar\xi - x_\alpha \cdot \frac{\sigma}{\sqrt n}}_{h_1} < a < \underbrace{\bar\xi + x_\alpha\frac{\sigma}{\sqrt n}}_{h_2}\Big\} = 1-2\alpha$$

2) Доверит. интервал для a при неизвестном σ.

$$\frac{\bar\xi - a}{s/\sqrt n} = \tau_{n-1}$$ распределение Стьюдента с $(n-1)$ степенью свободы





берём $(n-1)$

τ_{n-1} асимпт. нормально $(0,1)$ при $n\to\infty$.

$$P\left\{\left|\frac{\bar\xi-a}{s/\sqrt n}\right| < t_{\alpha,n-1}\right\} = \int_{-t_{\alpha,n-1}}^{t_{\alpha,n-1}} s_{n-1}(u)\,du = 1-2\alpha$$

$$P\Big\{\underbrace{\bar\xi - t_{\alpha,n-1}\frac{s}{\sqrt n}}_{h_1} < a < \underbrace{\bar\xi + t_{\alpha,n-1}\frac{s}{\sqrt n}}_{h_2}\Big\} = 1-2\alpha$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n(\xi_i-\bar\xi)^2;\quad S^2 = \frac1n\sum_{i=1}^n(\xi_i-\bar\xi)^2 = m_2$$

$$S_0^2 = \sum_{i=1}^n (\xi_i-\bar\xi)^2$$

$$S_0^2 = nS^2 = (n-1)s^2$$

Пример.

$n = 16$ измерений заряда e-на

$\xi_1,\dots,\xi_{16}$

$\bar\xi = 4{,}779\cdot 10^{-10}$

$s = 0{,}0098$

$1-2\alpha = 0{,}99 \Rightarrow \alpha = 0{,}005$

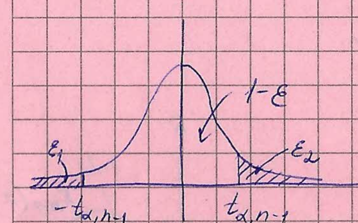
$t_{0{,}005,15} = 2{,}947$

$[4{,}772;\ 4{,}786]$ — доверительный интервал

$\cancel{P\{4{,}772 < e < 4{,}786\} = 0{,}99}$ так писать нельзя!

Мы имеем 3 константы, расположенные в определённом порядке. Запись бессмысленна.

$2t_{\alpha,n-1}\cdot\frac{s}{\sqrt n}$ — длина доверительного интервала



$\varepsilon_1+\varepsilon_2=\varepsilon$

При варьировании ε_1 и ε_2 меняется длина доверительного интервала.

Оптимально принять $\varepsilon_1=\varepsilon_2=\alpha$.

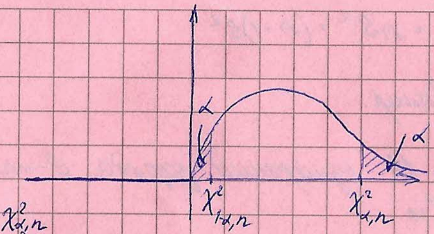
3) Доверит. интервал для σ при известном a.

$\xi_1,\dots,\xi_n \sim N(a,\sigma)$

$$\tilde S_0^2 = \sum_{i=1}^n(\xi_i-a)^2;\quad \frac{\tilde S_0^2}{\sigma^2} = \sum_{i=1}^n\left(\frac{\xi_i-a}{\sigma}\right)^2 = \chi^2_n$$

$$\frac{\xi_i-a}{\sigma}\sim N(0;1)$$

$$k_n(x) = \begin{cases} 0, & x<0 \\ \varphi(x)>0, & x\ge 0 \end{cases}$$



$$P\left\{\chi^2_{1-\alpha,n} < \frac{\tilde S_0^2}{\sigma^2} < \chi^2_{\alpha,n}\right\} = \int\limits_{\chi^2_{1-\alpha,n}}^{\chi^2_{\alpha,n}} k_n(x)\,dx$$

$$P\left\{\chi^2_n > \chi^2_{\alpha,n}\right\} = \alpha$$

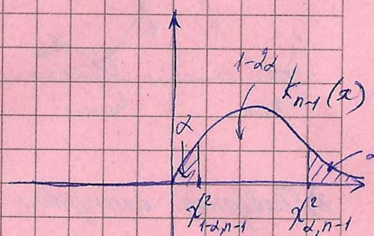
$$P\left\{\frac{\tilde S_0}{\chi_{\alpha,n}} < \sigma < \frac{\tilde S_0}{\chi_{1-\alpha,n}}\right\} = 1-2\alpha.$$

