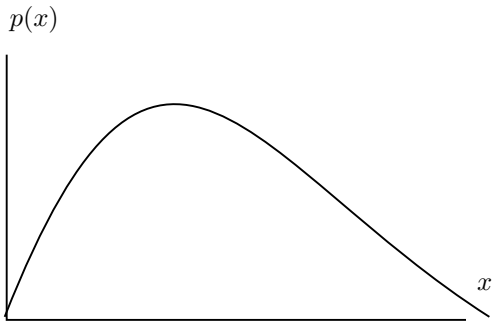
1 Занятие 1

1.1 Основные распределения в Мат Стат

1.1.1 Гамма распределение

$$\xi \sim \Gamma(\lambda, a)\ \lambda > 0, a > 0$$
$$P(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}\ \{(0, +\infty)\}$$



$$\Gamma(S+1) = S\Gamma(S) \qquad \Gamma(S) = \int_0^\infty x^{s-1} e^{-x} dx\ x > 0$$
$$M[\Gamma] = \int_{-\infty}^{\infty} \rho(x) dx = \int_0^\infty \frac{\lambda^a}{\Gamma(a)} (\frac{t}{\lambda})^a e^{-t} \frac{dt}{\lambda} == \frac{1}{\lambda\Gamma(a)} \int_0^\infty t^a e^{-t} dt = \frac{a}{\lambda}$$
$$D[\xi] = M[\xi^2] - M^2[\xi] = \frac{a^2 + a}{\lambda^2} - (\frac{a}{\lambda})^2 = \frac{a}{\lambda^2}$$

Теорема 1.1 (Свойство суммы)**.** $\xi_1, \ldots, \xi_n$ независимы, $\xi_i \sim \Gamma(\lambda, a_i)$, $\eta = \xi_1 + \cdots + \xi_n \sim \Gamma(\lambda, a_1 + \cdots + a_n)$

Доказательство. $\xi_1 \sim \Gamma(\lambda, a_1)$. $\xi_2 \sim \Gamma(\lambda, a_2)$ - независимые, $\eta = \xi_1 + \xi_2$

$$\Phi(y) = P(\eta < y) = P(\xi_2 + \xi_2 < y) = \iint_{x_1+x_2<y} p(x_1, x_2)dx_1dx_2 =$$
$$= \int_0^y dx_2 \int_0^{y-x_1} \frac{\lambda^{a_1}}{\Gamma(a_1)x_1^{a_1-1}e^{-\lambda x_1}} \frac{\lambda^{a_2}}{\Gamma(a_2)x_2^{a_2-2}e^{-\lambda x_2}} dx_2$$
$$\varphi(y) = \Phi'(y)$$

□

1.1.2 Распределение Пирсона χ^2

$\xi_i \sim N(0,1)$ - независимы, $\eta = \xi_1^2 + \cdots + \xi_n^2 = \chi^2$

$$\Phi(y) = P(\xi^2 < y) = \begin{cases} y \le 0 & : 0 \\ y > 0 & : P(-\sqrt{y} < \xi < \sqrt{y}) \end{cases}$$
$$\varphi(x) = \begin{cases} \frac{1}{2\sqrt{y}}F'(\sqrt{y}) + \frac{1}{2\sqrt{y}}F'(-\sqrt{y}),\ y > 0 \\ 0,\ y < 0 \end{cases}$$
$$p(x) = \frac{e^{x^2/2}}{\sqrt{2\pi}}$$
$$\varphi(y) = \begin{cases} \frac{1}{\sqrt{y}}\frac{e^{-y/2}}{\sqrt{2\pi}},\ y > 0 \\ 0,\ y \le 0 \end{cases}$$
$$p(x) = \frac{\lambda^a}{\Gamma(a)}x^{a-1}e^{-\lambda x}\{(0;+\infty)\} \qquad \lambda = \frac{1}{2} \qquad a = \frac{1}{2}$$
$$\xi^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2}) \qquad \xi_1^2 + \cdots + \xi_n^2 \sim \Gamma(\frac{1}{2}, \frac{n}{2}) = \chi^2(n)$$

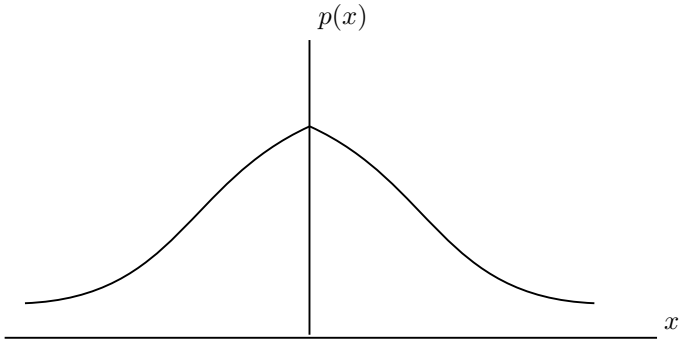
n - число степеней свободы

$$M[\eta] = \frac{a}{\lambda} = \frac{n/2}{1/2} = n$$
$$D[\eta] = \frac{a}{\lambda^2} = \frac{n/2}{1/4} = 2n$$

Теорема 1.2 (Свойство суммы)**.** $\xi_1, \ldots, \xi_m$ - независ, $\xi_i \sim \chi^2(n_i)$, $\xi_1 + \cdots + \xi_n \sim \chi^2(n_1 + \cdots + n_m)$

1.2 Распределение Стьюдента (Госсет)

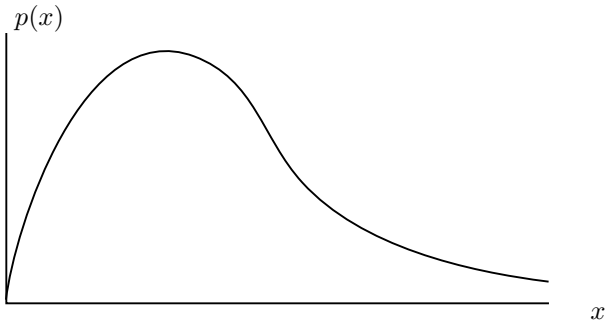
$\xi \sim N(0,1)$, $\eta \sim \chi^2(m)$ - независимы, $\frac{\xi}{\sqrt{\eta/m}} \sim t(m)$



$$p(x) = \frac{(m)^{m/2}\Gamma(\frac{m+1}{2})}{\sqrt{\pi}\Gamma(\frac{m}{2})(x^2+m)^{\frac{m+1}{2}}}$$

1.3 Распределение Фишера

$\xi \sim \chi^2(n)$, $\eta \sim \chi^2(m)$ - независимые, $\frac{\xi/n}{\eta/m} \sim F(n,m)$



1.4 Нормальное распределение

$$p(\vec{x}) = \frac{1}{(\sqrt{2\pi})^n}\frac{1}{\sqrt{detK}}e^{-\frac{1}{2}(\vec{x}-\vec{a})^T K^{-1}(\vec{x}-\vec{a})}$$

$$\vec{\xi} \sim N(\vec{a}, R)$$

Свойства:

- $\xi \sim N(0,1)$, $\eta = a\xi + b \sim N(b, a^2)$
- $\xi \sim N(\alpha, \sigma^2)$, $\eta = a\xi + b \sim N(a\alpha + b, \sigma^2 a^2)$
- $\xi \sim N(\vec{0}, E)$,$\vec{\eta} = A\vec{\xi} + \vec{b}$, $A : n \times n, detA \neq 0$

$$\Phi(t_1, \dots, t_n) = P(\eta_1 < t_1, \dots, \eta_n < t_n) = P(\vec{\eta} < \vec{t}) = P(A\vec{\xi} + \vec{b} < \vec{t}) =$$
$$= \int \cdots \int_{A\vec{x}+\vec{b}<\vec{t}} p(x_1, \dots, x_n) dx_1 \dots dx_n =$$
$$\vec{y} = A\vec{x} + \vec{b} \qquad J = \left|\frac{\partial \vec{x}}{\partial \vec{y}}\right| \qquad \frac{1}{J} = detA$$
$$= \int \cdots \int_{\vec{y}<\vec{t}} p(A^{-1}(\vec{y} - \vec{b})) \frac{1}{|detA|} dy_1 \dots dy_n$$
$$\varphi(\vec{t}) = p(A^{-1}(\vec{y} - \vec{b})) \frac{1}{|detA|}$$
$$\varphi(\vec{t}) = \frac{1}{|detA|} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}(A^{-1}(\vec{y}-\vec{b})^T)(A^{-1}(\vec{y}-\vec{b}))} =$$
$$= \frac{1}{|detA|} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}(\vec{t}-\vec{b})^T (A^T)^{-1} A^{-1} (\vec{t}-\vec{b})}$$
$$K = AA^T \qquad \vec{\eta} = A\vec{\xi} + \vec{b} \sim N(\vec{b}, AA^T)$$

- $\xi \sim N(\vec{a}, K)$, $\vec{\eta} = A\vec{\xi} + \vec{b} \sim N(A\vec{a} + \vec{b}, AKA^T)$, $A : n \times n$, $detA \neq 0$
- Для $A : m \times n$ два предыдущих свойства так же верны
- ξ, η - независ $\Rightarrow cov(\xi, \eta) = 0$, в другую сторону не верно

$$\begin{cases} \xi \sim N(a_1, \sigma_1^2) \\ \eta \sim N(a_2, \sigma_2^2) \\ \text{независимые} \end{cases} \Leftrightarrow (\xi, \eta) \sim N\left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\right)$$

$$\begin{cases} \xi \sim N(a_1, \sigma_1^2) \\ \eta \sim N(a_2, \sigma_2^2) \\ cov(\xi, \eta) = 0 \end{cases} \Leftarrow (\xi, \eta) \sim N\left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\right)$$

Лемма 1.1 (Лемма Фишера)**.** Пусть $\vec{\xi} \sim N(\vec{0}, E)$ и C ортогональная матрица, $\vec{\eta} = C\vec{\xi}$, тогда $\forall k = 1 \dots n-1$ сл. вел. $\varkappa = \sum_{i=1}^{n} \xi_i^2 - \eta_1^2 - \eta_2^2 - \cdots - \eta_k^2 \sim \chi^2(n-k)$ и вел $\varkappa, \eta_1, \eta_2, \dots, \eta_k$ независ.

Доказательство.

$$\vec{\eta} \sim N(\vec{0}, \underbrace{CC^T}_{E})$$
$$\eta_1^2 + \cdots + \eta_n^2 = \vec{\eta}^T\vec{\eta} = \vec{\xi}^T C^T C\vec{\xi} = \xi_1^2 + \cdots + \xi_n^2$$
$$\varkappa = \eta_{k-1}^2 + \cdots + \eta_n^2$$
$$\varkappa = \chi^2(n-k)$$

□

Теорема 1.3 (Фишера)**.** Пусть $\xi_1, \ldots, \xi_n$ независ и $\xi_i \sim N(a, \sigma^2)$, тогда:

1. $\varphi = \sqrt{n}\frac{\bar{\xi}-a}{\sigma} \sim N(0,1)$, $\bar{\xi} = \frac{1}{n}\sum_1^n \xi_i$
2. $\psi = \sum_{i=1}^n \frac{(\xi_i - \bar{\xi})^2}{\sigma^2} \sim \chi^2(n-1)$
3. φ и ψ независ.

Доказательство.

$$\varphi = \frac{1}{\sqrt{n}}\frac{\sum^n \xi_i - na}{\sigma} = \frac{1}{\sqrt{n}}\sum^n\left(\frac{\xi_i - a}{\sigma}\right)$$
$$\frac{\xi_i - a}{\sigma} = \frac{1}{\sigma}\xi_i - \frac{a}{\sigma} \sim N(\frac{a}{\sigma} - \frac{a}{\sigma}, \sigma^2\frac{1}{\sigma^2}) = N(0,1)$$
$$\varphi = \frac{1}{\sqrt{n}}\sum^n \eta_i = (\frac{1}{\sqrt{n}} \ldots \frac{1}{\sqrt{n}})\vec{\eta} \sim N(\vec{0}, AA^T) = N(0,1)$$

1) Доказан

$$\psi = \sum^n\left(\underbrace{\frac{\xi_i - a}{\sigma}}_{\eta_i \sim N(0,1)} - \underbrace{\frac{\bar{\xi} - a}{\sigma}}_{\bar{\eta}}\right)^2 = \sum^n(\eta_i - \bar{\eta})^2 = \sum^n(\eta_i^2 - 2\eta_i\bar{\eta} + (\bar{\eta})^2) =$$
$$= \sum^n \eta_i^2 - 2\bar{\eta}\sum^n \eta_i + n(\bar{\eta})^2 = \sum^n \eta_i^2 - n(\bar{\eta})^2$$

$$\eta_i \sim N(0,1) \qquad \zeta^2 = n\bar{\eta}^2 \qquad \zeta = \sqrt{n}\bar{\eta} = \frac{1}{\sqrt{n}}\sum^n \eta_i = A\vec{\eta} = \varphi$$

$A = \left(\frac{1}{\sqrt{n}} \ldots \frac{1}{\sqrt{n}}\right) \Rightarrow C$ - ортог. матрица (Грамма-Шмидта)

(A получается строчкой матрицы C и тогда ζ - одна из координат в другом базисе и применима Лемма Фишера)

По лемме Фишера $\psi \sim \chi^2(n-1)$, ψ и $A\bar{\eta}$ независ □

Теорема 1.4 (О проекции)**.** Пусть $\vec{\xi} \sim N(\vec{0}, \sigma^2 E)$, $L_1 : dim L_1 = m_1$ и $L_2 : dim L_2 = m_2$ два ортогональных подпространства $\mathbb{R}^n$, $\vec{\eta_1}$ - проекция $\vec{\xi}$ на L_1, норм. распр., независ. и $\frac{|\eta_1|^2}{\sigma^2} \sim \chi^2(dim L_1)$, $\frac{|\eta_2|^2}{\sigma^2} \sim \chi^2(dim L_2)$

Доказательство. $\vec{\eta_1} = A_1\vec{\xi} \sim N(\dots, \dots)$, $\vec{\zeta} = C\vec{\xi}$, C - ортогональная. $\vec{\zeta} \sim N(\vec{0}, C\sigma^2 E C^T) = N(\vec{0}, \sigma^2 E)$. Новый ортонормированный базис $e'_1 \dots e'_m$ в L_1, $e'_{m+1} \dots e'_n$ в L_2, $\vec{\eta_1} = \zeta_1 e'_1 + \dots + \zeta_m e'_m$, $\vec{\eta_2} = \zeta_{m+1} e'_{m+1} + \dots + \zeta_n e'_n$

$$\frac{\bar{\xi}}{\sigma} \sim N(\vec{0}, E) \qquad \frac{|\eta_1|^2}{\sigma^2} = \sum \frac{\xi_i^2}{\sigma^2} \sim \chi^2(m_1)$$

□

2 Порядковые случайные величины

$\xi_1, \xi_2, \dots, \xi_n$ назависс, $\xi_i \sim F(x)$

$$\eta = min(\xi_1, \dots \xi_n) \sim? \qquad \zeta = max(\xi_1, \dots \xi_n) \sim?$$

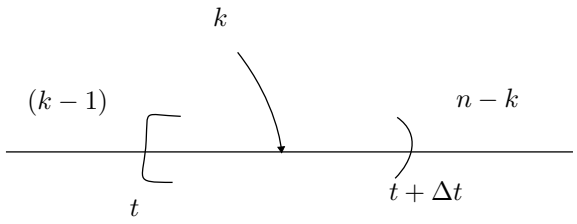
$$\Phi(y) = P(\eta < y) = 1 - P(\eta \geq y) = 1 - P(min(\xi_1, \dots, \xi_n) \geq y) =$$
$$= 1 - P(\xi_1 \geq y, \dots, \xi_n \geq y) = 1 - \prod_{i=1}^{n} P(\xi_i \geq y) =$$
$$= 1 - \prod_{i=1}^{n} (1 - P(\xi_i < y)) = 1 - (1 - F(y))^n$$

$$\Psi(z) = P(\zeta < z) = P(max(\xi_1, \dots, \xi_n) < z) =$$
$$= P(\xi_1 < z, \dots, \xi_n < z) = \prod_{i=1}^{n} P(\xi_i < z) = (F(z))^n$$

Порядковые величины:

$$\xi_{(1)} = min(\xi_1, \dots, \xi_n)$$
$$\xi_{(2)} = min(\xi_i : \xi_i \neq \xi_{(1)})$$
$$\xi_{(3)} = min(\xi_i : \xi_i \neq \xi_{(1)}, \xi_i \neq \xi_{(2)})$$
$$\dots$$
$$\xi_{(n)} = max(\xi_1, \dots, \xi_n)$$

Положим $F(x)$ - непрерывна:



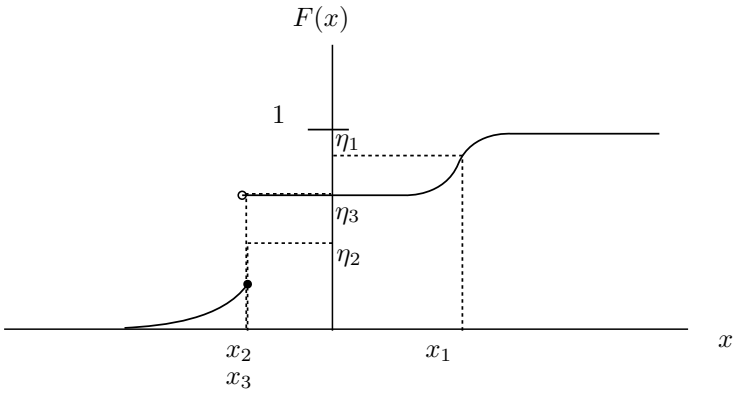
$$P(t \le \xi_{(k)} < t + \Delta t) =$$
$$= nP(t \le \xi < t + \Delta t)C_{n-1}^{k-1}(P(\xi < t))^{k-1}(P(\xi \ge t + \Delta t))^{n-k}C_{n-k}^{n-k} =$$
$$= \varkappa(t + \Delta t) - \varkappa(t)$$
$$n\frac{F(t + \Delta t) - F(t)}{\Delta t}C_{n-1}^{k-1}(F(t))^{k-1}(1 - F(t + \Delta t))^{n-k} = \frac{\varkappa(t + \Delta t) - \varkappa(t)}{\Delta t}$$
$$\varkappa(t) = np(t)C_{n-1}^{k-1}(1 - F(t))^{n-k}(F(t))^{k-1}$$

$\varkappa(t)$ - плотность распределения $\xi_{(k)}$, $p(t) = F'(t)$

$\xi_{(1)}$ и $\xi_{(n)}$ совместное распр

$$G(y, z) = P(\xi_{(1)} < y, \xi_{(n)} < z)$$
$$P(\xi_{(n)} < z) = P(\xi_{(1)} < y, \xi_{(n)} < z) + P(\xi_{(1)} \ge y, \xi_{(n)} < z)$$
$$\Psi(z) = (F(z))^n = P(\xi_{(1)} < y, \xi_{(n)} < z) + \prod_{i=1}^{n}\underbrace{P(y \le \xi_i < z)}_{F(z)-F(y)}$$
$$G(y, z) = P(\xi_{(1)} < y, \xi_{(n)} < z) = \begin{cases}(F(z))^n, & y > z\\ (F(z))^n - (F(z) - F(y))^n, & y \le z\end{cases}$$

3 Моделирование случайных величин



$\xi \sim F(x)$, $\eta \sim R(0,1)$, $F(x) = \eta_1 \to x_1$, для псевдослучайных чисел вихрь Мерсенна

4 Основные задачи статистики

Явление $\to$ математическая модель явления $\to$ вероятностная модель явления ξ_i.

Выборка - n наблюдений над явлением $\to$описательная статистика (непараметрическая), выбор классов (e.g. $\varepsilon \sim N(a, \sigma^2)$) $\to$ параметрическая статистика (e.g. $\varepsilon \sim N(a, \sigma^2)$, $a-?$, $b-?$)

Пример. Пытаемся понять как остывает чашка чая.

$$\frac{dT}{dt} = k(T - T_0) + \varepsilon$$

Вероятностная модель явления с двумя случайными величинами (погрешности измерений ε и внутри k).