4) Доверительный интервал для σ при неизвестном a.

$$S_0^2 = \sum_{i=1}^{n} (\xi_i - \bar\xi)^2$$

По т. Фишера, $\frac{n m_2}{\sigma^2} = \frac{nS^2}{\sigma^2} = \frac{S_0^2}{\sigma^2} = \chi^2_{n-1}$

$$\frac{S_0^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{\xi_i - \bar\xi}{\sigma}\right)^2 = \underbrace{\eta_1^2 + \ldots + \eta_{n-1}^2}_{\text{независимые, н.р. } (0,1)} = \chi^2_{n-1}$$



$$P\left\{\chi^2_{1-\alpha,n-1} < \frac{S_0^2}{\sigma^2} < \chi^2_{\alpha,n-1}\right\} = 1-2\alpha$$

$$P\left\{\frac{S_0}{\chi_{\alpha,n-1}} < \sigma < \frac{S_0}{\chi_{1-\alpha,n-1}}\right\} = 1-2\alpha$$

* Посмотреть, как строится гистограмма по выборке. Гистограмма – статистический аналог функции распределения.

Статистическая проверка гипотез.

В прикладных задачах требуется проверить предположения (гипотезы) по экспериментальным данным.

$\xi_1, \ldots, \xi_n \sim N(a, \sigma)$

$\bar\xi \sim N\left(a, \frac{\sigma}{\sqrt{n}}\right)$

$\xi_1, \ldots, \xi_n$

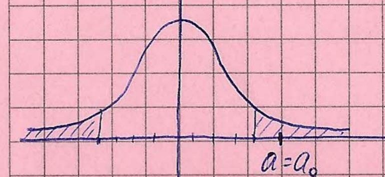
$M\xi_i = n_i$; $D\xi_i = \sigma^2$

$\bar\xi$ асимпт. $N\left(a, \frac{\sigma}{\sqrt{n}}\right)$

Проверить, имеет ли параметр некоторое значение Θ_0.

$\xi_1, \ldots, \xi_n$ $\quad F(x, \Theta)$

1) $\Theta = \Theta_0$ согласуется ли наша выборка с таким предположением о значении параметра.



гипотеза отвергается

Для различающихся $\bar\xi_{1,2}$ двух выборок установить, являются ли различия случайными и можно объединить выборки, или все различия не случайны.

$\xi_1, \xi_2, \ldots, \xi_{n_1}$ $\quad n_1$ $\quad \bar\xi = \frac{1}{n_1}\sum_{i=1}^{n_1} \xi_i$

$\eta_1, \eta_2, \ldots, \eta_{n_2}$ $\quad n_2$ $\quad \bar\eta = \frac{1}{n_2}\sum_{i=1}^{n_2} \eta_i$

$\xi_1, \xi_2, \ldots, \xi_{n_1}, \eta_1, \ldots, \eta_{n_2}$.

$a_1 = M\xi \qquad F(x, a_1) = \{\xi\}$

$a_2 = M\eta \qquad F(x, a_2) = \{\eta\}$

Пример 1.

Наша выборка – n независимых сл. величин, нормально распределённых с параметрами (a, σ), σ – неизвестно.

$H:\ a = a_0$

$1 - 2\alpha = P_0$

$$P\left\{\bar{\xi} - t_{\alpha,n-1}\frac{s}{\sqrt{n}} < a < \bar{\xi} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right\} = \int_{-t_{\alpha,n-1}}^{t_{\alpha,n-1}} s_{n-1}(x)\,dx = 1 - 2\alpha$$

Если $a_0 \in \left[\bar{\xi} - t_{\alpha,n-1}\frac{s}{\sqrt{n}};\ \bar{\xi} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right]$, то гипотеза H не отвергается (или проверяется).