Распределения полагаем нормальными и так далее.

1. Определение параметров и оценка их точности
2. Проверка статистических гипотиз

Характеристики модели θ, по выборке оценка $\bar{\theta}(\vec{x}_n)$, n - объём выборки.

Статистика - $\forall$ барелевская функция от $\vec{x}_n$ (борелевская $g : \mathbb{R} \to \mathbb{R}$ $\forall B \in \mathbb{B} \hookrightarrow g^{-1}(B) \in \mathbb{B}$).

Воспринимаем $\vec{x}_n$ с двух сторон:

1. конкретные наблюдения над явлением

2. независимые случайные величны с распределением, одинаковым со случайными величинами в вероятностной модели

4.1 Свойства оценок

Θ - множество значений θ, $\theta \in \Theta$, $\tilde{\theta}(\vec{x}_n)$ - оценка θ по выборке

1. несмещённость $\forall\theta \in \Theta \hookrightarrow M[\tilde{\theta}(\vec{x}_n)] = \theta$
2. состоятельность $\forall\theta \in \Theta \hookrightarrow \tilde{\theta} \xrightarrow{P} \theta$ (i.e. $\forall\varepsilon > 0\ P(|\tilde{\theta} - \theta| \geq \varepsilon) \underset{n\to\infty}{\to} 0$)
3. сравнение оценок $\tilde{\theta}_1$ эффективнее $\tilde{\theta}_2$, если $\forall\theta \in \Theta \hookrightarrow D[\tilde{\theta}_1] \leq D[\tilde{\theta}_2]$ и $\exists\theta \in \Theta :\ D[\tilde{\theta}_1] < D[\tilde{\theta}_2]$

Теорема 4.1 (Достаточное условие состоятельности)**.** Если $\tilde{\theta}$ является несмещённой оценкой θ и $D[\tilde{\theta}] \underset{n\to\infty}{\to} 0$, то $\tilde{\theta}$ является состоятельной оценкой ($\forall\theta \in \Theta$)

Доказательство. $M[\tilde{\theta}] = \theta$, по неравенству Чебышёва:

$$\forall\varepsilon > 0 \hookrightarrow P(|\tilde{\theta} - \underbrace{M[\tilde{\theta}]}_{\theta}| < \varepsilon) \geq 1 - \frac{D[\tilde{\theta}]}{\varepsilon^2} \underset{n\to\infty}{\to} 1$$

□

Задача (T1)**.** Пытаемся понять по двум серийным номерам сколько всего танков.

$\xi \sim R(0, \theta)$, $\theta > 0$ вер. модель., $\vec{x}_n$ - выбрка объёмом n

$$\tilde{\theta}_1 = 2\bar{x} = 2\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\tilde{\theta}_2 = \min x_i$$

$$\tilde{\theta}_3 = \max x_i$$

$$\tilde{\theta}_4 = x_1 + \frac{1}{(n-1)}\sum_{i=2}^{n} x_i$$

$$M[\xi] = \int_{-\infty}^{\infty} xp(x)dx = \int_0^{\theta} \frac{x}{\theta}dx = \frac{\theta}{2}$$

$$M[\xi^2] = \int_{-\infty}^{\infty} x^2p(x)dx = \int_0^{\theta} \frac{x^2}{\theta}dx = \frac{\theta^2}{3}$$

Рассматриваем $\tilde{\theta}_1$:

Несмещённость $\forall\theta > 0 M[\tilde{\theta}] = \theta$:

$M[\frac{\theta}{n}\sum_{i=1}^{n} x_i] = \frac{2}{n}\sum_{i=1}^{n} M[x_i] = 2M[\xi] = \theta$ несмещённая

$D[\tilde{\theta}_1] = D[\frac{2}{n}\sum x_i] = \frac{1}{n^2}\sum D[x_i] = \frac{4}{n}D[\xi] = \frac{\theta^2}{3n} \underset{n\to\infty}{\to} 0$, по достаточному условия оценка состоятельная

Рассматриваем $\tilde{\theta}_2$:

$$M[\tilde{\theta}_2] = \int_{-\infty}^{\infty} y\varphi(y)dy$$

$$\Phi(y) = 1 - (1 - F(y))^n \qquad \varphi(y) = n(1 - F(y))^{n-1}p(y)$$

$$M[\tilde{\theta}_2] = \int_0^{\theta} n(1 - \frac{y}{\theta})^{n-1}\frac{1}{\theta}ydy =$$

$$t = 1 - \frac{y}{\theta}$$

$$= -\int_1^0 nt^{n-1}(1-t)\theta dt = \int_0^1 n\theta t^{n-1}dt - \int_0^1 n\theta t^n dt =$$

$$= n\theta[1 - \frac{n}{n+1}] = \frac{\theta}{n+1} \quad \text{смещённая}$$

$$\tilde{\theta}_2' = (n+1)x_{min} = (n+1)\tilde{\theta}_2 \quad \text{несмещённая}$$

$$M[\tilde{\theta}_2^2] = \int_0^{\theta} n(1 - \frac{y}{\theta})^{n-1}\frac{1}{\theta}y^2dy =$$

$$= -\int_1^0 nt^{n-1}(1-t)^2\theta^2 dt = \frac{2\theta^2}{(n+1)(n+2)}$$

$$D[\tilde{\theta}_2] = \frac{2\theta^2}{(n+1)(n+2)} - \frac{\theta^2}{(n+1)^2} = \theta^2\left[\frac{2(n+1)-(n+2)}{(n+1)^2(n+2)}\right] =$$

$$= \theta^2\left[\frac{n}{(n+1)^2(n+2)}\right] \underset{n\to\infty}{\to} 0 \quad \text{достаточное не выполнятеся}$$

$$D[\tilde{\theta}_2'] = (n+1)^2 D[\tilde{\theta}_2] = \frac{\theta^2 n}{n+2} \not\to 0$$

Смотрим состоятельность $\tilde{\theta}_2'$ по определению

$$\forall\theta > 0\ \forall\varepsilon > 0 \hookrightarrow P(|\tilde{\theta}_2' - \theta| \geq \varepsilon) \underset{n\to\infty}{\to} 0$$

$$P(|\tilde{\theta}_2' - \theta| \geq \varepsilon) \geq P(\tilde{\theta}_2' > \theta + \varepsilon) =$$

$$= P((n+1)x_{min} \geq \theta + \varepsilon) = P(x_{min} \geq \frac{\theta+\varepsilon}{n+1}) =$$

$$= 1 - P(x_{min} < \frac{\theta+\varepsilon}{n+1}) = 1 - (1 - (1 - F(\frac{\theta+\varepsilon}{n+1}))^n) =$$

$$= (1 - (\frac{\theta+\varepsilon}{\theta(n+1)}))^n \underset{n\to\infty}{\to} e^{-\frac{\theta+\varepsilon}{\theta}} > 0$$

Не является состоятельной!

Смотрим состоятельность $\tilde{\theta}_2$ по определению:

$$P(\tilde{\theta}_2 < \theta - \varepsilon) + \underbrace{P(\tilde{\theta}_2 > \theta + \varepsilon)}_{=0,\text{т.к.}\tilde{\theta}_2 = x_{min}}$$

$$P(x_{min} < \theta - \varepsilon = \Phi(\theta - \varepsilon)) = 1 - (1 - \frac{\theta - \varepsilon}{\theta})^n = 1 - \left(\frac{\varepsilon}{\theta}\right)^n \underset{n\to\infty}{\to} 1$$

Не является состоятельной!

Рассматриваем $\tilde{\theta}_3 = x_{max}$:

$$M[\tilde{\theta}_3] = \int_{-\infty}^{+\infty} z\psi(z)dz = \int_0^\theta n\frac{z^n}{\theta^n}dz = \frac{n}{n+1}\theta \qquad \text{смешённая}$$

$$\Psi(z) = (F(z))^n \qquad \psi(z) = n(F(z))^{n-1}p(z) = n\left(\frac{z}{\theta}\right)^{n-1}\frac{1}{\theta}\{(0;\theta)\}$$

$$D[\tilde{\theta}_3]\frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \frac{n\theta^2}{(n+2)(n+1)^2}$$

$$D[\tilde{\theta}_3']\frac{(n+1)^2}{n^2}D[\tilde{\theta}_3] = \frac{\theta^2}{n(n+2)} \underset{n\to\infty}{\to} 0 \qquad \text{состоятельная}$$

Смотрим состоятельность $\tilde{\theta}_2'$ по определению

$$\forall\theta > 0\ \forall\varepsilon > 0$$

$$P(|\tilde{\theta}_2' - \theta| \geq \varepsilon) = P(x_{max} < \theta - \varepsilon) + \underbrace{P(x_{max} \geq \theta + \varepsilon)}_{=0} =$$

$$= (F(\theta - \varepsilon))^n = \begin{cases} 0 < \varepsilon < \theta: \ \left(\dfrac{\theta - \varepsilon}{\theta}\right)^n \underset{n\to\infty}{to} 0 \\ \varepsilon \geq \theta: \ (0)^n \underset{n\to\infty}{\to} 0 \end{cases}$$

Является состоятельной!

Рассматриваем $\tilde{\Theta}_4$:

$$M[\tilde{\theta}_4] = M[x_1 + \sum_{i=2}^n x_i] = M[x_1] + \frac{1}{n-1}\sum_{i=1}^n M[x_i] = \frac{\theta}{2} + \frac{\theta}{2} = \theta$$

$$D[\tilde{\theta}_4] = D[\xi] + \frac{1}{(n-1)^2}(n-1)D[\xi] = \frac{\theta^2}{12}\frac{n}{n-1} \underset{n\to\infty}{\not\to} 0$$

Достаточное усл. не работает.

Используем теорему $\xi_n \xrightarrow{p} \xi$, $\eta_n \xrightarrow{p} \eta$, $\xi_n + \eta_n \xrightarrow{p} \xi + \eta$.

И ЗБЧ Хинчина: $\xi_1, \dots, \xi_n$ незав., одинак распр. $\Rightarrow \frac{1}{n}\sum_{i=1}^{n}\xi_i \xrightarrow{p} M[\xi]$.

$$x_1 \xrightarrow{p} x_1 \qquad \frac{1}{n-1}\sum_{i=2}^{n} x_i \xrightarrow{p} \frac{\theta}{2}$$

$$\tilde{\theta}_4 \xrightarrow{p} x_1 + \frac{\theta}{2}$$

Не состоятельна!

Адекватные остались $\tilde{\theta}_1 = 2\bar{x}$, $\tilde{\theta}_3' = \frac{n+1}{n}x_{max}$

$$D[\tilde{\theta}_1] = \frac{\theta^2}{3n} > D[\tilde{\theta}_3'] = \frac{\theta^2}{n(n+2)}$$

Лучшая оценка $\tilde{\theta}_3$.

---

5 Оптимальность и эффективность оценок

Определение 5.1. Несмещённая оценка $\tilde{\theta}(\vec{x}_n)$ характеристики θ называется оптимальной $\tilde{\theta}_{opt}$ если для $\forall\theta \in \Theta \Rightarrow D[\tilde{\theta}_{opt}] = \inf D[\tilde{\theta}]$, inf по всем несмещённым оценкам θ.

Теорема 5.1 (Единственность оптимальной оценки)**.** Если оптимальная оценка существует, то она единственна.

Доказательство. Пусть $\tilde{\theta}_1$ и $\tilde{\theta}_2$ разные оптимальные оценки

$$\tilde{\theta}_3 = \frac{\tilde{\theta}_1 + \tilde{\theta}_2}{2} \qquad M[\tilde{\theta}_3] = \theta$$

$$D[\tilde{\theta}_3] = \frac{1}{4}D[\tilde{\theta}_1] + D[\tilde{\theta}_2] + \frac{1}{2}cov(\tilde{\theta}_1, \tilde{\theta}_2) = \frac{1}{2}D[\tilde{\theta}_1] + \frac{1}{2}cos(\tilde{\theta}_1, \tilde{\theta}_2)$$

$$D[a\xi + b\eta] = a^2 D[\xi] + b^2 D[\eta] + 2abcov(\xi, \eta)$$

$$|cov(\tilde{\theta}_1, \tilde{\theta}_2)| \le \sqrt{D\tilde{\theta}_1 D\tilde{\theta}_2} = D[\tilde{\theta}_1]$$

$$D[\tilde{\theta}_3] \le D[\tilde{\theta}_1] \qquad D[\tilde{\theta}_3] = D[\tilde{\theta}_1]$$

$$cov(\tilde{\theta}_1, \tilde{\theta}_2) = D[\tilde{\theta}_1] \quad \Rightarrow \quad r = 1 \Leftrightarrow \tilde{\theta}_1 = a\tilde{\theta}_2 + b$$

$$M[\tilde{\theta}_1] = M[\tilde{\theta}_2] = \theta \qquad D[\tilde{\theta}_1] = D[\tilde{\theta}_2]$$

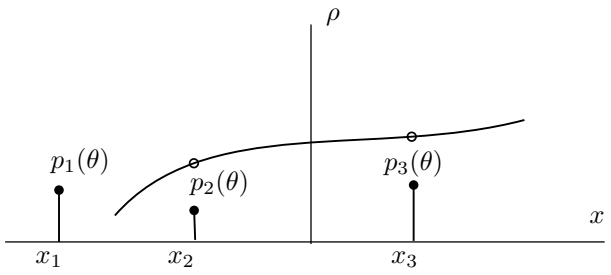
$$\begin{cases} \theta = a\theta + b \\ a^2 D[\tilde{\theta}_2] = D[\tilde{\theta}_2] \end{cases} \Rightarrow \begin{cases} a = 1 \\ b = 0 \end{cases}$$

$$\Rightarrow \tilde{\theta}_1 = \tilde{\theta}_2$$

Противоречие. □

Будем рассматриватть параметрические вероятностные модели:

$$\xi \sim \rho(x,\theta),\ \theta \in \Theta \subset \mathbb{R},\ x \in A(\theta)$$
$$\xi \sim \rho(x,\vec{\theta}),\ \vec{\theta} \in \Theta \subset \mathbb{R}^m,\ x \in A(\vec{\theta})$$
$$\rho(x,\theta) = \underbrace{p(x,\theta)\{E\}}_{\text{непр. часть}} + \underbrace{\sum_k p_k(\theta)\{x_k\}}_{\text{дискр. часть}}$$



5.1 Информация Фишера

$$I(\theta) = M\left[(\frac{\partial \ln \rho(x,\theta)}{\partial \theta})^2\right] =$$
$$= \int_E \left(\frac{\partial \ln p(x,\theta)}{\partial \theta}\right)^2 p(x,\theta)dx + \sum_k \left(\frac{\partial \ln p_k(x,\theta)}{\partial \theta}\right)^2 p_k(\theta)$$

$I(\vec{\theta})$ - информационная матрица Фишера

$$I_{ij}(\vec{\theta}) = M\left[\frac{\partial \ln p(x,\theta)}{\partial \theta_i}\frac{\partial \ln p(x,\theta)}{\partial \theta_j}\right]$$

Определение 5.2. Вероятностьная модель $\xi \sim \zeta(x,\theta),\ \theta \in \Theta \subset \mathbb{R},\ x \in A$ называется регулярной, если

1. $\rho(x,\theta)$ непр дифф по θ на Θ
2. $\frac{\partial}{\partial\theta}\int_A \rho(x,\theta)dx = \int_A \frac{\partial}{\partial\theta}\rho(x,\theta)dx$ на Θ
3. $I(\theta)$ непр на Θ и $I(\theta) > 0$ на Θ

Определение 5.3. Вероятностьная модель $\xi \sim \zeta(x, \vec{\theta})$, $\vec{\theta} \in \Theta \subset \mathbb{R}^m$, $x \in A$ называется регулярной, если

1. $\rho(x, \vec{\theta})$ непр дифф по $\vec{\theta}$ на Θ
2. $\frac{\partial}{\partial \theta_i} \int_A \rho(x, \vec{\theta}) dx = \int_A \frac{\partial}{\partial \theta_i} \rho(x, \vec{\theta}) dx$ на Θ, $i = 1, \ldots, m$
3. $I(\vec{\theta})$ положительно определена на Θ и $I_{ij}(\vec{\theta})$ непрер. на Θ

Определение 5.4. Вероятностная модель $\xi \sim \rho(x, \theta)$, $\vec{\theta} \in \Theta \subset \mathbb{R}^m$, $x \in A$ называется сильно регулярной, если эта модель регулярна и

1. $\rho(x, \theta)$ k раз непрерывно дифф по θ на Θ ($k \geq 2$)
2. $\frac{\partial^l}{\partial \theta^l} \int_A \rho(x, \theta) dx = \int_A \frac{\partial^l}{\partial \theta^l} \rho(x, \theta) dx$, $l = 1, \ldots, k$

Определение 5.5. Вероятностная модель $\xi \sim \zeta(x, \vec{\theta})$, $\vec{\theta} \in \Theta \subset \mathbb{R}^m$, $x \in A$ называется сильно регулярной, если эта модель регулярна и

1. $\rho(x, \vec{\theta})$ k раз непрерывно дифф по θ на Θ ($k \geq 2$)
2. все производные по $\vec{\theta}$ перестановочные с $\int$ по x

Определение 5.6. Статистика $\tilde{g}(\vec{x}_n)$ называется регулярной оценкой функции $g(\theta)$, если она является несмещённой оценкой и

$$\frac{\partial}{\partial \theta} \int_B \tilde{g}(\vec{x}_n) L(\vec{x}_n, \theta) d\vec{x}_n = \int_B \tilde{g}(\vec{x}_n) \frac{\partial}{\partial \theta} L(\vec{x}_n, \theta) d\vec{x}_n$$

где $L(\vec{x}_n, \theta)$ - плотность распределения случайного вектора $\vec{x}_n$ ($L(\vec{x}_n, \theta) = \prod_{i=1}^n \rho(x_i, \theta)$), $B = \underbrace{A \times A \times \cdots \times A}_{n}$

Теорема 5.2 (Достаточное условие регулярности оценки)**.** Если модель регулярна, $\tilde{g}(\vec{x}_n)$ является несмещ. оценкой $g(\theta)$ и $D[\tilde{g}(\vec{x}_n)]$ ограничена на $\forall$ компакте из Θ по θ, тогда оценка регулярна.

5.2 Неравенство Крамера-Рао

Теорема 5.3. Пусть модель является регулярной, $\tilde{g}(\vec{x}_n)$ является регулярной оценкой оценкой дифф функции $g(\theta)$. Тогда

$$\forall \theta \in \Theta \hookrightarrow D[\tilde{g}] \geq \frac{(g')^2(\theta)}{nI(\theta)}$$

Доказательство. $\xi \sim \rho(x, \theta)$, $\theta \in \Theta \subset \mathbb{R}$, $x \in A(\theta)$, $\vec{x}_n$ независ. выборка

$L(\vec{x}_n, \theta) = \prod_{i=1}^{n} \rho(x_i, \theta)$ - распр. выборки $\vec{x}_n$, $B = A \times \cdots \times A$

$$\frac{\partial}{\partial\theta} \int\limits_B \cdots \int L(\vec{x}, \theta) d\vec{x} = \frac{\partial}{\partial\theta} 1 = 0$$

в силу регулярности модели

$$\int\limits_B \cdots \int \frac{\partial}{\partial\theta} L d\vec{x} = 0$$

Домножаем и делим на L, там где $L = 0$ считаем что интеграл 0

$$\int_B \frac{\partial \ln L}{\partial\theta} L d\vec{x} = 0$$

$$M[\tilde{g}] = g(\theta)$$

$$\int_B \tilde{g}(\vec{x}_n) L(\vec{x}_n, \theta) d\vec{x}_n = g(\theta)$$

$$\frac{\partial}{\partial\theta} \int_B \tilde{g}(\vec{x}_n) L(\vec{x}_n, \theta) d\vec{x}_n = \frac{\partial}{\partial\theta} g(\theta)$$

$$\int_B \tilde{g}(\vec{x}_n) \frac{\partial}{\partial\theta} L d\vec{x}_n = g'(\theta)$$

$$\int_B \tilde{g}(\vec{x}_n) \frac{\partial \ln L}{\partial\theta} L d\vec{x}_n = g'(\theta)$$

$$\int_B (\tilde{g}(\vec{x}_n) - g(\theta)) \frac{\partial \ln L}{\partial\theta} L d\vec{x}_n = g'(\theta)$$

$\eta = \tilde{g}(\vec{x}_n) - g(\theta)$ - сл. вел

$\zeta = \frac{\partial \ln L(\vec{x}_n,\theta)}{\partial\theta}$ - сл. вел.

$$M[\eta] = 0 \qquad M[\zeta] = 0$$
$$M[\eta\zeta] = g'(\theta)$$
$$cov(\eta,\zeta) = M[\eta\zeta] - M[\eta]M[\zeta] = g'(\theta)$$
$$r = \frac{cov(\eta,\zeta)}{\sqrt{D\eta D\zeta}} \qquad |r| \le 1$$
$$\frac{cov^2(\zeta,\eta)}{D\zeta D\eta} \le 1$$
$$g'^2(\theta) \le D\zeta \underbrace{D[\tilde{g}]}_{D[\tilde{g}]}$$
$$D\zeta = M[\zeta^2] - M^2[\zeta] = M[\zeta^2]$$
$$D\zeta = D\left[\frac{\partial \ln L}{\partial\theta}\right] = D\left[\sum_{i=1}^{n}\frac{\partial \ln \rho(x_i,\theta)}{\partial\theta}\right] = \sum_{i=1}^{n} D\left[\frac{\partial \ln \rho(x_i,\theta)}{\partial\theta}\right] =$$
$$= nD\left[\frac{\partial \ln \rho(x_i,\theta)}{\partial\theta}\right] = nM\left[\left(\frac{\partial \ln \rho}{\partial\theta}\right)^2\right] - n\underbrace{M^2\left[\frac{\partial \ln \rho}{\partial\theta}\right]}_{0} = nI(\theta)$$

□

Следствие 5.1.

1. оценка параметра θ, $g(t) = \theta$,
$$D[\tilde{\theta}] \ge \frac{1}{nI(\theta)}$$
2. многомерный аналог нер. Крамера-Рао
$$D[\tilde{g}(\vec{x}_n)] \ge \frac{1}{n}\nabla^T g(\vec{\theta}) I^{-1}(\vec{\theta})\nabla g(\vec{\theta})$$

Определение 5.7 (Эффективная оценка)**.** Регулярная оценка $\tilde{g}(\vec{x}_n)$ функции $g(\theta)$ называется эффективной ($\tilde{g}_{eff}$), если $\forall\theta \in \Theta \hookrightarrow D[\tilde{g}_{eff}] = \inf D[\tilde{g}]$, inf берётся по всем регулярным оценкам.

Теорема 5.4. Если эффективная оценка $\exists$, то она единственна.

Доказательство. Так же как и оптимальная только нужно доказать что $\tilde{\theta}_3 = \frac{\tilde{\theta}_1+\tilde{\theta}_2}{2}$ - регулярная. □

Теорема 5.5 (Достаточное условие эффективности)**.** Пусть выполнены условия нер. Крамера-Рао и $D[\tilde{g}] = \frac{g'^2}{nI(\theta)}$, тогда $\tilde{g}$ эффективная оценка $g(\theta)$.

Теорема 5.6 (Теорема о частоте)**.** Частота появления события A в n независимых опытах является несмещённой, состоятельной и эффективной оценкой вероятности появления этого события.

Доказательство. (на примере)

$$\xi \sim \rho(x,\theta) = \theta\{1\} + (1-\theta)\{0\} \qquad \theta \in (0,1)$$
$$\xi \sim Bi(1,\theta) \qquad \nu = \frac{m}{n}$$
$$\vec{x}_n = (0,0,1,\dots) \qquad \nu = \frac{1}{n}\sum x_i = \bar{x} \qquad \tilde{\theta} = \bar{x}$$

1. несмещённость ($\xi \sim Bi(l,\theta)$, $M[\xi] = l\theta$, $D[\xi] = l\theta(1-\theta)$)

$$M[\bar{x}] = \frac{1}{n}M[\sum x_i] = M[\xi]$$

2. состоятельность

$$D[\tilde{\theta}] = D[\frac{1}{n}\sum x_i] = \frac{1}{n^2}nD[\xi] = \frac{1}{n}\theta(1-\theta) \underset{n\to\infty}{\to} 0$$

состоятельна по достаточному условию

3. эффективность, модель регулярна

$$I(\theta) = \frac{l}{\theta(1-\theta)}\Big|_{l=1} = \frac{1}{\theta(1-\theta)}$$

$\tilde{\theta} = \bar{x}$ - регулярная оценка?

$D[\tilde{\theta}] = \frac{1}{n}\theta(1-\theta)$ огран на $\forall$ компакте из $(0,1)$

Является регулярной ✓

Неравенство Крамера-Рао:

$$D[\tilde{\theta}] = \frac{1}{n}\theta(1-\theta) \geq \frac{1}{nI(\theta)} = \frac{\theta(1-\theta)}{n}$$

достигает нижней грани $\Rightarrow$ эффективная (в данном случае ещё и оптимальная)

□

6 Описательная стат. (непараметр. стат.)

$\vec{x}_n$ - выборка

Вероятностная модель - все распределения, кроме сингулярных и вырожденных.

1. Вариационный ряд - упорядоченная ваборка

$$x_{min} = x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)} = x_{max}$$

   $x_{(k)}$ - k-ая порядковая сл. вел

2. Размах выборки $l = x_{max} - x_{min}$

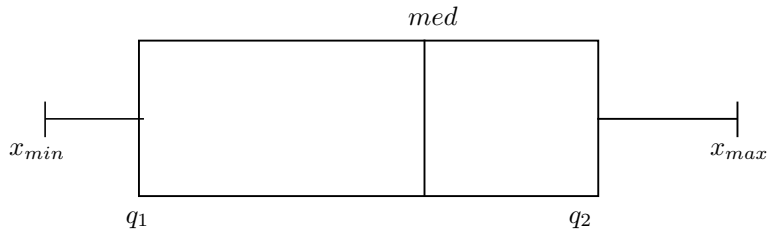
3. Медиана выборки med

$$mes = \begin{cases} x_{(k+1)}, \mu n = 2k+1 \\ \dfrac{x_{(k+1)} + x_{(k)}}{2},\ n = 2k \end{cases}$$

4. Мода - эл. выборки, который встречается чаще всего

5. Квартили q_1, q_2 (медианы половинок)

6. Boxplot



   $\varepsilon = q_2 - q_1$, если $x_{min} < q_1 - 1.5\varepsilon$ или $x_{max} > q_2 + 1.5\varepsilon$ рисуем усики до $q_1 - 1.5\varepsilon$ или $q_2 + 1.5\varepsilon$ соответственно, а дальше выбросы обозначаем точками для каждого значения

7. эмпирическая функция распределения

$$\tilde{F}(x) = \frac{m(x)}{n}$$

где $m(x)$ число элементов выборки, которые $< x$.

$$F(x) = P(\underbrace{\xi < x}_{A})$$

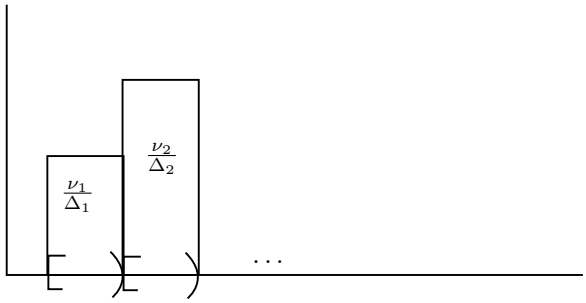
$\tilde{F}$ является несмещённой, состоятельной и эффективной оценкой $F(x)$ (по Т. о частоте).

8. Гистограмма

   Статистический ряд

$$\underbrace{[y_1, y_2)}_{\nu_1 = \frac{m_1}{n}}, \underbrace{[y_2, y_3)}_{\nu_2} \cdots \underbrace{[y_k, y_{k+1})}_{\nu_k}$$

   Эмперически $k = 1 + \log_2 n$



   $\nu_m = P(y_m \le \xi < y_{m+1}) = \int_{y_m}^{y_{m+1}} p(x)dx = p(\bar{x})\Delta_m$

9. числовые характеристики

   $\alpha_k = M[\xi^k]$ момент к-го порядка

$$\tilde{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \qquad \tilde{\alpha}_1 = \bar{x}$$

- несмещ: $M[\alpha_k] = \frac{1}{n} M[\sum x_i^k] = M[\xi^k] = \alpha_k$
- состоятельность: ЗБЧ Хинчина $\tilde{\alpha}_k \xrightarrow{p} \alpha_k = M[\xi^k]$

$\mu_k = M[(\xi - M\xi)^K]$ - центральный момент k-го порядка

$$\tilde{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

- состоятельность

$$\mu_k = M[\sum_{m=0}^{k} C_k^m \xi^m (-1)^{k-m} (M\xi)^{k-m}] =$$
$$= \sum_{m=0}^{k} C_k^m (-1)^{k-m} \alpha_m (\alpha_1)^{k-m}$$
$$\tilde{\mu}_k = \sum_{m=0}^{k} C_k^m (-1)^{k-m} \tilde{\alpha}_m (\tilde{\alpha}_1)^{k-m}$$

Теорема наследования сходимости:

$\xi_n \xrightarrow{p} \xi$, $f(x)$ непр на $\mathbb{R} \Rightarrow f(\xi_n) \xrightarrow{p} f(\xi)$

$\xi_n \xrightarrow{p} C$, $f(x)$ непр в точке $x = C \Rightarrow f(\xi_n) \xrightarrow{p} f(C)$

$\tilde{\alpha}_k \xrightarrow{p} \alpha_k$, $f(x_1, \ldots, x_n)$ непр $\Rightarrow \tilde{\mu}_k \xrightarrow{p} \mu_k$

- несмещённость

$$\mu_2 = D\xi \qquad \tilde{\mu}_2 = \tilde{\alpha}_2 - (\tilde{\alpha})^2$$
$$M[\tilde{\mu}_2] = M\left[\frac{1}{n}\sum x_i^2\right] - M\left[(\frac{1}{n}\sum x_i)^2\right] =$$
$$= M\xi^2 - (D[\bar{x}] + (M[\bar{x}])^2) = M\xi^2 - \frac{1}{n^2} n D\xi - (M\xi)^2 =$$
$$= D\xi\left[1 - \frac{1}{n}\right] = \mu_2 \frac{n-1}{n}$$

$S^2 = \frac{n}{n-1}\tilde{\mu}_2$, $M[S^2] = \mu_2$ несмещ оценка дисперсии

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

коэффициент асимметрии

$$\gamma = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}$$
$$\tilde{\gamma} = \frac{\tilde{\mu}_3}{\tilde{\mu}_2^{3/2}} \xrightarrow{p} \gamma$$

10. оценка распределения статистики

$$\vec{x}_n \qquad \tilde{g}(\vec{x}_n)$$

$\vec{x}_n$ становится вероятностной моделью, вытаскиваем 1000 подвыбоок с повторением элементов того же объёма $\vec{x}_n^*$, $\vec{x}_n^* \to \tilde{g}_i^*(\vec{x}_n^*)$, строим гистограмму из $\tilde{g}_1^* \dots \tilde{g}_{1000}^*$

7 Методы нахожд. параметров модели

$$\xi \sim \rho(x, \vec{\theta}),\ \vec{\theta} \in \Theta \subset \mathbb{R}^m,\ x \in A$$

Парметрическая модель, $\vec{x}_n$ выборка

1. Метод моментов (Пирсон)

   $$\alpha_k(\vec{\theta}) = M[\xi^k] \to \tilde{\alpha}_k = \alpha_k(\vec{\theta})$$

   Если можем решить систему получаем $\tilde{\vec{\theta}}$ оценку методом моментов (ОММ)

   Замечание. ОММГ (ОММ Групированная) для статистического ряда:

   $$\underbrace{[y_1, y_2)}_{\nu_1} \dots \underbrace{[y_k, y_{k+1})}_{\nu_k} \qquad \nu_i = \frac{m_i}{n}$$
   $$z_1 = \frac{y_1 + y_2}{2}, \dots, z_k = \frac{y_k + y_{k+1}}{2}$$
   $$\tilde{\alpha}_l = \sum_{i=1}^{k} z_i^l \nu_i$$

2. Оценка Методом Правдоподобия (Фишер):

монета 10 раз, 6 орлов, 4 решки

(a) $p = \frac{1}{2}$

(b) $p = \frac{1}{3}$ - орёл, $p = \frac{2}{3}$ - решка

$$\left(\frac{1}{2}\right)^{10} = 0.00098 \qquad \left(\frac{1}{3}\right)^{6} \left(\frac{2}{3}\right)^{4} = 0.00027$$

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

$fix\, \theta$, $L(\vec{x}_n)$ - плотность распределения выборки

$fix\, \vec{x}_n$, $L(\theta) = \prod_{i=1}^{n} \rho(x_i, \theta)$ - функция правдоподобия (здесь x_i - элемент выборки), пытаемся максимизировать $L(\theta) \to \max$

Замечание. ОМПГ для статистического ряда

$$\underbrace{[y_1, y_2)}_{\nu_1} \dots \underbrace{[y_k, y_{k+1})}_{\nu_k} \qquad \nu_i = \frac{m_i}{n}$$

$$P_1(\theta) = \int_{-\infty}^{y_2} \rho(x,\theta)dx$$

$$P_2(\theta) = \int_{y_2}^{y_3} \rho(x,\theta)dx$$

$$\dots$$

$$P_k(\theta) = \int_{y_k}^{+\infty} \rho(x,\theta)dx$$

$$L(\theta) = P_1^{m_1} \dots P_k^{m_k} \qquad L(\theta) \to \max$$

8 Асимптотические свойства оценок

$n \to \infty$

1. состоятельность $\forall\theta \in \Theta \hookrightarrow \tilde{g}(\vec{x}_n) \xrightarrow{p} g(\theta)$
2. асимптотическая несмещённость $\forall\theta \in \Theta \hookrightarrow M[\tilde{g}(\vec{x}_n)] \underset{n\to\infty}{\to} g(\theta)$
3. асимптотическая эффективность $\forall\theta \in \Theta \hookrightarrow nD[\tilde{g}(\vec{x}_n)] \underset{n\to\infty}{\to} \frac{g'^2(\theta)}{I(\theta)}$
4. асимпотическая нормальность $\forall\theta \in \Theta \hookrightarrow \sqrt{n}(\tilde{g}(\vec{x}_n) - g(\theta)) \xrightarrow{F} \eta \sim N(0, \sigma^2)$

8.1 Асимптотические свойства ОММ

$$\alpha_k(\vec{\theta}) = \tilde{\alpha}_k \to \tilde{\vec{\theta}} = f(\tilde{\alpha}_1, \dots, \tilde{\alpha}_k)$$

$$\tilde{\alpha}_k \xrightarrow{p} \alpha_k$$

f - непр в $\alpha_1, \dots, \alpha_k$, тогла по теореме наследования $\tilde{\vec{\theta}} \xrightarrow{p} f(\alpha_1, \dots, \alpha_k) = \vec{\theta}$ состоятельная

Теорема о наследовании нормальности:

$\sqrt{n}(\xi_n - a) \rightsquigarrow N(0, \sigma^2)$ и $f(x) \in C'(r)$ и $f'(a) \neq 0$, тогда

$$\sqrt{n}(g(\xi_n) - g(a)) \rightsquigarrow N(0, g'^2(a)\sigma^2)$$

ЦПТ:

$$\frac{\sum x_i^k - nM[\xi^k]}{\sqrt{nD[\xi^k]}} \rightsquigarrow N(0,1)$$

$$\frac{\tilde{\alpha}_k - a_k}{\sqrt{\alpha_{2k} - a_k^2}}\sqrt{m} \rightsquigarrow N(0, \underbrace{\alpha_{2k} - a_k^2}_{\sigma_k^2})$$

$$(f(\tilde{\alpha}_k) - f(\alpha_j))\sqrt{n} \rightsquigarrow N(0, f'^2(\alpha_k)\sigma_k^2)$$

$$(\tilde{\theta} - \theta)\sqrt{n} \rightsquigarrow N(0, f'^2(\alpha_k)\sigma_k^2)$$

8.1.1 Многомерное ЦПТ

$$\alpha = \begin{pmatrix} \alpha_{s_1} \\ \vdots \\ \alpha_{s_k} \end{pmatrix} \qquad \tilde{\alpha} = \begin{pmatrix} \tilde{\alpha}_{s_1} \\ \vdots \\ \tilde{\alpha}_{s_k} \end{pmatrix} \qquad \tilde{\alpha}_{s_1} = \frac{1}{n}\sum_{i=1}^{n} x_i^{s_1}$$

$$(\tilde{\alpha} - \alpha)\sqrt{n} \rightsquigarrow N(\vec{0}, K) \qquad K_{ij} = \alpha_{s_i+s_j} - \alpha_{s_i}\alpha_{s_j}$$

$f(x) \in C'(R^k)$ и $\nabla f(\alpha) \neq 0$

$$(f(\tilde{\alpha}) - f(\alpha))\sqrt{n} \rightsquigarrow N(0, \nabla^T f(\alpha) K \nabla f(\alpha))$$

8.2 Асимптотические свойства ОМП

Теорема 8.1. Пусть вероятностная модель сильно регулярна и множество Θ открыто.

Тогда ОМП является состоятельной, асимп. несмещ., асимп. эффект. и асимп. нормальной.

Доказательство. $L(\theta) \to \max$, $\ln L \to \max$, $\frac{d\ln L(\tilde{\theta})}{d\theta} = 0$ Ряд Тёйлора с остаточным членом в форме Лагранжа:

$$\underbrace{\frac{d\ln L(\theta)}{d\theta}}_{0} = \frac{d\ln L}{d\theta}(\theta) + \frac{d^2\ln L}{d\theta^2}(\theta^*)(\tilde{\theta} - \theta)$$

$$\tilde{\theta} - \theta = -\frac{(\ln L)'(\theta)}{(\ln L)''(\theta^*)}$$

$$\frac{d\ln L}{d\theta} = \frac{d}{d\theta}(\ln \prod p(x_i, \theta)) = \sum_{i=1}^{n} \frac{d\ln p(x_i, \theta)}{d\theta}$$

ЗБЧ Хинчина: $\frac{1}{n}\sum \frac{d\ln p}{d\theta} \xrightarrow{p} M\left[\frac{d\ln p(\xi,\theta)}{d\theta}\right]$

$$\int_{-\infty}^{+\infty} p(x,\theta)dx = 1$$
$$\int_{-\infty}^{+\infty} \frac{d}{d\theta}p(x,\theta)dx = 1$$
$$\int_{-\infty}^{+\infty} \frac{d\ln p}{d\theta}pdx = 0$$
$$M\left[\frac{d\ln p}{d\theta}\right] = 0$$

$$\frac{d}{d\theta}\int_{-\infty}^{+\infty} \frac{d\ln p}{d\theta}pdx = 0$$
$$\int_{-\infty}^{+\infty} \frac{d^2\ln p}{d\theta^2}pdx + \int_{-\infty}^{+\infty} \frac{d\ln p}{d\theta}\frac{dp}{d\theta}dx = 0$$
$$M\left[\frac{d^2\ln p}{d\theta^2}\right] + \underbrace{M\left[\left(\frac{d\ln p}{d\theta}\right)^2\right]}_{I(\theta)>0} = 0$$
$$\frac{d^2\ln L}{d\theta^2} = \sum_{i=1}^{n}\frac{d^2\ln p(x_i,\theta^*)}{d\theta^2}$$
$$\frac{1}{n}\sum \frac{d\ln p(x_i,\theta^*)}{d\theta^2} \xrightarrow{p} -I(\theta^*) \neq 0$$

Состоятельность доказана: $\theta^* \xrightarrow{p} \theta$

$$\frac{\sum \frac{d\ln p}{d\theta} - \overbrace{nM[\frac{d\ln p}{d\theta}]}^{=0}}{\sqrt{nD[\frac{d\ln p}{d\theta}]}} \rightsquigarrow N(0,1)$$

Лемма Слуцкого: $\xi_n \xrightarrow{F} \xi$, $\eta_n \xrightarrow{p} C$, $x_n\eta_n \xrightarrow{F} \xi C$

$$\tilde{\theta} - \theta = -\frac{\frac{\sum \frac{d \ln p}{d\theta}}{\sqrt{nI(\theta)}}\sqrt{nI(\theta)}}{\frac{1}{-nI(\theta)}\sum \frac{d^2 \ln p}{d\theta^2}(-nI(\theta))}$$

$$a = \frac{\sum \frac{d \ln p}{d\theta}}{\sqrt{nI(\theta)}} \rightsquigarrow N(0,1) \qquad b = \frac{1}{-nI(\theta)}\sum \frac{d^2 \ln p}{d\theta^2} \rightsquigarrow 1$$

$$(\tilde{\theta} - \theta)\sqrt{nI(\theta)} = \frac{a}{b} \rightsquigarrow N(0,1)$$

$$(\tilde{\theta} - \theta)\sqrt{n} \rightsquigarrow N(0, \frac{1}{I(\theta)})$$

$$D[(\tilde{\theta} - \theta)\sqrt{n}] = nD[\tilde{\theta}] \underset{n\to\infty}{\to} \frac{1}{I(\theta)}$$

Асимптотическая эффективность. □

Следствие 8.1. $\tilde{\theta} \xrightarrow{p} \theta$ - сост. $\tilde{\theta}$ ОМП, $\theta \in \Theta \subset \mathbb{R}$

$$\sqrt{n}(\tilde{\theta} - \theta) \rightsquigarrow N(0, \frac{1}{I(\theta)})$$

- $g(\theta) \in C'(\Theta)$ и $g'(\theta) \neq 0$ на Θ

$$\sqrt{n}(g(\theta) - g(\theta)) \rightsquigarrow N(0, g'^2(\theta)\frac{1}{I(\theta)})$$

$$g(\tilde{\theta}) \xrightarrow{p} g(\theta)$$

  $g(\tilde{\theta})$ сост. оценка $g(\theta)$, асим. несмещ., асим. эффект., асим. норм.