Если действительно $a = a_0$, то

$$\frac{\bar{\xi} - a_0}{s/\sqrt{n}} = \tau_{n-1}$$

$$P\left\{\left|\frac{\bar{\xi} - a_0}{s/\sqrt{n}}\right| \geq t_{\alpha,n-1}\right\} = 2\alpha$$

$$P\left\{\frac{|\bar{\xi} - a_0|}{1} \geq t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right\} = 2\alpha \quad (1)$$

$\xi_1, \ldots, \xi_n$ – реализация выборки (результат эксперимента).

$\bar{\xi}_э = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ — среднее по экспериме́нт. данным.

$|\bar{\xi}_э - a_0| \geq t_{\alpha,n-1}\frac{s}{\sqrt{n}} \qquad H:\ a = a_0$ отвергается

Соотношение (1) показывает, что если гипотеза H верна, то событие $|\bar{\xi}_э - a_0| \geq t_{\alpha,n-1}\frac{s}{\sqrt{n}}$ маловероятно.

Если для данной выборки обнаруживается, что $|\bar{\xi}_э - a_0| \geq t_{\alpha,n-1}\frac{s}{\sqrt{n}}$ выполняется, то гипотезу H следует отвергнуть.

11.04.2011 год. Лекция № 5.

$\xi_1, \xi_2, \dots, \xi_n$ $\quad \xi_i$ н.р. (a, σ) или (a, σ^2)

$M\xi_i = a; \; D\xi_i = \sigma^2$

$\eta = \dfrac{\bar\xi - a}{\sigma/\sqrt{n}} \sim N(0;1)$ $\quad \sigma$ – известная величина

$\tau_{n-1} = \dfrac{\bar\xi - a}{s/\sqrt{n}}$ $\quad$ Стьюдента с $(n-1)$ степ. свободы
σ – неизвестно

$$P\{h_1(\xi_1,\dots,\xi_n) < a < h_2(\xi_1,\dots,\xi_n)\} = 1-2\alpha$$

$$P\left\{-u_\alpha < \frac{\bar\xi - a}{\sigma/\sqrt{n}} < u_\alpha\right\} = \frac{1}{\sqrt{2\pi}}\int_{-u_\alpha}^{u_\alpha} e^{-\frac{t^2}{2}}dt = 1-2\alpha$$

$$P\left\{-t_{\alpha,n-1} < \frac{\bar\xi - a}{s/\sqrt{n}} < t_{\alpha,n-1}\right\} = 1-2\alpha$$

$H: \; a = a_0$

Если H верна, то $\dfrac{\bar\xi - a_0}{\sigma/\sqrt{n}} \sim N(0;1)$

$\dfrac{\bar\xi - a_0}{s/\sqrt{n}} \sim S$ с $(n-1)$ сс

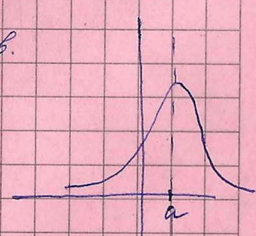
$|\xi_э - a_0| > t_{\alpha,n-1}$ $\Rightarrow$ отвергаем гипотезу

Пример.

$\xi_1,\dots,\xi_{n_1} \sim N(a_1,\sigma_1)$ $\quad a_1, a_2$ – неизв.

$\eta_1,\dots,\eta_{n_2} \sim N(a_2,\sigma_2)$ $\quad \boxed{H: \sigma_1 = \sigma_2}$

$\lambda = \dfrac{\sigma_1^2}{\sigma_2^2} = 1.$



Обозначения:

$\xi_1,\dots,\xi_n$ – выборка

$\bar\xi = \dfrac{1}{n}\sum_{i=1}^{n}\xi_i$ – среднее значение (несмещ. оценка)

$m_2 = S^2 = \dfrac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar\xi)^2$ $\quad$ смещённая оценка

$s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar\xi)^2$

$S_0^2 = \sum_{i=1}^{n}(\xi_i - \bar\xi)^2$

$\tilde S_0^2 = \sum_{i=1}^{n}(\xi_i - a)^2$

Лемма Фишера

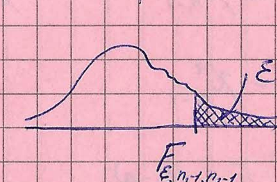
$$\frac{s_1^2}{\sigma_1^2} = \frac{1}{(n_1-1)\sigma_1^2}\sum_{i=1}^{n_1}(\xi_i - \bar\xi)^2 = \frac{1}{n_1-1}\chi^2_{n_1-1}$$

$$\frac{s_2^2}{\sigma_2^2} = \frac{1}{(n_2-1)\sigma_2^2}\sum_{j=1}^{n_2}(\eta_j - \bar\eta)^2 = \frac{1}{n_2-1}\chi^2_{n_2-1}$$

$$F(\xi,\eta,\lambda) = \frac{s_1^2\cdot\sigma_2^2}{\sigma_1^2\cdot s_2^2} = \frac{s_1^2}{s_2^2}\cdot\frac{1}{\lambda} = F_{n_1-1,n_2-1}$$

Зададим $\varepsilon > 0$

$$P\{\underbrace{F_{n_1-1,n_2-1}}_{\text{сл. величина}} \ge \underbrace{F_{\varepsilon,n_1-1,n_2-1}}_{\text{число}}\} = \varepsilon$$



Можно построить доверит. интервал для λ.

$\lambda = 1$ $\quad P(h_1(\xi,\eta) < 1 < h_2(\xi,\eta)) = 1-\varepsilon$

$F_{\text{эксп}}(\xi, \eta, \lambda = 1) > F_{\varepsilon, n_1-1, n_2-1}$ => отвергаем гипотезу

Пример 2. Даны две выборки:

$\xi_1, \dots, \xi_{n_1} \sim N(a_1, \sigma)$ σ – неизвестно

$\eta_1, \dots, \eta_{n_2} \sim N(a_2, \sigma)$ $\boxed{H: a_1 = a_2}$

Строим доверительный интервал для

$\lambda = a_1 - a_2$ (нам важно, чтобы $\lambda = 0$).

$\bar\xi - \bar\eta$ норм. распр. $M(\bar\xi - \bar\eta), D(\bar\xi - \bar\eta)$

$M\bar\xi - M\bar\eta = M(\bar\xi - \bar\eta) = a_1 - a_2$

$$D(\bar\xi - \bar\eta) = D\bar\xi + D\bar\eta = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \sigma^2 \cdot \frac{n_1 + n_2}{n_1 n_2}$$

$$\frac{\bar\xi - \bar\eta - (a_1 - a_2)}{\sigma\sqrt{\frac{n_1+n_2}{n_1 n_2}}} \sim N(0; 1)$$

$\xi, \xi_1, \dots, \xi_n$ независимы и н.р. $(0; 1)$

$$\frac{\xi}{\sqrt{\frac{1}{n}\chi_n^2}} = \tau_n$$

$$\chi_n^2 = \xi_1^2 + \dots + \xi_n^2$$

$$\frac{n_1 S_1^2}{\sigma^2} \sim \chi^2_{n_1-1} \qquad \chi^2_{n_1-1} + \chi^2_{n_2-1} = \chi^2_{n_1+n_2-2}$$

$$\frac{n_2 S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

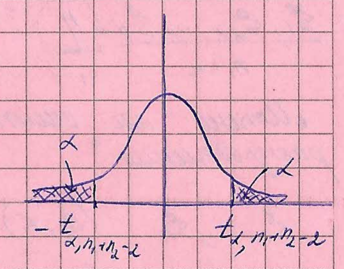
$$\frac{\dfrac{\bar\xi - \bar\eta - (a_1 - a_2)}{\sigma\sqrt{\frac{n_1+n_2}{n_1 n_2}}}}{\sqrt{\frac{1}{n_1+n_2-2}\underbrace{\left(\frac{n_1 S_1^2}{\sigma^2} + \frac{n_2 S_2^2}{\sigma^2}\right)}_{\chi^2_{n_1+n_2-2}}}} = \tau_{n_1+n_2-2}$$

$$\frac{\bar\xi - \bar\eta - (a_1 - a_2)}{\sqrt{\frac{1}{n_1+n_2-2}(n_1 S_1^2 + n_2 S_2^2)} \cdot \sqrt{\frac{n_1+n_2}{n_1 n_2}}} = \tau_{n_1+n_2-2}$$

$$P(\underline{h_1(\xi, \eta)} < a_1 - a_2 < \underline{h_2(\xi, \eta)}) = 1 - 2\alpha$$