- многомерный аналог

$$\sqrt{n}(\tilde{\vec{\theta}} - \vec{\theta}) \rightsquigarrow N(\vec{0}, I^{-1}(\vec{\theta}))$$

$$g(\vec{\theta}) \in C'(\mathbb{R}^m) \qquad \nabla g(\vec{\theta}) \neq 0 \; \theta \in \Theta \subset \mathbb{R}^m$$

$$\sqrt{n}(g(\tilde{\vec{\theta}}) - g(\vec{\theta})) \rightsquigarrow N(\vec{0}, \nabla^T g(\vec{\theta}) I^{-1}(\vec{\theta}) \nabla g(\vec{\theta}))$$

Пример. $\xi \sim R(0,\theta)$, $\theta > 0$, ОМП: $\tilde{\theta} = x_{max}$

$\tilde{\theta}' = \frac{n+1}{n} x_{max}$ несмещ

$M[x_{max}] = \frac{n}{n+1}\theta \underset{n\to\infty}{\to} \theta$ асим. несм.

$x_{max} \xrightarrow{p} \theta$ сост.

$D[x_{max}] = \frac{n\theta^2}{(n+1)^2(n+2)}$, $I(\theta) = \frac{1}{\theta^2}$, $nD[x_{max}] \underset{n\to\infty}{\not\to} \frac{1}{I(\theta)}$, значит не является эффективной (на самом деле оценка сверхэффективная, модель нерегулярна

условия теоремы не выполнены)

Пусть асим. норм.

$$\sqrt{n}(\underbrace{\tilde{\theta}}_{x_{max}} - \theta) \rightsquigarrow N(0, \sigma^2)$$

$$P(\sqrt{n}(x_{max} - \theta) < x) \underset{n\to\infty}{\to} \Phi(x) \; \forall x \in \mathbb{R}$$

$$P(\sqrt{n}(x_{max} - \theta) < 0) = 1 \to 1 \qquad \Phi(0) = \frac{1}{2}$$

Противоречие $\Rightarrow$ не является асим. нормальной.

9 Доверительный интервал

Определение 9.1. Доверительным интервалом величины h вероятностной модели называется случайный интервал, который накрывает значение h с вероятностью, не меньшей β.

$$I = (g_1(\vec{x}_n), g_2(\vec{x}_n))$$
$$P((g_1, g_2) \ni \theta) \geq \beta$$

β - доверительная вероятность, чаще всего $0.9, 0.95, 0.99$.

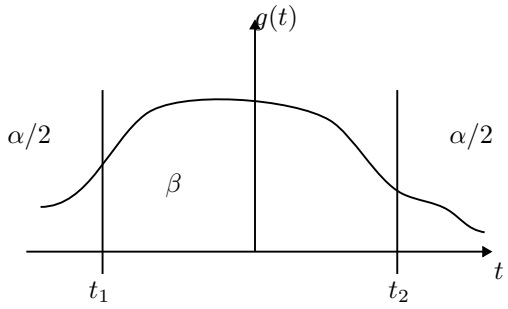
9.1 Методы построения доверит. инт.

- точный
- асимптотический
  - по ОММ
  - по ОМП
- численные
  - параметрический бутстрап
  - непараметрический бутстрап

9.1.1 Точный метод

$h, \vec{x}_n \to f(g, \vec{x}_n) \sim g(t)$ (созерцание)

1. $g(t)$ - плотность распр. (сл. вел. непр.)



$\alpha + \beta = 1$, квантиль $F(x_p) = p$, x_p - квантиль порядка p, $F(x_p) = \int_{-\infty}^{x_p} q(t)dt$

$$t_1 = g_{\alpha/2} = g_{\frac{1-\beta}{2}} \qquad t_2 = g_{\beta+\frac{\alpha}{2}} = g_{\frac{1+\beta}{2}}$$
$$P(t_1 < f(h, \vec{x}_n) < t_2) = \beta$$
$$g_1(\vec{x}_n) < h < g_2(\vec{x}_n)$$

2. $g(t)$ содержит дискр. части, сдвигаем β так чтобы получить точное равенство

Пример. $\xi \sim N(\theta_1, \theta_2^2)$, $\theta_1 \in \mathbb{R}$, $\theta_2 > 0$

$$\tilde{\theta}_1 = \bar{x} \qquad \tilde{\theta}_2^2 = S^2$$

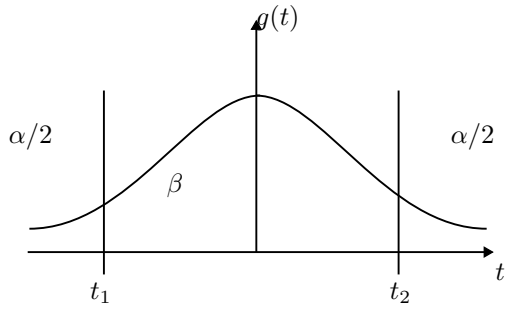
Теорема Фишера:

$$\sqrt{n}\frac{\bar{x} - \theta_1}{\theta_2} \sim N(0,1) \qquad \frac{S^2(n-1)}{\theta_2^2} \sim \chi^2(n-1)$$

$$t_1 = \chi^2_{\frac{1-\beta}{2}}(n-1) \qquad t_2 = \chi^2_{\frac{1+\beta}{2}}(n-1)$$
$$P(t_1 < \frac{S^2(n-1)}{\theta_2^2} < t_2) =; B$$
$$\frac{S^2(n-1)}{t_2} < \theta_2^2 < \frac{S^2(n-1)}{t_1}$$
$$\sqrt{\frac{S^2(n-1)}{t_2}} < \theta_2 < \sqrt{\frac{S^2(n-1)}{t_1}}$$

$$\frac{\sqrt{n}\frac{\bar{x}-\theta_1}{\theta_2}}{\sqrt{\frac{S^2(n-1)}{\theta_2^2(n-1)}}} \sim t(n-1)$$

$$\sqrt{n}\frac{\bar{x}-\theta_1}{S} \sim t(n-1)$$



$$t_1 = t_{\frac{1-\beta}{2}}(n-1) \qquad t_2 = t_{\frac{1+\beta}{2}}(n-1)$$

$$P(t_1 < \sqrt{n}\frac{\bar{x}-\theta_1}{S} < t_2) = \beta$$

$$\bar{x} - \frac{St_2}{\sqrt{n}} < \theta_1 < \bar{x} - \frac{St_1}{\sqrt{n}}$$

9.2 Асимптотический метод

$h, \vec{x}_n \to f(h, \vec{x}_n) \rightsquigarrow g(t)$

По ОММ

$$\sqrt{n}(g(\tilde{\vec{\theta}}) - g(\vec{\theta})) \rightsquigarrow N(\vec{0}, \nabla^T g(\vec{\theta}) I^{-1}(\vec{\theta}) \nabla g(\vec{\theta}))$$

$$\alpha = \begin{pmatrix} \alpha_{s_1} \\ \vdots \\ \alpha_{s_k} \end{pmatrix} \qquad \tilde{\alpha} = \begin{pmatrix} \tilde{\alpha}_{s_1} \\ \vdots \\ \tilde{\alpha}_{s_k} \end{pmatrix} \qquad K_{ij} = \alpha_{s_i + s_j} - \alpha_{s_i}\alpha_{s_j}$$

$$(\tilde{\alpha}_j - \alpha_j)\sqrt{n} \rightsquigarrow N(0, \alpha_{2k} - \alpha_k^2)$$

$$\frac{\tilde{a}_k - \alpha_k}{\sqrt{\alpha_{2k} - \alpha_k^2}}\sqrt{n} \rightsquigarrow N(0,1)$$

$$\tilde{\alpha}_k \xrightarrow{p} \alpha_k \qquad \sqrt{\tilde{\alpha}_{2k} - \tilde{\alpha}_k^2} \xrightarrow{p} \sqrt{\alpha_{2k} - \alpha_k^2}$$

$$\frac{\sqrt{\alpha_{2k} - \alpha_k^2}}{\sqrt{\tilde{\alpha}_{2k} - \tilde{\alpha}_k^2}} \xrightarrow{p} 1$$

Лемма Слуцкого $\xi_n \xrightarrow{F} \xi$, $\eta_n \xrightarrow{p} C$, $\xi_n\eta_n \xrightarrow{F} C\xi$

$$\frac{\tilde{a}_k - \alpha_k}{\sqrt{\alpha_{2k} - \alpha_k^2}}\sqrt{n}\frac{\sqrt{\alpha_{2k} - \alpha_k^2}}{\sqrt{\tilde{\alpha}_{2k} - \tilde{\alpha}_k^2}} \rightsquigarrow N(0,1)$$

$$\boxed{\frac{\sqrt{n}(g(\tilde{\alpha}) - g(\alpha))}{\sqrt{\nabla^T g(\tilde{\alpha}) K(\tilde{\alpha}) \nabla g(\tilde{\alpha})}} \rightsquigarrow N(0,1) \qquad \text{ОММ}}$$

$$\boxed{\frac{\sqrt{n}(g(\tilde{\vec{\theta}}) - g(\vec{\theta}))}{\sqrt{\nabla^T g(\tilde{\vec{\theta}}) I^{-1}(\tilde{\vec{\theta}}) \nabla g(\tilde{\vec{\theta}})}} \rightsquigarrow N(0,1) \qquad \text{ОМП}}$$

Пример. $\xi \sim \rho(x) = x\theta\{(0,1)\} + (1 - \frac{\theta}{2})\{2\}$, $\theta \in (0,2)$

$\vec{x}_n = \{0.53, 0.84, 0.1, 0.83, \underbrace{2, \ldots, 2}_{16}\}$, $n = 20$

1. ОММ $\tilde{\theta} = \frac{3}{2}(2 - \bar{x})$, $\tilde{\theta} = 0.4275$, $S = 0.6$, $\beta = 0.95$

$$\tilde{\theta} = \frac{3}{2}(2 - \tilde{\alpha}_1) = g(\tilde{\alpha}_1) \qquad \theta = g(\alpha_1)$$

$$\frac{\sqrt{n}(g(\tilde{\alpha}_1) - g(\alpha_1))}{\sqrt{\nabla^T g(\tilde{\alpha}) K(\tilde{\alpha}) \nabla g(\tilde{\alpha})}} \rightsquigarrow N(0,1)$$

$$\nabla g = 0\frac{3}{2} \qquad K_{11} = \alpha_2 - \alpha_1^2$$

$$\frac{\sqrt{n}(\tilde{\theta} - \theta)}{\sqrt{-\frac{3}{2}(-\frac{3}{2})(\tilde{\alpha}_2 - \tilde{\alpha}_1^2)}} \rightsquigarrow N(0,1)$$

$$\tilde{\mu}_2 = \tilde{\alpha}_2 - \tilde{\alpha}_1^2 = 0.342 \qquad \tilde{\mu}_2 = S^2\frac{n-1}{n}$$

$$t_1 = u_{\frac{1-0.95}{2}} = u_{0.025} = -1.96 \qquad t_2 = u_{\frac{1+0.95}{2}} = u_{0.025} = 1.96$$

$$-1.96 < \frac{\sqrt{20}(0.4275 - \theta)}{\sqrt{\frac{9}{4} \cdot 0.342}} < 1.96$$
$$0.0435 < \theta < 0.811 \qquad l = 0.768$$

2. ОМП $\tilde{\theta} = 2(1 - \nu) = 2(1 - \frac{16}{20}) = 0.4$

$$\frac{\sqrt{n}(\tilde{\theta} - \theta)}{\sqrt{I^{-1}(\theta)}} \rightsquigarrow N(0, 1)$$
$$I(\theta) = \frac{1}{\theta(2 - \theta)}$$
$$\frac{\sqrt{n}(\tilde{\theta} - \theta)}{\sqrt{\tilde{\theta}(2 - \tilde{\theta})}} \rightsquigarrow N(0, 1)$$
$$-1.96 < \frac{\sqrt{20}(0.4 - \theta)}{\sqrt{0.4 \cdot 1.6}} < 1.96$$
$$0.049 < \theta < 0.751 \qquad l = 0.702$$

ОМП довер. инт. дисперсии

$$D\xi = \frac{11}{12}\theta - \frac{4}{9}\theta^2$$
$$\tilde{D\xi} = \frac{11}{12} \cdot 0.4 - \frac{4}{9} \cdot 0.4^2 = 0.296$$
$$\frac{\sqrt{n}(g(\tilde{\theta}) - g(\theta))}{\sqrt{\nabla^T g(\tilde{\theta}) I^{-1}(\tilde{\theta}) \nabla g(\tilde{\theta})}} \rightsquigarrow N(0, 1)$$
$$\nabla g(\theta) = \frac{11}{12} - \frac{8}{9}\theta$$
$$-1.96 < \frac{\sqrt{20}(\tilde{D\xi} - D\xi)}{\sqrt{(\frac{11}{12} - \frac{8}{9} \cdot 0.4)^2 \cdot 0.4 \cdot 1.6}} < 1.96$$
$$0.1 < D\xi < 0.492$$

9.3 Численные методы

1. Непараметрические методы

   $h, \vec{x}_n \to \tilde{h}$, берём выборку за вероятностную модель, из выборки формируем подвыборку $N = 1000$, $\tilde{h} - h = \tilde{\Delta}$

   (a) $\vec{x}_n^*$ с повторение элем. $\Delta_1^* = \tilde{h}^* - \tilde{h}$

   (b) ...

   (c) ???

(d) ...

(e) $\vec{x}_n^*$ с повторение элем. $\Delta_{1000}^* = \tilde{h}^* - \tilde{h}$

(f) Profit!

Вариационный ряд $\Delta_{(1)}^*, \ldots, \Delta_{(1000)}^*$

$$K_1 = \left[\frac{1-\beta}{2} \cdot 1000\right] \qquad K_2 = \left[\frac{1+\beta}{2} \cdot 1000\right]$$
$$t_1 = \Delta_{(k_1)} \qquad t_2 = \Delta_{(k_2)}$$
$$P(t_1 < \tilde{h}^* - \tilde{h} < t_2) \approx \beta$$
$$P(t_1 < h^* - h < t_2) \approx \beta$$
$$\tilde{h} - t_2 < h < \tilde{h} - t_1$$

2. Параметричекий бутстрап

   $h, \tilde{h}, \Delta = \tilde{h} - h$, $\xi \sim \rho(x, h)$, $\vec{x}_n \to \tilde{h}$ - сост. и несм. оценка

   $\xi \sim \rho(x, \tilde{h})$ моделируем выборки $\vec{x}_n^*$, $N = 50000$

   $\Delta_i^* = \tilde{h}^* - \tilde{h}$, $\Delta_{(1)}^* \ldots \Delta_{(N)}^*$

$$K_1 = \left[\frac{1-\beta}{2} \cdot N\right] \qquad K_2 = \left[\frac{1+\beta}{2} \cdot N\right]$$
$$t_1 = \Delta_{(k_1)}^* \qquad t_2 = \Delta_{(k_2)}^*$$
$$P(t_1 < \tilde{h}^* - \tilde{h} < t_2) \approx \beta$$
$$P(t_1 < h^* - h < t_2) \approx \beta$$

Пример. $\vec{x}_n = \{0.53, 0.84, 0.1, 0.83, \underbrace{2, \ldots, 2}_{16}\}$, $n = 20$

1. Непараметрический бутсрап $\tilde{\theta} = 0.4$ ОМП $\tilde{\theta} = 2(1 - \nu)$
   - $\vec{x}_n^* = 0.53, \ldots$, $m = 14$, $\Delta_1^* = \tilde{\theta}^* - \tilde{\theta} = 0.2$
   - ...
   - $\Delta_{1000}^* = 0$

$$\beta = 0.95 \qquad k_1 = 25 \qquad k_2 = 975$$
$$t_1 = \Delta_{(25)}^* = -0.3 \qquad t_2 = \Delta_{(975)}^* = 0.4$$
$$P(-0.3 < \tilde{\theta}^* - \tilde{\theta} < 0.4) \approx 0.95$$
$$P(-0.3 < \tilde{\theta} - \theta < 0.4) \approx 0.95$$
$$0 < \theta < 0.7 \qquad l = 0.7$$

2. Параметрический бутстрап $\tilde{\theta} = 0.4$

$$\xi \sim \rho(x, \theta) = x\theta\{(0,1)\} + (1 - \frac{\theta}{2})\{2\} = x \cdot 0.4\{(0,1)\} + 0.8\{2\}$$

$N = 10000$, делаем выборки из модели и дальше так же как и непарам.

9.4 Доверительный интервал для частоты

$\nu = \frac{m}{n}$ хорошая оценка $P(A)$

Интегральная теорема Муавра-Лапласа:

$$\frac{m - np}{\sqrt{np(1-p)}} \rightsquigarrow N(0,1)$$
$$\frac{\nu - p}{\sqrt{p(1-p)}}\sqrt{n} \rightsquigarrow N(0,1)$$
$$\frac{\nu - p}{\sqrt{\nu(1-\nu)}}\sqrt{n} \rightsquigarrow N(0,1)$$

9.5 Доверительный интервал функции распределения

F, $\tilde{F}(x) = \frac{m}{n}$, где m - кол-во элементов меньше x

$$\frac{\tilde{F}(x) - F(x)}{\sqrt{\tilde{F}(x)(1 - \tilde{F}(x))}}\sqrt{n} \rightsquigarrow N(0,1)$$

10 Проверка статистических гипотез

Определение 10.1. Гипотеза - любое высказывание о вероятностной модели.

Определение 10.2. Простая гипотеза - однозначное определение вероятностной модели.

Определение 10.3. Сложная гипотеза - неоднозначное определение вер. модели.

$\rho \sim N(0, a)$, $H : a = 2$ - простая, $H : a > 2$ - сложная

Определение 10.4. H_0 - основная гипотеза, H_1 - альтернативная (отклонение от основной)

$H_0 : a = 2$, $H_1 : a > 2$

10.1 Принцип от маловеороятного

Пусть H_0 - верна. Событие A. $P(A|H_0)$ - мала. Событие A наблюдаемо. Отвергаем H_0, иначе нет оснований отвергнуть H_0.

11 Критерии согласия

11.1 Критерий Пирсона

Вероятностная модель, $\vec{x}_n$

$H_0 : \xi \sim F(x)$ - простая гипотеза

$H_1 : \bar{H}_0$ - сложная

Полная группа событий: $A_1 \dots A_k$ - конечная группа событий, $A_1 + \dots + A_n = \Omega$, $A_i A_j = \emptyset$ $i \neq j$, $P(A_i) > 0$

$H_0 : P_i = P(A_i)$, $\tilde{P}_i = \nu_i = \frac{m_i}{n}$

$$\Delta = n \sum_{i=1}^{k} \frac{(p_i - \nu)^2}{p_i}$$

$$\Delta = n \sum_{i=1}^{k} \frac{(p_i - \frac{m_i}{n})^2}{p_i} = \sum_{i=1}^{k} \frac{(np_i - m_i)^2}{np_i}$$

Теорема 11.1. Если H_0 верна, то $\Delta \rightsquigarrow \chi^2(k-1)$

Замечание. Нормальная аппроксимация при $n \geq 50$, $np_i \geq 5$ (можно мухлевать: $np_i \geq 1$ и в 20% случаев можем разрешить $np_i < 5$)

Пример. H_0 : красные автомобили штрафуют в 2 раза чаще остальных

$H_1 : \bar{H}_0$

Выборка: 150 штрафов, 90 красные, A_{red}, A_{other}

	A_r	A_o
p_i	2/3	1/3
np_i	100	50
np_i	90	60

$$\tilde{\Delta} = \frac{(100-90)^2}{100} + \frac{(50-60)^2}{50} = 3$$

α - уровень значимости, $\alpha = .1; \boxed{.05}; .01$

$H_0 : \Delta \rightsquigarrow \chi^2(1)$, $k = 2$

p-value $= P(\Delta \geq \tilde{\Delta}|H_0) = \int_3^\infty q(t)dt = 0.083$

Нет веских оснований отвергнуть H_0.

Замечание. Правила использования p-value:

1. Если p-value $\leq \alpha$, то H_0 отвергается, результаты значимы, p-value - мера значимости
2. Если p-value $> \alpha$, то нет оснований отвергать H_0, результаты незначимы.

   Либо гипотеза верна (отклонения объясняются случайными факторами), либо критерий недостаточно мощный.

p-value не является вероятностью H_0 и не является вероятностью случайных факторов!

Пример (Закон Бенфорда)**.** В больших массивах данных, полученных естественным путём, следуют определённому распределению.

$$P_i = \log_{10}(1 + \frac{1}{d_i}) \qquad d_i = 1, \ldots, 9$$

Рассмотрим числа Фибоначи (первые 100 штпук)

	1	2	3	4	5	6	7	8	9
m_i	29	18	13	9	8	6	5	7	5
p_i	0.3	0.18	0.12	0.1	0.08	0.07	0.06	0.05	0.04
np_i	3	18	12	10	8	7	6	5	4

8 и 9 объединяем чтобы попадать под условия приминимости

$$\tilde{\Delta} = \frac{(30-29)^2}{30} + \cdots + \frac{(9-12)^2}{9} = 1.53 \qquad \Delta \rightsquigarrow \chi^2(7)$$

$$\text{p-value} = P(\Delta \geq \tilde{\Delta}|H_0) = \int_{1.53}^\infty q(t)dt = 0.981$$

Нет оснований отвергать H_0.

11.2 Критерий Колмогорова (непр. распр.)

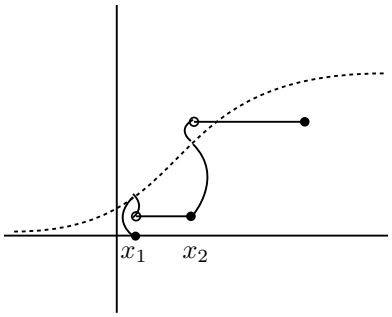
$H_0 : \xi \sim F(x)$, $H_1 : \bar{H_0}$, $\vec{x}_n$

$$\Delta = \sqrt{n} \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)|$$

- где $\tilde{F}$ - империческая функция распределения.

Теорема 11.2. Если H_0 верна, то $\Delta \rightsquigarrow K(x)$ $K(x)$ - функ. распределения Колмогорова

$$K(x) = P(\Delta < x) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2}\{(0,\infty)\}$$



$$\sup_{\mathbb{R}}|\tilde{F}(x) - F(x)| = \max_{i=1,\dots,n}\max(|\tilde{F}(x_i - 0) - F(x_i)|, |\tilde{F}(x_i + 0) - F(x_i)|)$$

Пример. $H_0 : \xi \sim R(0,1)$, $H_1 : \bar{H_0}$, $\vec{x}_n = (0.2; 0.5; 0.8; 0.9)$

$$\tilde{\Delta} = \sqrt{4}\max(0.2; 0,25; 0.3; 0.15) = 2 * 0.3 = 0.6$$

$H_0 : \Delta \rightsquigarrow K(x)$

p-value $= P(\Delta \geq \tilde{\Delta}|H_0) = 1 - K(\tilde{\Delta}) = -2\sum_{k=1}^{\infty}(-1)^j e^{-2k^2\tilde{\Delta}^2} = 0.8642$

4 элемента - вообще беда, используем бутстрап, так как считаем в вероятности, что верна H_0 по хорошему бутстрап параметрический

- $x_4^* \to \Delta_1^* = \sqrt{n}\sup|\tilde{F}^*(x) - F(x)|$
- ...
- $x_4^* \to \Delta_N^* = \sqrt{n}\sup|\tilde{F}^*(x) - F(x)|$

$N = 10000 - 50000$, вариационный ряд $\Delta_{(1)}^* \dots \Delta_{(N)}^*$

p-value $= \frac{K}{N}$, K - число $\Delta_{(i)}^* \geq \tilde{\Delta}$

p-value $= 0.47$

11.3 Критерий Пирсона для слож. гипотезы

$H_0 : \xi \sim F(x, \vec{\theta})$, $\vec{\theta} \in \Theta \subset \mathbb{R}^m$, $H_1 : \bar{H}_0$, $\vec{x}_n$, $A_1 \dots A_k$ - полная группа событий

$$P_i(\vec{\theta}) = P(A_i) \qquad P_i = \nu_i$$
$$\Delta = \sum_{i=1}^{n} \frac{(np_i(\vec{\theta}) - m_i)^2}{np_i(\vec{\theta})}$$

Теорема 11.3. Если H_0 верная и $\tilde{\vec{\theta}}$ есть ОМПГ, то $\Delta \rightsquigarrow \chi^2(k-1-m)$

Пример. $H_0 : \xi \sim R(0, \theta)$, $\theta > 0$, $H_1 : \bar{H}_0$

	$[0,1)$	$[1,2)$	$[2,3)$	$[3,4)$	$[4,\infty)$
m_1	25	10	15	30	20
p_i	1/5	1/5	1/5	1/5	1/5
np_i	20	20	20	20	20

$$P_i = \int_{a_i}^{b_i} \frac{1}{\theta} dx \qquad P_1 = P_2 = P_3 = P_4 = \frac{1}{\theta}$$
$$P_5 = \int_4^{\infty} p(x, \theta) dx = \int_4^{\theta} \frac{1}{\theta} dx = \frac{\theta - 4}{\theta}$$
$$L(\theta) = \left(\frac{1}{\theta}\right)^{25+10+15+30} \left(\frac{\theta - 4}{\theta}\right)^{20} \to \max$$
$$\ln L = -80 \ln \theta + 20 \ln(\theta - 4) - 20 \ln \theta = -100 \ln \theta + 20 \ln(\theta - 4) \to \max$$
$$(\ln L)' = -\frac{100}{\theta} + \frac{20}{\theta - 4} = 0 \qquad \tilde{\theta} = 5$$

$$\Delta \rightsquigarrow \chi^2(5 - 1 - 1) = \chi^2(3)$$
$$\tilde{\Delta} = \frac{(25 - 20)^2}{20} + \dots + \frac{(20 - 20)^2}{20} = 12.5$$
$$\text{p-value} = P(\Delta \geq \tilde{\Delta} | H_0) = \int_{12.5}^{\infty} q(t) dt = 0.00585$$

Отвергаем H_0.

Пример. H_0 : распределение Эрланга $\xi \sim p(x) = \frac{1}{\lambda^2} x e^{-x \ln x} \{(0, +\infty)\}$, $\lambda > 0$, $H_1 : \bar{H}_0$

	$[0; 2.5)$	$[2.5; 5)$	$[5; 7.5)$	$[7.5; 10)$	$[10; 12.5)$	$[12.5; 15)$
m_i	12	17	12	4	3	2
np_i	12.9	16.4	10.4	5.5	2.6	2.2

$$\int_a^b p(x)dx = \frac{a}{\lambda}e^{-a/\lambda} - \frac{b}{\lambda}e^{-b/\lambda} + e^{-a/\lambda} - e^{-b/\lambda}$$

$$P_1 = \int_0^{2.5} p(x)dx \qquad \dots \qquad P_6 = \int_{12.5}^{\infty} p(x)dx$$

$$L(\lambda) = P_1^{12} \dots P_6^2 \to \max$$

Считаем экстремум численно.

$$\tilde{\lambda} = 2.531$$

После объединения и пересчёта:

$$\tilde{\lambda} = 2.542 \qquad \tilde{\Delta} = 0.756$$

$$\text{p-value} = P(\Delta \geq \tilde{\Delta}|H_0) = 0.86$$

11.4 Критерий Колмогорова для сл. гип. (непр. распр.)

$H_0 : \xi \sim F(x, \vec{\theta})$, $\vec{\theta} \in \Theta \subset \mathbb{R}^m$, $H_1 : \bar{H}_0$, $\vec{x}_n$, $A_1 \dots A_k$ - полная группа событий

$$\Delta = \sqrt{n} \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x, \vec{\theta})|$$

Параметрический бутстрап: $\vec{x}_n \to \tilde{\vec{\theta}}$ - сост. оценка (любой метод)

$$\tilde{\Delta} = \sqrt{n} \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x, \tilde{\vec{\theta}})|$$

$\xi \sim F(x, \tilde{\vec{\theta}})$, $N = 10000 - 50000$

- $\vec{x}_n^* \to \tilde{\vec{\theta}}^* \to \Delta_1^* = \sqrt{n} \sup\limits_{x \in \mathbb{R}} |\tilde{F}^*(x) - F(x, \tilde{\vec{\theta}})|$
- ...

Вариационный ряд $\Delta_{(1)}^* \dots \Delta_{(N)}^*$, p-value $= \frac{K}{N}$, K - число элементов $\Delta_{(i)}^* \geq \tilde{\Delta}$

Пример (Эрланг)**.**

$$\tilde{\Delta} = 0.3982 \qquad \text{p-value} = 0.9284$$

11.5 Проверка гипотезы однородности

$A_1 \dots A_k$ - полная группа

- 1 выборка $m_{11} \dots m_{1k}$
- ...
- l выборка $m_{l1} \dots m_{lk}$

H_0 : все выборки из одного распределения, $H_1 : \bar{H}_0$

$$n = n_1 + \cdots + n_l \qquad \nu_j = \frac{\sum_{i=1}^{l} m_i j}{n}$$

$$\Delta_1 = \sum_{j=1}^{k} \frac{(m_{1j} - n_1\nu_j)^2}{n_1\nu_j} \qquad \Delta_s = \sum_{j=s}^{k} \frac{(m_{sj} - n_s\nu_j)^2}{n_s\nu_j}$$

$$\Delta = \Delta_1 + \cdots + \Delta_l$$

Теорема 11.4. Если H_0 верна, то $\Delta \rightsquigarrow \chi^2((k-1)(l-1))$

Пример (Коллоквиум 2023)**.**

	2	3	4	5	
1гр	0	1	10	5	16
2гр	1	0	9	1	11
3гр	2	4	6	4	16
4гр	1	0	8	6	15
	4	5	33	16	

2 и 5	3 и 4
5	11
2	9
6	10
7	8
$\frac{20}{58}$	$\frac{38}{58}$

$$\Delta_1 = 0.074 \qquad \Delta_2 = 1.294 \qquad \Delta_3 = 0.064 \qquad \Delta_4 = 0.986$$

$$\tilde{\Delta} = 2.418 \qquad \chi^2(1 \cdot 3)$$

$$\text{p-value} = P(\Delta \geq \tilde{\Delta} | H_0) = 0.49$$

11.6 Проверка гипотезы независимости

(ξ, η), $H_0 : \xi$ и η - независимы, $H_1 : \bar{H}_0$, A_i и B_j - полные группы событий

	A_1	$\ldots$	A_k
B_1	m_{11}	$\ldots$	m_{1k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_l	m_{l1}	$\ldots$	m_{lk}

$$P_1 = \frac{m_{11} + \cdots + m_{1k}}{n} \qquad q_1 = \frac{m_{11} + \cdots + m_{l1}}{n}$$

$$\Delta = \sum_{i,j} \frac{(m_{ij} - np_iq_j)^2}{np_iq_j}$$

Теорема 11.5. Если H_0 верна, то $\Delta \rightsquigarrow \chi^2((k-1)(l-1))$.

Пример. $n = 246$, зависимость успеваемости от родного города

	Москва	Подмоск	Остальное	
2	13	10	17	$\frac{40}{246}$
3	35	38	19	$\frac{92}{246}$
4	21	21	26	$\frac{68}{246}$
5	16	11	19	$\frac{46}{246}$
	$\frac{85}{246}$	$\frac{80}{246}$	$\frac{81}{246}$	

$$\tilde{\Delta} = 11.05 \qquad \text{p-value} = \int_{11.05}^{\infty} q(t)dt = 0.087$$

Нет оснований отвергать H_0.

11.7 Проверка гипотезы случайности

H_0 : элементы выборки независимы и одинакого распределены, $H_1 : \bar{H_0}$, I_n - число инверсий в выборке

Теорема 11.6. Если H_0 верно, то

$$\Delta = \frac{I_n - \frac{n(n-1)}{4}}{\sqrt{\frac{n^3}{36}}} \rightsquigarrow N(0,1)$$

Пример. $n = 10$, $\vec{x}_n = (4, 10, 8, 8, 6, 7, 7, 9, 8, 9)$

$$I_n = 0 + 8 + 3 + 3 + \cdots + 0 = 15$$

$$\tilde{\Delta} = \frac{15 - \frac{10 \cdot 9}{4}}{\sqrt{\frac{10^3}{36}}} = -1.42$$

$$\text{p-value} = P(|\Delta| \geq |\tilde{\Delta}| \big| H_0) = 2P(\Delta \geq |\tilde{\Delta}| \big| H_0) = 2\int_{1.42}^{\infty} q(t)dt = 0.155$$

Нет оснований отвергать H_0.

11.8 Критерий Смирнова для проверки гипотезы однородности (непр. расп.)

$H_0 : F(x) = G(x)$, $H_1 : \bar{H_0}$ Выборки $\vec{x}_n$, $\vec{y}_m$

$$\Delta = \sqrt{\frac{nm}{n+m}} \sup_R |\tilde{F}(x) - \tilde{G}(x)|$$

где $\tilde{F}$ и $\tilde{G}$ - эмпирические функции распределения

Теорема 11.7. Если H_0 верна, то $\Delta \rightsquigarrow K(x)$

12 Использование доверительных интервалов для проверки гипотез

Вероятностная модель, h - характеристика вероятностной модели, $\tilde{h}$ - оценка

$H_0 : h \in T_0$, $H_1 : h \in T_1$, $T_0 \cap T_1 = \emptyset$

$\tilde{h} \ \rightarrow$ доверительный интервал (точный, асимптотический или непараметрический бутстраповский)

Если $I \subset T_1$, то H_0 отвергаем

Пример. $H_0 : h \leq 3$, $H_1 : h > 3$

- Если $\tilde{h} \leq 3$ сразу говорим что нет оснований отвергать гипотезу
- Если $\tilde{h} > 3$, считаем доверительный интервал (правосторонний): $\tilde{h} \rightarrow I = [h_0, +\infty)$

Пример. $H_0 : h = 3$, $H_1 : h \neq 3$, в таком случае интервал двухсторонний: $I = (\tilde{h} - \delta, \tilde{h} + \delta)$

13 Методы построения критериев

$G_{\text{кр}}$ (G_{cr}) - критическая область

H_0 отвергаем $\Leftrightarrow \vec{x}_n \in G_{cr}$

$P(\vec{x}_n \in G \,|\, H_0) \leq \alpha$

Ошибка I рода $\alpha_1 = P(H_1|H_0)$.

Ошибка II рода $\alpha_0 = P(H_0|H_1)$.

Если H_0 и H_1 сложные, то $\alpha_1 = \sup\limits_{H_0} P(\vec{x}_n \in G|H_0)$, $\alpha_2 = \sup\limits_{H_1} P(\vec{x}_n \notin G|H_1)$

Определение 13.1 (Мощность критерия)**.**

$$W(H_1) = P(\vec{x}_n \in G|H_1)$$

Для простой гипотезы $W = 1 - \alpha_2$

Определение 13.2 (Состоятельность критерия)**.** Критерий называется состоятельным, если

$$\forall H_1 \hookrightarrow W(H_1) \underset{n\to\infty}{\to} 1$$

Замечание. α_1 - наиболее опасная ошибка

$\alpha_1 \le \alpha$ обязательно, $\alpha_2 \to \min$, решаем оптимизационную задачу $\to$ G_{opt}

Пример. H_0 : болен, H_1 : здоров

$\alpha_1 = P(H_1|H_0)$ - признали больного здоровым - беда

$\alpha_2 = P(H_0|H_1)$ - признали здорового больным - не так критично

13.1 Теорема Неймана-Пирсона для проверки простой H0 против простой H1

$\alpha_2 \to \min$, $\alpha_1 \le \alpha$

$$W = 1 - \alpha_2 = P(\vec{x}_n \in G|H_1) \to \max$$
$$\int_G L_1(\vec{x}_n)d\vec{x}_n = \int_G \underbrace{\frac{L_1}{L_0}}_{l} L_0 d\vec{x}_n \to \max$$
$$G : l \ge C \qquad \int_G L_0 d\vec{x}_n \le L$$

l - отношение правдоподобия

Пример. $n = 25$, $\vec{x} = 9.3$, $\alpha = 0.02$

$H_0 : \xi \sim N(10, 4)$, $H_1 : \xi \sim N(9, 4)$

$$l = \frac{L_1}{L_0} = \frac{\prod_{i=1}^n p_1(x_i)}{\prod_{i=1}^n p_0(x_i)} = \frac{\left(\frac{1}{\sqrt{2\pi}\cdot 2}\right)^n e^{-\frac{1}{8}\sum_{i=1}^n (x_i-9)^2}}{\left(\frac{1}{\sqrt{2\pi}\cdot 2}\right)^n e^{-\frac{1}{8}\sum_{i=1}^n (x_i-10)^2}} =$$
$$= e^{-\frac{1}{8}\sum_{i=1}^n (2x_i-10)} = e^{-\frac{1}{4}n\bar{x}+\frac{19}{8}n} \ge C$$
$$-\frac{1}{4}n\bar{x} + \frac{19}{8}n \ge \ln C \qquad G : \bar{x} \le A$$
$$P(\bar{x} \le A|H_0) \le \alpha$$

По Т. Фишера:

$$\frac{\bar{x} - a}{\sigma}\sqrt{n} \sim N(0, 1)$$
$$P(\frac{\bar{x} - 10}{2}\sqrt{25} \le \frac{A - 10}{2}\sqrt{25}) \le \alpha$$

$\frac{5}{2}(A-10) = U_\alpha$ - квантиль порядка α для нормального распределения

$$\frac{A-10}{2}5 = U_\alpha = -2.054$$
$$G : \bar{x} \le 9.18$$

$\bar{x} \notin G$ нет оснований отвергунть H_0

$$W = P(\bar{x} \in G|H_1) = P(\frac{\bar{x}-9}{2}5 \le \frac{9.18-9}{2}5) = \int_{-\infty}^{0.45} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0.67$$
$$\alpha_1 = 0.02 \qquad \alpha_2 = 1 - W = 0.33$$