$0 \in [h_1, h_2] \Rightarrow a_1 = a_2$

ноль



$$\zeta = \frac{\bar\xi - \bar\eta - 0}{\sqrt{\frac{1}{n_1+n_2-2} \cdot \frac{n_1+n_2}{n_1 n_2} \cdot (n_1 S_1^2 + n_2 S_2^2)}} = \tau_{n_1+n_2-1} \qquad \zeta$$

дзета

Задача (из Золотова "")

ξ_1				ξ_5
0,75	0,72	0,73	0,74	0,72
η_1			η_4	
0,74	0,76	0,75	0,73	

Рассмотреть выборки на предмет одинаковой воспроизводимости

I) $\bar\xi = 0{,}73$ $S_1^2 = 0{,}000170$ $n_1 - 1 = 4$

II) $\bar\eta = 0{,}74$ $S_2^2 = 0{,}000125$ $n_2 - 1 = 3$

$F_{\text{таб}} = 9{,}1$ $F_{\text{э}} = 1{,}36$

$\varepsilon = 0{,}05$

$F_{\text{э}} < F_{\text{таб}}$ => оснований отвергать гипотезу нет

$2\alpha = 0{,}01$ $n_1 + n_2 - 2 = 7$ $\tau_{\text{таб}} = 3{,}5$

$|\zeta_{\text{э}}| = 1{,}86 < 3{,}5$

Значит, наша гипотеза о том, что выборки принадлежат к одному распределению, не отвергается.

Отбраковка аномальных результатов (выявление промахов).

$$\underbrace{\xi_1, \xi_2, \dots, \xi_n}_{n_1 = n}, \underbrace{\eta}_{n_2 = 1} \ ?$$

Можно ли считать, что $\{\xi\}$ и $\{\eta\}$ взяты из одного распределения?

$$\left.\begin{matrix} \xi_1, \dots, \xi_n & (a_1, \sigma) \\ \eta & (a_2, \sigma) \end{matrix}\right\} \quad a_1 = a_2 \ ?$$

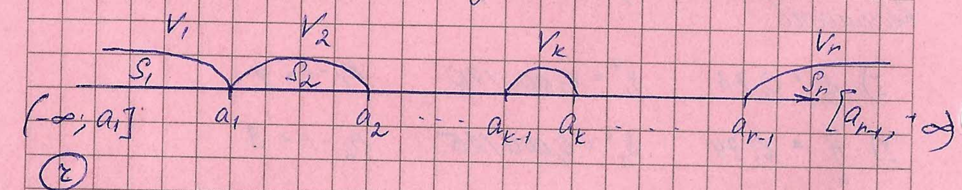
$$\tau_{n-1} \qquad S_2^2 = 0 \qquad \cancel{\frac{1}{n_2}\sum_{i=1}^{n_2}(\eta_i - \bar{\eta})^2} \to 0$$

Задача.

Критерий согласия.

$\xi_1, \xi_2, \dots, \xi_n$ 1) $F(x)$ — полностью известна

2) $F(x, \theta)$ — известно с точностью



Значения выборки наносят на числовую ось и считают, сколько значений попало в каждый интервал. Число попавших значений обозначают ν_i.