Сколько нужно n чтобы имело смысл

$$\alpha_1 = 0.01 \qquad \alpha_2 = 0.1$$
$$G \cdot \bar{x} \le B$$
$$\alpha_1 = P(\bar{x} \le B|H_0) = P(\frac{\bar{x}-10}{2}\sqrt{n} \le \frac{B-10}{2}\sqrt{n}) = 0.01$$
$$\alpha_2 = P(\bar{x} > B|H_0) = P(\frac{\bar{x}-9}{2}\sqrt{n} \le \frac{B-10}{2}\sqrt{n}) = 0.1$$
$$\begin{cases} \frac{B-10}{2}\sqrt{n} = U_{0.01} = -2.33 \\ \frac{B-9}{2}\sqrt{n} = U_{0.9} = 1.28 \end{cases}$$
$$B = 9.36 \qquad n = 53$$

Пример. $\xi \sim \xi(x) = (2-2\theta)x\{(0,1)\} + \theta\{0\}$

$H_0 : \theta = \frac{1}{3}$, $H_1 : \theta = \frac{1}{2}$, $n = 1$, $\alpha = 0.35$

$$l = \frac{L_1}{L_0} = \frac{\rho_1(x)}{\rho_0(x)}$$

x	$(0,1)$	2
l	$\frac{x}{4x/3} = \frac{3}{4}$	$\frac{1/2}{1/3} = \frac{3}{2}$
H_0, p	$\frac{2}{3}$	$\frac{1}{3}$
H_1, p	$\frac{1}{2}$	$\frac{1}{2}$

$$G : l \ge C \qquad P(l \ge C|H_0) \le 0.35$$

$G : l \ge \frac{3}{2}$, $\alpha_1 = \frac{1}{3}$, $\alpha_2 = P(l < C|H_1) = \frac{1}{2}$

Пример. Те же условия что и в прошлом но $n = 2$ l:

	$(0,1)$	2
$(0,1)$	$\frac{1 \cdot 1 \cdot x_1 \cdot x_2}{\frac{4}{3}x_1\frac{4}{3}x_2} = \frac{9}{16}$	$\frac{x_2 \cdot \frac{1}{2}}{\frac{4}{3}x_2\frac{1}{3}} = \frac{9}{8}$
2	$\frac{9}{8}$	$\frac{\frac{1}{2}\frac{1}{2}}{\frac{1}{3}\frac{1}{3}} = \frac{9}{4}$

H_0:

	$(0,1)$	2
$(0,1)$	$\frac{4}{9}$	$\frac{2}{9}$
2	$\frac{2}{9}$	$\frac{1}{9}$

H_1:

	$(0,1)$	2
$(0,1)$	$\frac{1}{4}$	$\frac{1}{4}$
2	$\frac{1}{4}$	$\frac{1}{4}$

$$G : l \geq C \qquad P(l \geq C|H_0) \leq 0.35$$

$G : l \geq \frac{9}{4}$, $\alpha_1 = \frac{1}{9}$, $\alpha_2 = P(l < C|H_1) = \frac{3}{4}$, $W = \frac{1}{4}$

Асимптотический критерий $n \to \infty$

$$l = \frac{L_1}{L_0} = \prod_{i=1}^{n} \frac{\rho_1(x_i)}{\rho_0(x_i)} \geq C$$

$$\ln l = \sum_{i=1}^{n} \ln \underbrace{\frac{\rho_1(x_i)}{\rho_0(x_i)}}_{\eta_i} \geq \ln C$$

ЦПТ

Параметрический бутстрап

m - число появлений 2

$$l = \frac{L_1}{L_0} = \frac{\prod x_i \cdot \left(\frac{1}{2}\right)^2}{\left(\frac{4}{3}\right)^{n-m} \prod x_i \left(\frac{1}{3}\right)^m}$$

$$l = \left(\frac{3}{4}\right)^{n-m} \left(\frac{3}{2}\right)^n \geq C \qquad P(l \geq C|H_0) \leq \alpha$$

$$\rho_0(x) = \frac{4}{3} x\{(0,1)\} + \frac{1}{3}\{2\}$$

- $\vec{x}_n^* \to m_1^* \to l_1^*$
- ...
- $\vec{x}_n^* \to m_N^* \to l_N^*$

Вариационный ряд $l_{(1)}^* \dots l_{(N)}^*$, $k = [(1-\alpha)N]$, $C = l_{(k)}^*$

$n = 1$, $\alpha = 0.02$, $N = 50000$, $C = 4$

$W = P(l \geq C|H_1)$, те же самые процедуры с другим распределением $W = \frac{k}{N}$, $W = 0.17$

Пример (T11 из задания)**.** $H_0 : \xi \sim p_0(x) = 1\{(0,1)\}$ $H_1 : \xi \sim p_0(x) = \frac{e}{e-1}e^{-x}\{(0,1)\}$

---

a) $n = 1$, α

$$l = \frac{L_1}{L_0} = \frac{\frac{e}{e-1}e^{-x}}{1} \geq C \qquad e^{-x} \geq B \qquad x \leq A$$

$$P(x \leq A|H_0) = \alpha \qquad \int_0^A 1dx = A = \alpha$$

$$G : x \leq \alpha \qquad \alpha_1 = \alpha$$

$$W = P(x \leq A|H_1) = \int_0^\alpha \frac{e}{e-1}e^{-x}dx = \frac{e}{e-1}(1 - e^{-\alpha}) \qquad \alpha_2 = 1 - W$$

---

b) $n = 2$, α

$$l = \frac{L_1}{L_0} = \frac{\left(\frac{e}{e-1}\right)^2 e^{-x_1}e^{-x_2}}{1 \cdot 1} \geq C \qquad x_1 + x_2 \leq A$$

$$P(x_1 + x_2 \leq A|H_0) = \alpha \qquad \iint\limits_{x_1+x_2\leq A} 1dx_1dx_2 = \frac{A^2}{2} = \alpha \qquad A = \sqrt{2\alpha}$$

$$G : x_1 + x_2 \leq \sqrt{2\alpha} \qquad \alpha_1 = \alpha$$

$$W = P(x_1 + x_2 \leq A|H_1) = \iint\limits_{x_1+x_2\leq A} \left(\frac{e}{e-1}\right)^2 e^{-x_1-x_2}dx_1dx_2 =$$

$$= \left(\frac{e}{e-1}\right)^2 \int_0^A dx_1 \int_0^{A-x_1} e^{-x_1}e^{-x_2}dx_2 =$$

$$= \left(\frac{e}{e-1}\right)^2 \int_0^A e^{-x_1}(1 - e^{-(A-x_1)})dx_1 =$$

$$= \left(\frac{e}{e-1}\right)^2 \int_0^A (e^{-x_1} - e^{-A})dx_1 = \left(\frac{e}{e-1}\right)^2 (1 - e^{-A} - Ae^{-A}) \qquad A = \sqrt{2\alpha}$$

$$\alpha = 1 - W$$

---

c)

$$l = \frac{L_1}{L_0} = \prod \frac{p_1(x_i)}{p_0(x_i)} \geq C$$

$$\ln l = \sum \ln \underbrace{\frac{p_1(x_i)}{p_0(x_i)}}_{\eta_i} \geq \ln C$$

$$\frac{\sum \eta_i - nM\eta_i}{\sqrt{nD\eta_i}} \rightsquigarrow N(0,1)$$

$$P(\ln l \geq \ln C | H_0) = \alpha \qquad \eta = \ln(\frac{e}{e-1}e^{-x}) = \ln\frac{e}{e-1} - x$$

$$\ln l = \sum \ln\frac{e}{e-1} - \sum x_i \geq \ln C \qquad G : \sum x_i \leq A$$

$$P(\sum x_i \leq A | H_0) = \alpha \qquad P(\frac{\sum x_i - nMx}{\sqrt{nDx}} \leq \frac{A - nMx}{\sqrt{nDx}} | H_0) = \alpha$$

$$Mx = \frac{1}{2} \qquad Dx = \frac{1}{2}(b-a)^2 = \frac{1}{12}$$

$$\frac{A - n\frac{1}{2}}{\sqrt{n\frac{1}{2}}} = u_\alpha \qquad A = n\frac{1}{2} + \sqrt{n\frac{1}{12}}u_\alpha$$

$$G : \sum x_i \leq n\frac{1}{2} + u_\alpha\sqrt{\frac{n}{12}} \qquad \alpha_1 = \alpha$$

$$W = P(\sum x_i \leq A | H_1) = P(\frac{\sum x_i - nMx}{\sqrt{nDx}} \leq \frac{A - nMx}{\sqrt{nDx}} | H_1)$$

$$Mx = \int_0^1 x\frac{e}{e-1}e^{-x}dx = \frac{e-2}{e-1} \qquad Mx^2 = \frac{2e-5}{e-1} \qquad Dx = \frac{e^2-3e+1}{(e-1)^2}$$

$$W = \int_{-\infty}^{B} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \qquad B = \frac{\frac{n}{2} + U_\alpha\sqrt{\frac{n}{12}} - n\frac{e-2}{e-1}}{\sqrt{n\frac{e^2-3e+1}{(e-1)^2}}} \qquad \alpha_2 = 1 - W$$

$$B = \frac{\sqrt{n}\cdot 0.082 + U_\alpha\sqrt{\frac{1}{12}}}{\sqrt{\frac{e^2-3e+1}{(e-1)^2}}} \underset{n\to\infty}{\to} \infty$$

Критерий состоятелен.

$\alpha = 0.05$, $n = 1$, $x \leq 0.05$, $W = 0.077$, $\alpha_2 = 0.923$

$\alpha = 0.05$, $n = 2$, $x_1 + x_2 \leq 0.316$, $W = 0.102$, $\alpha_2 = 0.898$

$\alpha = 0.05$, $n = 10$, $\sum x_i \leq 3.5$, $B = -0.78$, $W = 0.22$, $\alpha_2 = 0.78$

---

d) $G : x_{max} \le C$

$$P(\vec{x}_n \in G|H_0) = \alpha \qquad P(x_{max} \le G|H_0) = \alpha$$
$$\xi \sim F(x) \qquad \xi_{max} \sim (F(x))^n$$
$$P(x_{max} \le C|H_0) = \underbrace{(F_0(C))}_{C}{}^n = \alpha \qquad C = \sqrt[n]{\alpha}$$
$$G : x_{max} \le \sqrt[n]{\alpha} \qquad \alpha_1 = \alpha$$
$$W = P(\vec{x}_n \in G|H_1) = P(x_{max} \le C|H_1)$$
$$H_1 : \xi \sim p(x) = \frac{e}{e-1}e^{-x}\{(0,1)\} \qquad F_1(x) = \frac{e}{e-1}(1-e^{-x})$$
$$W = (F_1(C))^n = \left(\frac{e}{e-1}(1-e^{-\sqrt[n]{\alpha}})\right)^n \qquad \alpha_2 = 1 - W$$

Состоятельность $W \to 1$ при $n \to \infty$:

$$e^{-(e^{\frac{1}{n}\ln\alpha})} = e^{-(1-\frac{1}{n}\ln\alpha + o(\frac{1}{n}))}$$
$$\left(\frac{e}{e-1}(1-e^{-1}e^{-\frac{1}{n}\ln\alpha+o(\frac{1}{n})})\right)^n = \left(\frac{e}{e-1}(1-e^{-1}(1-\frac{1}{n}\ln\alpha+o(\frac{1}{n})))\right)^n =$$
$$= \left[\frac{e}{e-1}\left(\frac{e-1}{e} + \frac{\ln\alpha}{ne} + o(\frac{1}{n})\right)\right]^n =$$
$$= \left[1 + \frac{\ln\alpha}{e-1}\frac{1}{n} + o(\frac{1}{n})\right]^n \underset{n\to\infty}{\to} e^{\frac{\ln\alpha}{e-1}} = \alpha^{\frac{1}{e-1}} \ne 1$$

Не является состоятельной

13.2 Проверка сложных гипотез. КОП-критерий отношения правдоподобия

$\xi \sim F(x, \vec{\theta}, \vec{\theta} \in \Theta \subset \mathbb{R}^m$

$H_0 : \vec{\theta} \in \Omega_0$, $H_1 : \vec{\theta} \in \Omega_1$, $\Omega_0 \cap \Omega_1 = \emptyset$, $\Omega = \Omega_0 \cup \Omega_1$

$$l = \frac{L_1}{L_0} \ge C \qquad l(x) = \frac{\sup_{\Omega_1} L}{\sup_{\Omega_0} L}$$
$$G : \ l(x) = \frac{\sup_{\Omega} L}{\sup_{\Omega_0} L} \ge C \qquad P(l \ge C|H_0) = \alpha$$

(сверху в числителе омега без индекса!)

Теорема 13.1. Если H_0 верна, то

$$2\ln l \rightsquigarrow \chi^2(dim\Omega - dim\Omega_0)$$

Пример. $\xi \sim p(x) = \theta e^{-\theta x}\{(0, +\infty)\}$, $\theta > 0$

$H : \theta = \theta_0$, $H_1 : \theta > \theta_0$

$\Omega_0 = \{\theta_0\}$, $\theta_1 = \{\theta : \theta > \theta_0\}$, $\Omega = [\theta_0, +\infty)$

Размерности соответсвенно 0, 1, 1

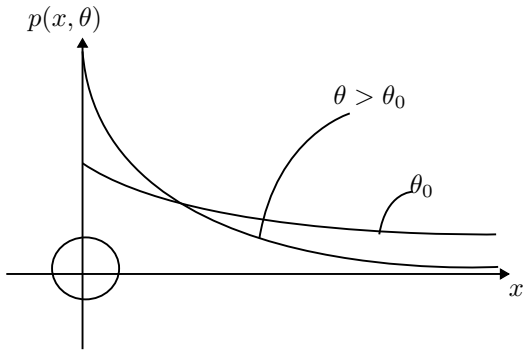
$\underline{n = 1}$

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

$$\sup_{\theta_0} L = \theta_0 e^{-\theta_0 x} \qquad \sup_{\Omega} L = \begin{cases} \theta_0 < \dfrac{1}{x} : \ \dfrac{1}{x} e^{-1} \\ \theta_0 > \dfrac{1}{x} : \ \theta_0 e^{-\theta_0 x} \end{cases}$$

$$l = \begin{cases} \theta_0 < \dfrac{1}{x} : \ \dfrac{1}{x} \dfrac{e^{-1}}{\theta_0 e^{-\theta_0 x}} \\ \theta_0 > \dfrac{1}{x} : \ 1 \end{cases} \geq C$$

Давим логикой

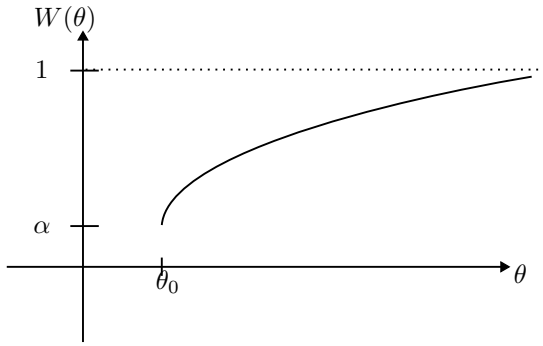


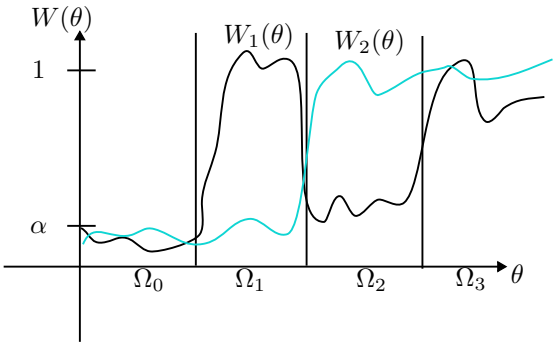
Рассматривая записимость распределения от θ, видим что у альтернативной гипотезы вероятность попасть в окрестность нуля больше, оттлакиваемся от этого

$$G : x \leq C \qquad P(x \leq C | H_0) = \alpha$$

$$\int_0^C \theta_0 e^{-\theta_0 x} dx = 1 - e^{-\theta_0 C} = \alpha \qquad C = -\frac{1}{\theta_0} \ln(1 - \alpha)$$

$$W(\theta) = P(x \leq C | H_1) = \int_0^C \theta e^{-\theta x} dx = 1 - e^{-\theta C} = 1 - e^{\frac{\theta}{\theta_0} \ln(1-\alpha)}$$





13.3 Проверка гипотез о параметрах нормальной модели

$\xi \sim N(\theta_1, \theta_2^2)$, $\theta_1 \in R$, $\theta_2 > 0$

13.3.1 Проверка гипотез о значениях параметров

a) $H_0 : \theta_1 = a$, $H_1 : \theta_1 \neq a$; $\theta_1 > a$; $\theta_1 < a$

b) $H_0 : \theta_2^2 = b$, $H_1 : \theta_2^2 \neq b$; $\theta_2^2 > b$; $\theta_2^2 < b$

Теорема Фишера:

$$\frac{\bar{x} - a}{\sigma}\sqrt{n} \sim N(0,1) \qquad \frac{S^2(n-1)}{\sigma^2} \sim \chi^{2(n-1)} \qquad \frac{\bar{x} - a}{S}\sqrt{n} \sim t(n)$$

Пример. $\xi \sim N(\theta, 0,001)$, $H_0 : \theta_1 = 43$, $H_1 : \theta \neq 43$, $n = 50$, $\bar{x} = 42.972$,

$\alpha = 0.05$

$$\Delta = \frac{\bar{x} - 43}{.1}\sqrt{50} \qquad G : |\Delta| \geq C \qquad P(|\Delta| \geq C|H_0) = \alpha$$

$$2P(\Delta \geq C|H_0) = \alpha \qquad C = u_{1-\frac{\alpha}{2}}$$

$$G : |\Delta| \geq u_{1-\alpha/2} = 1.96$$

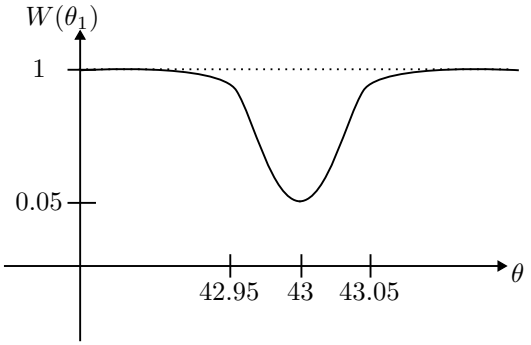
$$G : \begin{bmatrix} \bar{x} \geq 43.027 \\ \bar{x} \leq 42.9723 \end{bmatrix}$$

$\bar{x} \in G$, отвергаем H_0

$$W(\theta_1) = P(\vec{x}_n \in G|H_1) = P(\bar{x} \geq 43.027|H_1) + P(\bar{x} \leq 42.9723|H_1) =$$

$$= P(\frac{\bar{x} - \theta_1}{0.1}\sqrt{50} \geq \underbrace{\frac{43.027 - \theta_1}{0.1}\sqrt{50}}_{a_1}) + P(\frac{\bar{x} - \theta_1}{0.1}\sqrt{50} \leq \underbrace{\frac{42.9723 - \theta_1}{0.1}\sqrt{50}}_{a_2}) =$$

$$= \int_{a_1}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx + \int_{-\infty}^{a_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$



Парктические методы:

1. p-value:

$$\text{p-value} = P(|\Delta| \geq |\tilde{\Delta}||H_0) =$$

$$(\Delta = \frac{\bar{x} - 43}{0.1}\sqrt{50} \qquad \tilde{\Delta} = -1.98)$$

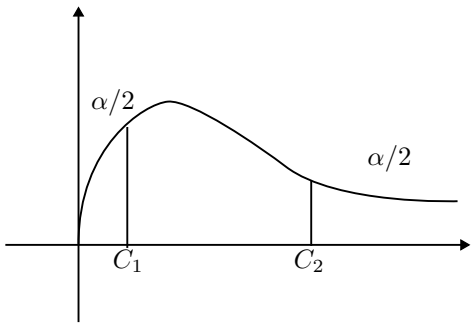
$$= 2P(\Delta \geq 1.98) = 2\int_{1.98}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0.0477$$

2. Доверительный интервал:

$$\beta = 1 - \alpha = 0.95$$
$$-1.96 \le \frac{\bar{x} - a}{0.1}\sqrt{50} \le 1.96$$
$$I = (42.9443; 42.9997) \qquad a \in I$$

$a = 43$, H_0 отвергаем

Пример. $\xi \sim N(\theta_1, \theta_2^2)$, $H_0 : \theta_2^2 = 400$, $H_1 : \theta_2^2 \neq 400$, $S^2 = 784$, $n = 20$, $\Delta = \frac{S^2(n-1)}{\theta_2^2}$



$$P(\Delta \le C_1 | H_0) + P(\Delta \ge C_2 | H_0) = \alpha$$
$$C_1 = \chi^2_{\alpha/2}(n-1) \qquad C_2 = \chi^2_{1-\alpha/2}(n-1)$$
$$\chi^2_{0.025}(19) = 8.907 \qquad \chi^2_{0.975}(19) = 32.87$$
$$\tilde{\Delta} = 37.24$$

Попали в критическую область, отвергаем H_0.

$$\text{p-value} = P(\Delta \ge \tilde{D} | H_0) = 0.0074$$

Оставляем гипотезу если $\frac{\alpha}{2} \le$ p-value $\le 1 - \frac{\alpha}{2}$, в данном случае не отвергаем.