1) H: $F(x)$; $p_1 = P\{\xi \in S_1 | H\} = F(a_1)$

$p_2 = P\{\xi \in S_2 | H\} = F(a_2) - F(a_1)$

$p_k = P\{\xi \in S_k | H\} = F(a_k) - F(a_{k-1})$

$p_r = P(\xi \in S_r | H) = 1 - F(a_{r-1})$

По т. Бернулли,

$$\frac{\nu_k}{n} \xrightarrow[n \to \infty]{P} p_k$$

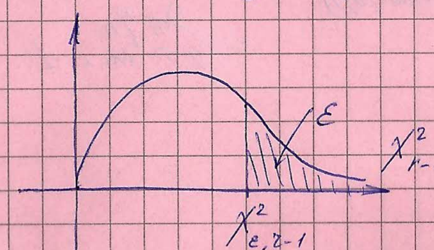
$$\sum \left(\frac{\nu_k}{n} - p_k\right)^2$$

сумма квадратов будет маленькой, поэтому добавляют ортонормирующий множитель.

$$\chi^2 = \sum_{k=1}^{r} \frac{n}{p_k}\left(\frac{\nu_k}{n} - p_k\right)^2 = \sum_{k=1}^{r} \frac{(\nu_k - np_k)^2}{np_k} \qquad (*)$$

Т. Пирсона: Если H верна, то сл. величина $(*)$ при любой функции $F(x)$ имеет в пределе распределение χ^2_{r-1}, $n \to \infty$.

Распределение не зависит от разбиения на интервалы, если гипотеза верна.



23.05.2011 г.

§9.6 п. 3 Общие понятия о стат. проверке гипотез

⑩ Построение приближ. вероятн. интервала для числа успехов в схеме Бернулли.

(n, p) – по выборке t построить дов. интервал.

$$\eta_n = \frac{\frac{\mu_n}{n} - p}{\sqrt{\frac{npq}{n^2}}}$$ асимпт. нормально $(0;1)$

$$\frac{\mu_n - np}{\sqrt{npq}} \sim N(0,1)$$

$$\zeta_n = \frac{\frac{\mu_n}{n}-p}{\sqrt{\frac{1}{n}\cdot\frac{\mu_n}{n}\left(1-\frac{\mu_n}{n}\right)}} = \frac{\frac{\mu_n}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \cdot \frac{\sqrt{\frac{p(1-p)}{n}}}{\sqrt{\frac{1}{n}\cdot\frac{\mu_n}{n}\left(1-\frac{\mu_n}{n}\right)}} = \xi_n \eta_n$$

ас. норм. $(0,1)$

$\lambda_n \eta_n$ по вер-ти сх. к 1

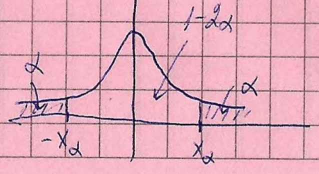
$$\xi_n = \frac{\frac{\mu_n}{n}-p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$\eta_n \xrightarrow[n\to\infty]{P} 1 \qquad \frac{\mu_n}{n} \xrightarrow[n\to\infty]{P} p$$

$P \to 1$

$\zeta_n = \xi_n \cdot \eta_n \Rightarrow \zeta_n$ распр. так же, как и ξ_n.

$$P\left\{\left|\frac{\frac{\mu_n}{n}-p}{\sigma_n}\right| < x_\alpha\right\} \simeq \frac{1}{\sqrt{2\pi}}\int_{-x}^{x} e^{-\frac{t^2}{2}}dt = 1-2\alpha$$



$$\frac{1}{\sqrt{2\pi}}\int_{x_\alpha}^{\infty} e^{-\frac{t^2}{2}}dt$$

$$P\left\{\frac{\mu_n}{n} - x_\alpha\sigma_n < p < \frac{\mu_n}{n} + x_\alpha\sigma_n\right\} \simeq 1-2\alpha$$

Общие понятия о проверке статистических гипотез.

В прак. зад. часто требуется по эмпир. данным проверить те или иное предп-ие отн-но свойств з-на распр. наблюдаемой сл. величины.

1) $F(x)$ или $F(x, \theta)$ критерий хи-квадрат.

2) По наблюдённым значениям установить, имеет ли параметр распред. θ некоторого значения $\theta = \theta_0$ или $\theta_0 \in [\theta_1, \theta_2]$ (согл. ли данные с нашим предположением).

3) Провер. гипотеза о равенстве дисперсий и средних

Статистические гипотезы

Гипотеза однозн. фиксирует з-н распр-ия $\Rightarrow$ $\Rightarrow$ простая гипотеза; в противном случае – сложная.

(a, σ) изв. / неизв. $\quad H: a = a_0$

гипотеза о том, что $\theta_0 \in [\theta_1, \theta_2]$ – слож.

Задан нек. стат. критерий для проверки гипотезы H_0, если сформ. правило, согласно которому приним. решение: согласуется ли набл. значение с этой гипотезой (H_0 – нулевая или основная), или эта гипотеза должна быть отвергнута.

$\xi = (\xi_1, \xi_2, \dots, \xi_n)$

$X = (x_1, x_2, \dots, x_n)$ наблюд. значения нек. сл. вел. ξ

Выделяют в выборочном пр-ве S – крит. мн-во (площадь его является теми значениями, которые принимает выборка).

R^n. $S \subset R^n$

$X \in S$ — считаются нехарак. для H_0

$X \in \bar{S}$ дополнение — характерны для H_0

H_0 отвергается, если конкр. реализация выборки попадает в мн-во S.

$D(x_1, x_2, \dots, x_n)$

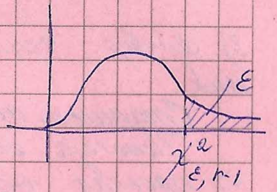
эти нер-ва и определяют крит. мн-во S.

$D(x_1, x_2, \dots, x_n) > C \Rightarrow H$ отвергается

статист.

$S = \{(x_1, \dots, x_n) : D(x_1, \dots, x_n) > C\}$.

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} \to \chi^2_{r-1}$$



$$\{\chi^2_{выб} > \chi^2_{\varepsilon, r-1}\} = \{(x_1, \dots, x_n) : \chi^2 > \chi^2_{\varepsilon, r-1}\}$$

Для каждого крит. возможны ошибки 2 видов:

1) Ош. $1^{ого}$ рода: гипотеза H_0 отвергнута, хотя на самом деле она верна.

α – вер-ть ошибки первого рода.

$n = 100$; отвергаем гипотезу о "честности" монеты, если все 100 раз выпадет, например, герб.

$$\alpha = \left(\frac{1}{2}\right)^{100}$$

Когда гипотеза сложная, то в кач-ве ошибки $1^{ого}$ рода можно определить

$$\alpha = \sup_{\theta \in H_0} P\{X \in S | H_0\}$$

2) Ошибка $2^{ого}$ рода

Гипотеза H_0 ложная, но она не отвергнута

β – вер-ть ошибки $2^{ого}$ рода.

Пример.

н.р. (a, σ) $\quad H_0 : a = a_0$

$H_1 : a = a_1, \quad a_1 \neq a_0, \quad a_0 < a_1$

H_0 : н.р. (a_0, σ)

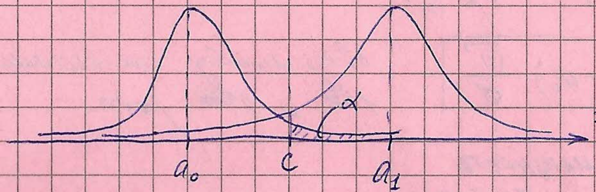
H_1 : н.р. (a_1, σ)

Сл. вел. может принимать только значения a_0 или a_1, а мы хотим проверить, какое именно из них принимает сл. вел.

$x_1, x_2, x_3, \dots, x_n \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

$H_0 : \bar{x} \sim N(a_0, \frac{\sigma}{\sqrt{n}})$

$H_1 : \bar{x} \sim N(a_1, \frac{\sigma}{\sqrt{n}})$



$\bar{x} = D(x_1, \dots, x_n)$

$\bar{x} > C \Rightarrow H_0$ отверг.

$\bar{x} \leq C \Rightarrow H_0$ не отв.

$S = \{(x_1, \dots, x_n) : \bar{x} > C\}$

α – задано; из усл-ия заданного α определяют границу – константу C.