13.3.2 Проверка гипотез о равенстве параметров

а) $\xi \sim N(\varphi, \theta_2^2)$, $\eta \sim N(\psi, \theta_2^2)$, независимы, θ_2^2 одинакова, но неизвестна

$H_0 : \varphi = \psi$, $H_1 : \varphi \neq \psi, \varphi > \psi, \varphi < \psi$

$$x_n \qquad y_m$$

$$\frac{\bar{x}-\varphi}{\theta_2}\sqrt{n} \sim N(0,1) \qquad \frac{\bar{y}-\psi}{\theta_2}\sqrt{m} \sim N(0,1)$$

$$\bar{x}-\varphi \sim N(0, \frac{\theta_2^2}{n}) \qquad \bar{y}-\psi \sim N(0, \frac{\theta_2^2}{m})$$

$$\Delta = \bar{x}-\varphi-(\bar{y}-\psi) \sim N(0, \frac{\theta_2^2}{n}+\frac{\theta_2^2}{m})$$

$$\tilde{\xi} \sim N(\vec{0}, K) \qquad L\vec{\xi} \sim N(\vec{0}, LKL^T)$$

$$\Delta \sim N(0, \theta_2^2 \frac{n+m}{nm})$$

$$\frac{\bar{x}-\bar{y}-(\varphi-\psi)}{\theta_2\sqrt{\frac{n+m}{nm}}} \sim N(0,1)$$

$$\frac{S_x^2(n-1)}{\theta_2^2} \sim \chi^2(n-1) \qquad \frac{S_y^2(m-1)}{\theta_2^2} \sim \chi^2(m-1)$$

$$\frac{S_x^2(n-1)-S_y^2(m-1)}{\theta_2^2} \sim \chi^2(n+m-2)$$

Теорема 13.2. Если H_0 верна, то

$$\frac{\bar{x}-\bar{y}}{\sqrt{\frac{n+m}{nm}}\sqrt{\frac{S_x^2(n-1)+S_y^2(m-1)}{n+m-2}}} \sim t(n+m-2)$$

Асимптотическое приближение $n \to \infty$, $m \to \infty$

$$\frac{\bar{x}-\varphi}{\theta_2'}\sqrt{n} \sim N(0,1) \qquad \frac{\bar{y}-\psi}{\theta_2''}\sqrt{m} \sim N(0,1)$$

По лемме Слуцкого:

$$\frac{\bar{x}-\varphi}{S_x}\sqrt{n} \rightsquigarrow N(0,1) \qquad \frac{\bar{y}-\psi}{S_y}\sqrt{m} \rightsquigarrow N(0,1)$$

$$\bar{x}-\varphi-(\bar{y}-\psi) \rightsquigarrow N(0, \frac{S_x^2}{n}+\frac{S_y^2}{m})$$

Теорема 13.3. Если H_0 верна ($\varphi = \psi$), то

$$\frac{\bar{x}-\bar{y}}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} \rightsquigarrow N(0,1)$$

b) $\xi \sim N(a, \varphi^2)$, $\eta \sim N(b, \psi^2)$ независ.

$H_0 : \varphi^2 = \psi^2$, $H_1 : \varphi^2 \neq \psi^2, \varphi^2 > \psi^2, \varphi^2 < \psi^2$

$$\begin{matrix} & x_n & y_m \\ & \frac{S_x^2(n-1)}{\varphi} \sim \chi^2(n-1) & \frac{S_y^2(m-1)}{\psi^2} \sim \chi^2(m-1) \end{matrix}$$

$$\frac{S_x^2}{\varphi^2}\frac{\psi^2}{S_y^2} \sim F(n-1, m-1)$$

Теорема 13.4. Если H_0 верна, то

$$\frac{S_x^2}{S_y^2} \sim F(n-1, m-1)$$

Пример. $n = 20$, $n_1 = 10$, $n_2 = 10$, A/B тестирование

1 группа сразу: $12, 17.5, 16, 9.5, 15, 15.5, 13, 16.5, 14.5, 17$

2 группа сразу: $16, 12, 15.5, 15, 14, 18.5, 18, 13, 17.5, 17$

$$\begin{matrix} \bar{x} = 14.65 & S_x = 2.49 \\ \bar{y} = 15.65 & S_y = 2.17 \end{matrix}$$

1. норм. закон распределения
2. дисперсии равны

$H_0 : M\xi = M\eta$, $H_1 : M\xi \neq M\eta$

$$\frac{\bar{x} - \bar{y}}{\sigma}\sqrt{\frac{nm}{n+m}} \sim t(n+m-2) \qquad \sigma = \sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}$$

$$\tilde{\Delta} = -1.044$$

$$\text{p-value} = P(|\Delta| \geq |\tilde{\Delta}||H_0) = 2P(\Delta \geq 1.044) = 2\int_{1.044}^{+\infty} q(t)dt = 0.31$$

Нет оснований отвергать гипотезу

1) Критерий Колмогорова для двух выборок p-value = 0.3, p-value = 0.68

2) $H_0 : D\xi = D\eta$, $H_1 : D\xi \neq D\eta$

$$\Delta = \frac{S_x^2}{S_y^2} \sim F(n-1, m-1)$$

$$\tilde{\Delta} = 1.147 \qquad g(t) : F(9, 9)$$

$$\text{p-value} = P(\Delta \geq \tilde{\Delta}|H_0) = \int_{1.147}^{\infty} g(t)dt = 0.42$$

3) Проверка гипотезы случайности:

$$\Delta = \frac{I_n - \frac{n(n-1)}{4}}{\sqrt{\frac{n^3}{36}}} \rightsquigarrow N(0,1)$$

Для y: $I_m = 19$, $\tilde{\Delta} = -0.626$, p-value $= P(|\Delta| \geq |\tilde{\Delta}|) = 2P(\Delta \geq 0.626) = 0.531$, нет оснований отвергать гипотезу.

Бутстрап - непараметрический бутстрап + метот доверительных инт.

$$h = M\xi - M\eta \qquad \tilde{h} = \bar{x} - \bar{y} = -1 \qquad \Delta = \tilde{h} - h$$

$$\begin{cases} \vec{x}_n \to \vec{x}_n^* \\ \vec{y}_m \to \vec{y}_m^* \end{cases} \to \Delta_i^* = \bar{x}^* - \bar{y}^* - \tilde{h}$$

$$K_1 = [N\frac{1-\beta}{2}] \qquad K_2 = [N\frac{1+\beta}{2}]$$

$$P(\Delta^*_{(k_1)} < \Delta^* < \Delta^*_{(k_2)}) \approx \beta$$

$$\Delta^*_{(k_1)} < \Delta^* < \Delta^*_{(k_2)} \to (-1, 2)$$

$0 \in I$, нет оснований отвергнуть гипотезу

13.4 Множественная проверка гипотез

$H_{01}, \ldots, H_{0m}$, всем соответсвует α, $P(A_i) = \alpha$

$$P(A_1 + \cdots + A_m) = \bar{P}(\bar{A}_1 \ldots \bar{A}_m) =$$
$$= 1 - P(\bar{A}_1 \ldots \bar{A}_m) = 1 - (1-\alpha)^m \approx 1 - (1 - \alpha m) = \alpha m$$

13.5 Метод Бонферрони

Проверяем гипотезы на уровне α/m, адекватно при $m \leq 5$

13.6 Метод Холма-Бонферрони

Считаем p-value для всех гипотиз и упорядовиваем

$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$

$p_{(1)}$ сравниваем с α/m:

если $p_{(1)} \geq \alpha/m$, то нет оснаваний отвернуть все гипотезы

если $p_{(1)} < \alpha/m$, отвергаем $H_{0(1)}$, $p_{(2)}$ сравн. с $\alpha/(m-1)$, ..., $p_{(m)}$ сравн. с α

14 Исследование зависимостей

14.1 Методы статистики

1. Традиционный анализ (ANCOVA)

Обнаружение зависимостей в количественных данных

2. Дисперсионный анализ (ANOVA)

   Обнаружение зависимости в качественных данных

3. Регрессионный анализ

   Обнаружение и определение формы зависимости

$\vec{\xi} = (\xi_1, \ldots, \xi_k)$ - факторы (регрессоры)

η - отклик

$\eta = f(\vec{x}) + \varepsilon(\vec{x})$ - апроксимация зависмости

$\vec{x}$ - значение $\vec{\xi}$, $\varepsilon(\vec{x})$ - сл. вел.

Базис $\Psi_1^{(\vec{x})}, \ldots, \Psi_p^{(\vec{x})}$, $f(\vec{x}) = \sum_{m=1}^{p} \beta_m \Psi_m(\vec{x})$

Выбор базиса зависит от цели:

1. Сглаживание данных
2. Факторный анализ: $\Psi : 1, \underbrace{x_1, \ldots, x_k}_{\text{факторы}}, \underbrace{x_1 x_2, \ldots, x_i^2}_{\text{взаим. влиян.}}$

Пример. Популяция $\xi = x$, $\eta = \dot{x}$

1. развитие $\eta = kx + \varepsilon_1$, вероят. модель.
2. ограниченность ресурсов $\eta = kx(a - x) + \varepsilon_2$
3. агрессивность внешней сред $\eta = kx(a - x)(x - b) + \varepsilon_3$

$\varepsilon(\vec{x}) \sim N(0, \sigma^2)$

$\beta_1 = M\eta$ при $x_i = 0$ - среднее влияние внешних факторов Вероятностная модель для одного отклика:

$$\eta = \sum_{m=1}^{p} \beta_m \Psi_m(\vec{x}) + \varepsilon(\vec{x}) \text{ - линейная регрессия}$$

Выборка $(y_i, \vec{x}_i)$ $i = 1, \ldots, n$

$$\Psi = \begin{pmatrix} \Psi_1(\vec{x}_1) & \ldots & \Psi_p(\vec{x}_1) \\ \ldots & \ldots & \ldots \\ \Psi_1(\vec{x}_n) & \ldots & \Psi_p(\vec{x}_n) \end{pmatrix} \text{ - матрица наблюдений}$$

$\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}$ - вектор отклика - сл. вел. $\qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ - значение век. откл.

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1(\vec{x}_1) \\ \vdots \\ \varepsilon_n(\vec{x}_n) \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \text{ - вектор ошибок - сл. вел.}$$

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_1 \\ \vdots \\ \tilde{\beta}_n \end{pmatrix} \qquad e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \text{ - значение } \varepsilon$$

Вероятностная модель для вектора отклика

$$\eta = \Psi\beta + \varepsilon$$

Оценка по выборке

$$Y = \Psi\tilde{\beta} + e$$

$\varepsilon \sim N(\vec{0}, \sigma^2 E)$ - сильно регулярная, открытое множество $\sigma > 0$

$$L = \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{-\frac{1}{2\sigma^2} e^T e} \to \max$$

$$e^T e \to \min$$

$$(Y - \Psi\tilde{\beta})^T (Y - \Psi\tilde{\beta}) \to \min$$

$$Y - \Psi\tilde{\beta} \perp \Psi t \qquad (\Psi t)^T (Y - \Psi\tilde{\beta}) = 0$$

$$t^T (\underbrace{\Psi^T Y - \Psi^T \Psi\tilde{\beta}}_{0}) = 0 \ \forall t$$

$F = \Psi^T\Psi$ - полож. определённая (предположение)

F^{-1} - матрица Фишера

$$\tilde{\beta} = F^{-1}\Psi^T Y$$

Свойства $\tilde{\beta}$: сост., асим. несмещ., асим. нормальн., асим. эффект (см. Замечание после примера)

Пример. Выборка: $x = (1, 2, 3, 4, 5)$, $y = (6, 5, 4, 5, 6)$

Базис берём x, $\frac{1}{x}$

$$\Psi = \begin{pmatrix} 1 & 1 \\ 2 & \frac{1}{2} \\ 3 & \frac{1}{3} \\ 4 & \frac{1}{4} \\ 5 & \frac{1}{5} \end{pmatrix} \qquad Y = \begin{pmatrix} 6 \\ 5 \\ 4 \\ 5 \\ 6 \end{pmatrix}$$

$$F = \Psi^T\Psi = \begin{pmatrix} 55 & 5 \\ 5 & 1.464 \end{pmatrix} \qquad F^{-1} = \begin{pmatrix} 0.0264 & -0.09 \\ -0.09 & 0.991 \end{pmatrix}$$

$$\tilde{\beta} = F^{-1}\Psi^T Y = \begin{pmatrix} 0.954 \\ 5.153 \end{pmatrix}$$

$$y = 0.954x + \frac{5.153}{x} + e$$

Распределение $\tilde{\beta}$ (как случайной вел)

$$\tilde{\beta} = F^{-1}\Psi^T\eta = F^{-1}\Psi^T(\Psi\beta + \varepsilon) = F^{-1}\underbrace{\Psi^T\Psi}_{F}\beta + F^{-1}\Psi^T\varepsilon$$

$$\underbrace{F^{-1}\Psi^T}_{L}\varepsilon \sim N(\vec{0}, LEL^T) = N(\vec{0}, F^{-1})$$

$$\tilde{\beta} \sim N(\beta, \sigma^2 F^{-1})$$

Замечание (Уточнение свойств $\tilde{\beta}$)**.**

1. состоятельная
2. несмещ
3. нормальная
4. эффективная (trust me bro, без доказательства)

Пример.

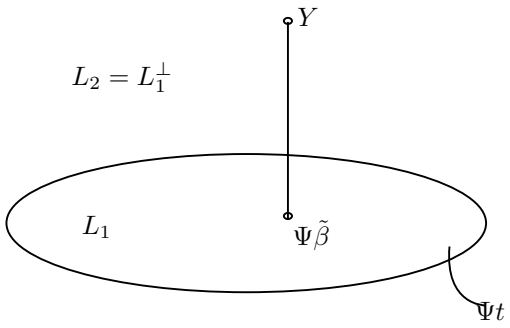
$$\tilde{\beta} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \sigma^2 \begin{pmatrix} 0.0264 & -0.09 \\ -0.09 & 0.991 \end{pmatrix}\right)$$

$$D\tilde{\beta}_1 = 0.0264\sigma^2 \qquad D\tilde{\beta}_2 = 0.991\sigma^2 \qquad cov(\tilde{\beta}_1, \tilde{\beta}_2) = \sigma^2(-0.09)$$

14.1.1 Оценка парметра сигма квадрат

RSS - остаточная сумма квадратов (residual sum of squares)

$RSS = e^T e$ - числа



$$\eta_{L_1} = \Psi\tilde{\beta} \qquad \eta_{L_2} = \eta - \eta_{L_1}$$
$$\eta = \Psi\beta + \varepsilon$$
$$(\Psi\beta + \varepsilon)_{L_1} = \Psi\beta + \varepsilon_{L_1} = \Psi\tilde{\beta}$$
$$(\Psi\beta + \varepsilon)_{L_2} = \varepsilon_{L_2} = \eta_{L_2} RSS = |\eta_{L_2}|^2 \text{ - как сл. вел}$$
$$RSS = |\varepsilon_{L_2}|^2$$

Теорема о процекциях: $\varepsilon \sim N(\vec{0}, \sigma^2 E)$, L_1 и L_2 лин. подпр. $L_2 = L_1^{\perp}$, тогда ε_{L_1}, ε_{L_2} норм. распр., незав и $\frac{|\varepsilon_{L_1}|^2}{\sigma^2} \sim \chi^2(dim\, L_1)$, $\frac{|\varepsilon_{L_2}|^2}{\sigma^2} \sim \chi^2(dim\, L_2)$

$$\frac{RSS}{\sigma^2} \sim \chi^2(n-p)$$
$$\Psi\beta + \varepsilon_{L_1} = \Psi\tilde{\beta} \mid \cdot F^{-1}\Psi^T$$
$$F^{-1}\Psi^T\Psi\beta + F^{-1}\Psi^T\varepsilon_{L_1} = F^{-1}\Psi^T\Psi\tilde{\beta}$$
$$\tilde{\beta} - \beta = F^{-1}\Psi^T\varepsilon_{L_1}$$

RSS и $\tilde{\beta} - \beta$ независ. как случайные величины

$$\tilde{\sigma^2} = \frac{RSS}{n-p}$$
$$M[\tilde{\sigma^2}] = M[\frac{RSS}{n-p}] = \frac{\sigma^2}{n-p} M[\frac{RSS}{\sigma^2}] = \sigma^2 \quad \text{несмещ.}$$
$$D[\tilde{\sigma^2}] = \frac{\sigma^4}{(n-p)^2} D[\frac{RSS}{\sigma^2}] = \frac{2\sigma^4}{n-p} \underset{n\to\infty}{\to} 0 \quad \text{сост.}$$

Пример.

$$Y = \Psi\tilde{\beta} + e$$

$$e = \begin{pmatrix} 6 \\ 5 \\ 4 \\ 5 \\ 6 \end{pmatrix} - \Psi \begin{pmatrix} 0.954 \\ 5.153 \end{pmatrix} = \begin{pmatrix} -0.107 \\ 0.5155 \\ -0.58 \\ -0.104 \\ 0.2 \end{pmatrix}$$

$$RSS = e^T e = 0.664 \qquad \tilde{\sigma^2} = \frac{0.664}{5-2} = 0.221$$

Прогноз:

$$x_0 = 2.5 \qquad \tilde{y}_0 = \Psi(x_0)\tilde{\beta}$$

$$\Psi(x_0) = (\Psi_1(x_0)\ \Psi_2(x_0)) = (2.5\ \frac{1}{2.5})$$

$$\tilde{y}_0 = (2.5\ \frac{1}{2.5}) \begin{pmatrix} 0.954 \\ 5.153 \end{pmatrix} = 4.45$$

14.1.2 Доверительный интервал

$$\eta_0 = \underbrace{\Psi(x_0)}_{\Psi_0}\beta + \underbrace{\varepsilon(x_0)}_{\varepsilon_0}$$

$$\tilde{y_0} \sim N(\Psi_0\beta, \sigma^2\Psi_0 F^{-1}\Psi_0^T)$$

$$\eta_0 \sim N(\Psi_0\beta, \sigma^2)$$

$$\tilde{y_0} - \eta_0 \sim N(0, \sigma^2(1 + \Psi_0 F^{-1}\Psi^T))$$

$$\begin{cases} \dfrac{\tilde{y_0} - \eta_0}{\sigma\sqrt{1 + \Psi_0 F^{-1}\Psi^T}} \sim N(0,1) \\ \dfrac{RSS}{\sigma^2} \sim \chi^2(n-p) \end{cases} \text{ независ.}$$

$$\frac{\dfrac{\tilde{y_0} - \eta_0}{\sqrt{1 + \Psi_0 F^{-1}\Psi_0}}}{\sqrt{\dfrac{RSS}{n-p}}} \sim t(n-p)$$

$$t_{\frac{1-\beta}{2}}(n-p) < \frac{\tilde{y_0} - \eta_0}{\sqrt{1 + \Psi_0 F^{-1}\Psi_0}}\sqrt{\frac{n-p}{RSS}} < t_{\frac{1+\beta}{2}}(n-p)$$

$$\tilde{y_0} - \Delta < \eta_0 < \tilde{y_0} + \Delta$$

$$\Delta = t_{\frac{1+\beta}{2}}(n-p)\frac{\sqrt{RSS(1 + \Psi_0 F^{-1}\Psi_0)}}{\sqrt{n-p}}$$

Пример.

$$\beta = 0.95 \qquad t_{0.975}(3) = 3.182$$
$$\Delta = 3.182\sqrt{\frac{(1+0.144)0.664}{3}} = 1.6$$
$$\tilde{y_0} = 4.45$$
$$(2.85, 6.05)$$

14.2 Проверка гипотез

14.2.1 Проверка значимости коэфф. регрессии

$H_0 : \beta_i = 0$, $H_1 : B_i \neq 0$

$$\tilde{\beta_i} \sim N(\beta_i, \sigma^2 F_{ii}^{-1})$$
$$\frac{\tilde{\beta_i} - \beta_i}{\sigma\sqrt{F_{ii}^{-1}}} \sim N(0,1) \qquad \frac{RSS}{\sigma^2} \sim \chi^2(n-p)$$
$$\frac{\tilde{\beta_i} - \beta}{\sqrt{RSS \cdot F_{ii}^{-1}}\sqrt{n-p}} \sim t(n-p)$$

Теорема 14.1. Если H_0 верна, то

$$\Delta = \frac{\tilde{\beta_i}}{\sqrt{RSS \cdot F_{ii}^{-1}}}\sqrt{n-p} \sim t(n-p)$$

Пример.

$$\tilde{\beta_1} = 0.954 \qquad RSS = 0.664 \qquad F_{11}^{-1} = 0.0264$$
$$\tilde{\Delta} = 12.48 \qquad \Delta \sim t(3)$$
$$\text{p-value} = P(|\Delta| \geq |\tilde{\Delta}| \big| H_0) = 2P(\Delta \geq 12.48) = 2\int_{12.48}^{\infty} g(t)dt = 0.0011$$

Отвергаем H_0, β_1 значимый

14.2.2 Проверка равенства коэфф.