$\frac{\bar{x} - a_0}{\sigma / \sqrt{n}}$

$$\alpha = P\{\bar{x} > C\} = P\left\{\frac{\bar{x} - a_0}{\sigma/\sqrt{n}} > \frac{C - a_0}{\sigma/\sqrt{n}}\right\} = \frac{1}{\sqrt{2\pi}} \int_{x_\alpha}^{\infty} e^{-\frac{t^2}{2}} dt,$$

x_α – корень ур-ия $\frac{1}{\sqrt{2\pi}}\int\limits_{+x_\alpha}^{\infty} e^{-\frac{t^2}{2}}\,dt = \alpha$

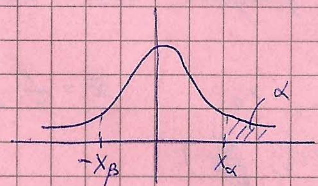
$\frac{C-a_0}{\sigma/\sqrt{n}} = x_\alpha$ $\quad \boxed{C_\alpha = a_0 + x_\alpha \frac{\sigma}{\sqrt{n}}}$

H_0 неверна, то верна H_1 $\quad \bar{x} \sim N\left(a_1, \frac{\sigma}{\sqrt{n}}\right)$

С какой вер-тью мы примем H_0, если она неверна?

$$\beta = P\{\bar{x} \le \underset{a_0 + x_\alpha\frac{\sigma}{\sqrt{n}}}{\underset{\shortparallel}{C_\alpha}} \mid H_1\} = P\left\{\underset{\text{н.р. }(0,1)}{\frac{\bar{x} - a_1}{\sigma/\sqrt{n}}} \le \frac{a_0 + x_\alpha\frac{\sigma}{\sqrt{n}} - a_1}{\sigma/\sqrt{n}}\right\}$$ вер-ть ош. 2-ого рода

$\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{-x_\beta} e^{-\frac{t^2}{2}}\,dt$



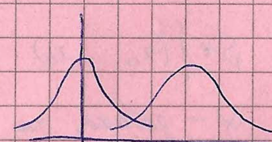
$\alpha = \beta$

$a_0 + x_\alpha\frac{\sigma}{\sqrt{n}} - a_1 = -x_\beta\frac{\sigma}{\sqrt{n}}$

$\boxed{x_\alpha + x_\beta = (a_1 - a_0)\frac{\sqrt{n}}{\sigma}}$ связь между ошибками 1-ого и 2-ого рода

$W = 1-\beta$ мощность критерия

$S = \{(x_1,\dots,x_n) : \bar{x} > a_0 + x_\alpha \frac{\sigma}{\sqrt{n}}\}$



β – min

$W = 1-\beta$ – max $\Rightarrow$ критерий наиболее мощный

* Чистяков: теорема Неймана–Пирсона

Для дискретного распр-ия: $P\{x \in S \mid H_0\} \le \alpha$

(вер-ть не всегда может быть точно равна числу)

$(x_1, x_2, \dots, x_n)$ $\quad \left.\begin{matrix} H_0 : p_0(x_1,\dots,x_n) \\ H_1 : p_1(x_1,\dots,x_n)\end{matrix}\right\} =$

$= \begin{cases} p_0(x_1)\,p_0(x_2)\dots p_0(x_n) \\ p_1(x_1)\,p_1(x_2)\dots p_1(x_n)\end{cases}$

Среди всех критериев различ. гипотез H_0 и H_1 с заданной ошибкой 1-ого рода α наиболее мощным критерием является критерий с крит. мн-вом

$$S_C = \{(x_1,\dots,x_n) : \underset{(a_1,\sigma)}{p_1(x_1,\dots,x_n)} \ge C\,\underset{(a_0,\sigma)}{p_0(x_1,\dots,x_n)}\}$$

C выбирается так, чтобы ошибка 1-ого рода α.

$$S_C = \left\{(x_1,\dots,x_n) : \prod_{i=1}^{n} \cancel{\frac{1}{\sqrt{2\pi}\sigma}}\, \underset{a_1,\bar{x}}{e^{-\frac{(x_i-a_1)^2}{2\sigma^2}}} \ge C \prod_{i=1}^{n} \cancel{\frac{1}{\sqrt{2\pi}\sigma}}\, \underset{a_0,\bar{x}}{e^{-\frac{(x_i-a_0)^2}{2\sigma^2}}}\right\} \qquad \bar{x} > C_1(c)$$

$\{\bar{x} > a_0 + x_\alpha \frac{\sigma}{\sqrt{n}}\}$