$H_0 : \ \beta_i = \beta_j$, $H_1 : \ \beta_i \neq \beta_j, \beta_i > \beta_k, \beta_i < \beta_j$

$$\tilde{\beta_i} \sim N(\beta_i, \sigma^2 F_{ii}^{-1})$$
$$\tilde{\beta_j} \sim N(\beta_j, \sigma^2 F_{jj}^{-1})$$
$$\frac{\tilde{\beta_i} - \beta_i - (\tilde{\beta_j} - \beta_j)}{\sigma} \sim N(0, F_{ii}^{-1} + F_{ij}^{-1})$$

Теорема 14.2. Если H_0 верна, то

$$\Delta = \frac{\tilde{\beta}_i - \tilde{\beta}_j}{\sqrt{RSS(F_{ii}^{-1} + F_{jj}^{-1})}}\sqrt{n-p} \sim t(n-p)$$

Пример. $H_0 : \beta_1 = \beta_2$, $H_1 : \beta_1 \neq B_2$

$$\tilde{\Delta} = -8.85 \qquad \Delta \sim t(3)$$

$$\text{p-value} = 2P(|\Delta| \geq |\tilde{\Delta}|) = 2P(\delta \geq 8.85) = 2\int_{8.85}^{\infty} g(t)dt = 0.003$$

Различие β_i значимо

Замечание. Нормировка:

$$y_i = \frac{x_i - x_{i\,min}}{x_{i\,max} - x_{i\,min}} \qquad z_i \in [0,1]$$

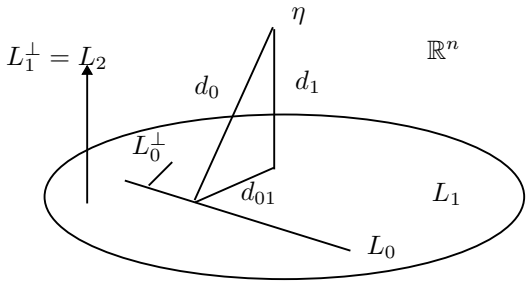
14.3 Сравнение регрессий

14.3.1 Вспомогательная задача

$\eta = \gamma + \varepsilon \qquad \varepsilon \sim N(\vec{0}, \sigma^2 E)$

γ - лежит в подпр L_1, $L_0 \subset L_1$

$H_0 : \gamma \in L_0$, $H_1 : \gamma \in L_1/L_0$



$$d_1 = |\eta - \eta_{L_1}|^2 \qquad d_0 = |\eta - \eta_{L_0}|^2 \qquad d_{01} = d_0 - d_1$$

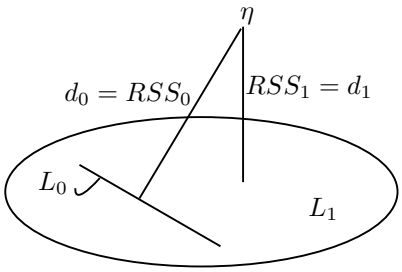
$$\mathbb{R}^n = L_0 \oplus L_0^{\perp} \oplus L_2$$

Теорема о проекц: $\varepsilon \sim N(0, \sigma^2 E)$ и $L_1 \oplus L_2 \oplus L_3$, тогда $\varepsilon_{L_1}, \varepsilon_{L_2}, \varepsilon_{L_3}$ - независ и $\frac{|\varepsilon_{L_i}|^2}{\sigma^2} \sim \chi^2(dim\, L_i)$

$$\eta_{L_1} = \gamma + \varepsilon_{L_1}$$
$$H_0 : \eta_{L_0} = \gamma + \varepsilon_{L_0}$$
$$d_1 = |\gamma + \varepsilon - \gamma - \varepsilon_{L_1}|^2 = |\varepsilon - \varepsilon_{L_1}|^2$$
$$d_0 = |\gamma + \varepsilon - \gamma - \varepsilon_{L_0}|^2 = |\varepsilon - \varepsilon_{L_0}|^2$$
$$\frac{d_1}{\sigma^2} \sim \chi^2(n - n_1) \qquad \frac{d_{01}}{\sigma^2} \sim \chi^2(n_1 - n_0)$$
$$\Delta = \frac{\dfrac{d_0 - d_1}{n_1 - n_0}}{\dfrac{d_1}{n - n_1}} \sim F(n_1 - n_0, n - n_1)$$

14.3.2 Сравнение вложенных регрессий

$$\eta = \Psi\beta + \varepsilon$$
$$H_0 : \beta_i = 0, i \in I \qquad H_1 : \exists i \in I : \beta_i = 0$$



$$\Delta = \frac{\dfrac{RSS_0 - RSS_1}{p_1 - p_0}}{\dfrac{RSS_1}{n - p_1}} \sim F(p_1 - p_0, n - p_1)$$

14.3.3 Сравнение невложенных регрессий

$$\Psi_1(x) \dots \Psi_{p_1}(x) \qquad \varphi_1(x) \dots \varphi_{p_2}(x)$$

Объединяем в общий базис $\eta = \Psi\beta + \varepsilon$

$$\begin{aligned} A: \quad & H_0 : \beta_i \text{ для 2го базиса } = 0 \\ & H_1 : \exists i : \ \beta_i \neq 0 \\ B: \quad & H_0 : \beta_i \text{ для 1го базиса } = 0 \\ & H_1 : \exists i : \ \beta_i \neq 0 \end{aligned}$$

Если верна H_0, говорим $A(B) <$ длин.

1. $A <$ длин $< B$, A - лучший базис
2. длин $< A <> B$ - лучший объединённый базис
3. $A <> B <$ длин - нельзя ничего сказать

Пример.

$$\{x, \frac{1}{x}\} \qquad \{x, \frac{1}{x^2}\}$$

$$RSS = 0.664 \qquad RSS = 1.93$$

$$\text{длин } \{\frac{1}{x^2}, x, \frac{1}{x}\}$$

$$\Psi^T\Psi = F = \begin{pmatrix} 1.08 & 2.28 & 1.19 \\ 2.28 & 55 & 5 \\ 1.19 & 5 & 1.45 \end{pmatrix} \qquad \tilde{\beta} = F^{-1}\Psi^T\Psi = \begin{pmatrix} -0.65 \\ 0.92 \\ 5.77 \end{pmatrix}$$

$$RSS = 0.6471$$

$$A: \ \{x, \frac{1}{x}\} \text{ vs } \{\frac{1}{x^2}, x, \frac{1}{x}\}$$

$$\tilde{D}_A = \frac{(0.664 - 0.6471)/(3-2)}{0.6471/(5-3)} = 0.052 \qquad \tilde{\Delta}_A \sim F(1,2)$$

$$\text{p-value} = P(\Delta \geq \tilde{\Delta}_A | H_0) = \int_{0.052}^{\infty} q(t)dt = 0.84$$

$$B: \ \{x, \frac{1}{x^2}\} \text{ vs } \{\frac{1}{x^2}, x, \frac{1}{x}\}$$

$$\tilde{D}_B = \frac{(1.93 - 0.6471)/(3-2)}{0.6471/(5-3)} = 3.965 \qquad \tilde{\Delta}_B \sim F(1,2)$$

$$\text{p-value} = P(\Delta \geq \tilde{\Delta}_B | H_0) = \int_{3.965}^{\infty} q(t)dt = 0.185$$

$$A < \text{длин} \qquad B < \text{длин}$$

14.3.4 Проверка значимости всей регрессии (факторный анализ)

$$\eta = \Psi\beta + \varepsilon \qquad \{1, \dots\}$$
$$H_0 : \beta_i = 0,\ i \neq 1 \text{ (не 0 только при 1)}$$
$$H_1 : \beta_i \neq 0,\ i \neq 1 \text{ (хотя бы один)}$$
$$\eta = \beta_1 - \varepsilon$$
$$e^T e = \sum_{i=1}^{n} (y_i - \beta_1)^2 \to \min \qquad \tilde{B}_1 = \bar{y}$$
$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = TSS \text{ (total sum of squares)}$$
$$\Delta = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} \sim F(p-1, n-p)$$

Определение 14.1 (Коэффициент детерминации)**.**

$$R^2 = \frac{TSS - RSS}{TSS}$$

- доля дисперсии отклика, которая объясняется уравнением линейной регресси

Замечание.

Вся регрессия значима, но при этом все коэфф. не явл. значимыми - значит суммарно значимы

Регрессия незначима, но коэфф. значимы - вероятно не учли множественную проверку гипотез и откинули лишнее

Пример. $n = 15$, x_1 БРС (24), x_2 семестровая (16), y экзамен

19	22	14	17	21	24	23	22	15	24	15	18	15	20	21
9	15	10	3	6	8	13	4	3	10	9	3	13	3	16
6	7	5	3	6	8	8	5	6	8	4	3	5	5	5

$$\eta = \beta_1 + \beta_2\xi_1 + \beta_3\xi_2 + \varepsilon$$

$$\Psi = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 19 & 22 & \dots & 21 \\ 9 & 15 & \dots & 16 \end{pmatrix}^T \qquad T = \begin{pmatrix} 6 & 7 & \dots & 5 \end{pmatrix}^T$$

$$F^{-1} = \begin{pmatrix} 2.31 & -0.11 & -0.011 \\ -0.11 & 0.0062 & -0.001 \\ -0.011 & -0.001 & 0.0036 \end{pmatrix} \qquad \tilde{B} = F^{-1}\Psi^T\Psi = \begin{pmatrix} -0.3 \\ 0.25 \\ 0.13 \end{pmatrix}$$

$$y = -0.3 + 0.25x_1 + 0.13x_2 + e$$

$$RSS = e^T e = 18.54$$

Статистический аналзи модели:

1. значимость коэффициентов

$$H_0 : \beta_i = 0 \qquad H_1 : \beta_i \neq 0$$

$$\Delta = \frac{\tilde{B}_i}{\sqrt{RSS \cdot F_{ii}^{-1}}}\sqrt{n-p} \sim t(n-p)$$

$$\tilde{\Delta}_1 = \frac{-0.3}{\sqrt{18.54 \cdot 2.31}}\sqrt{12} = -0.16$$

$$\text{p-value}_1 = P(|\Delta| \geq |\tilde{\Delta}||H_0) = 2\int_{0.16}^{\infty} q(t)dt = 0.876 \qquad \beta_1 \text{ не знач.}$$

$$\beta_1 \text{ знач.} \qquad \beta_1 \text{ не знач.}$$

2. значимость всей регрессии

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 37.33 \qquad \bar{y} = 5.67$$

$$R^2 = \frac{37.33 - 18.54}{37.33} = 0.503$$

$$\Delta = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} \sim F(p-1, n-p) \qquad \tilde{\Delta} = 6.072$$

$$\Delta \sim F(2, 12) \qquad \text{p-value} = P(\delta \geq \tilde{\Delta}|H_0) = 0.015$$

$$H_0 : \beta_1 \neq 0, \beta_i = 0 \qquad \text{отвергаем}$$

Регресия значима

3. Прогноз $x_1 = 15$, $x_2 = 8$

$$\tilde{y} = -0.3 + \dots = 4.49$$

Доверительный интервал:

$$\Delta = t_{\frac{1+\beta}{2}}(n-p)\sqrt{1+\Psi_0 F^{-1}\Psi_0^T}\sqrt{\frac{RSS}{n-p}}$$

$$\beta = 0.95 \qquad \Psi_0 = (1\ 15\ 8)$$

$$t_{0.975}(12) = 2.18 \qquad \delta = 2.94$$

$$(1.55; 7.43)$$

4. Проверка предположений регрессии:

$$\varepsilon \sim N(0, \sigma^2)$$

$$e = (0.33; -0.21; \ldots; -2.09)$$

(a) Гипотеза случайности e

$$\Delta = \frac{I_n - \frac{n(n-1)}{4}}{\sqrt{\frac{n^3}{36}}} \rightsquigarrow N(0,1)$$

$$\tilde{\Delta} = 0.67 \qquad I_n = 59 \qquad \text{p-value} = 0.52$$

(b) Гипотеза нормальности ε

$$\tilde{\sigma}^2 = \frac{RSS}{n-p} = \frac{18.54}{12} = 1.545$$

Критерий Колмогорова

$$\Delta = \sqrt{n}\sup_R |\tilde{F}(x) - F(x, 0, \sigma^2)|$$

Параметрический бутстрап:

$$\tilde{\sigma}^2 = \frac{1}{n-1}\sum (e_i - \bar{e})^2 = 1.15 \ \rightarrow \ N(0, 1.15)$$

$$\vec{x}_1 5^* \rightarrow \tilde{F}^*(x),\ \tilde{\sigma}^{2*} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

$$\Delta_i^* = \sqrt{15}\sup_R |\tilde{F}(x) - F(x, 0, \tilde{\sigma}^{2*})|$$

$N = 50000$, вар. ряд $\Delta^*_{(1)} \ldots \Delta^*_{(N)}$, m - колво $\Delta^*_{(k)} \geq \tilde{\Delta}$, p-value $= \frac{m}{N}$

$$\text{p-value} = 0.722$$

14.4 Проверка адекватности модели

$$fix\,\vec{x} \qquad \vec{\xi} = (\xi_1, \ldots, \xi_k)$$

$$y_1, y_2, \ldots, y_l \qquad \eta = f(\vec{x}) + \varepsilon \quad f = \sum \beta_i \Psi_i(\vec{x})$$

$$\tilde{\sigma}_1^2 = \frac{1}{l-1} \sum_{i=1}^{l} (y_i - \bar{y})^2$$

$$\varepsilon \sim N(0, \sigma_2^2)$$

$$\frac{\tilde{\sigma}_1^2 (l-1)}{\sigma_1^2} \sim \chi^2(l-1)$$

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_2^2 > \sigma_1^2$$

$$\frac{RSS}{\sigma_2^2} \sim \chi^2(n-p)$$

$$\Delta = \frac{\dfrac{RSS}{n-p}}{\dfrac{\tilde{\sigma}_1^2}{\sigma_1^2}} \sim F(n-p, l-1)$$

Пример. Для проверки берём (14 10 5) и (15 9 4) (дальше их не используем)

$$\tilde{\sigma}_1^2 = (5 - 4.5)^2 + (4 - 4.5)^2 = 0.5$$

$$n = 13 \qquad RSS = 17.88 \qquad \tilde{\Delta} = \frac{17.88/10}{0.5} = 3.58$$

$$\Delta \sim F(10, 1) \qquad \text{p-value} = \int_{3.58}^{\infty} q(t)dt = 0.39$$

14.5 Проверка предсказательной способности модели

Cross-validation (CV) - разбиваем на коробочки (обычно 5), одну используем для проверки, остальные - для обучения, каждый раз выбранную коробку и потом берём среднее

Пример. $n = 15$, по $n = 14 \to$ регрессия $CVSS_i = (y_i - \bar{y}_i)^2$

$$\vec{CVSS} = (0.15; 0.07; \ldots; 8.21) \qquad CVSS = \sum CVSS_i$$

$$R_{CV}^2 = \frac{TSS - CVSS}{TSS} = 0.172$$

R_{CV}^2 - насколько хорошо прогнозирует

14.6 Упрощение регрессии

Исключаем незничительные коэфф.

$$\{x_1\} \qquad \tilde{\beta}_1 = 0.29$$
$$y = 0.29x_1 + e$$
$$RSS_0 = 23.35 \qquad RSS_1 = 18.54$$
$$\Delta = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1)} \sim F(p_1 - p_0, n - p_1)$$
$$\tilde{\Delta} = \frac{(23.35 - 18.54)/(3 - 1)}{18.54/(12)} = 1.56$$
$$\text{p-value} = 0.25$$

14.7 Проблемы регрессии

14.7.1 Выбросы

Выбросы - как в boxplot, один из вариантов брать $\varepsilon \sim t(m)$

14.7.2 Мультиколлинеарность

$\xi_1, \ldots, \xi_k$ - могут коррелировать

$$\Psi = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} \qquad 2x_1 = x_2$$
$$F = \Psi^T \Psi = \begin{pmatrix} 3 & 6 \\ 6 & 12 \end{pmatrix} \qquad det\, F = 0$$

Пытаемся бороться:

$$\xi_1 = b_1 + b_2\xi_2 + \cdots + b_k\xi_k + \varepsilon_1 \;\rightarrow\; R^2$$

Если $R^2 > 0.7$ считаем что фактор определяется остальными и отбрасываем.

$$\xi_2 = b_1\xi_1 + b_2 + \cdots + b_k\xi_k + \varepsilon_2 \;\rightarrow\; R^2$$
$$\ldots$$

(Если ξ_1 отбросили в дальнейшие регрессии не включаем) Частный коэффициент корреляции

$$\eta = b_1 + b_2\xi_1 + \varepsilon \ \rightarrow \ y = \underbrace{\tilde{b}_1 + \tilde{b}_2 x_1}_{\tilde{y}} + e_1$$

$$\xi_2 = a_1 + a_2\xi_1 + \varepsilon_2 \ \rightarrow \ x = \underbrace{\tilde{a}_1 + \tilde{a}_2 x_1}_{\tilde{x}_2} + e_2$$

$$e_1 = y - \tilde{y} \qquad e_2 = x_2 - \tilde{x}_2$$

$$\rho_{\eta\xi_2} = \frac{\tilde{cov}(e_1, e_2)}{\sqrt{\tilde{D}e_1 \tilde{D}e_2}}$$

$$\tilde{D}e_1 = \frac{1}{n}\sum_{i=1}^{n}(e_{1i} - \bar{e}_1)^2 \qquad \tilde{D}e_2 = \frac{1}{n}\sum_{i=1}^{n}(e_{2i} - \bar{e}_2)^2$$

$$\tilde{cov}(e_1, e_2) = \frac{1}{n}\sum_{i=1}^{n}(e_{1i} - \bar{e}_1)(e_{2i} - \bar{e_2})$$

Пример. 20 лет, ν - урожай, ξ_1 - осадки, ξ_2 - темп.

$$r_{\eta\xi_1} = 0.8 \qquad r_{\eta\xi_2} = -0.4 \qquad r_{\xi_1\xi_2} = -0.75$$
$$\rho_{\eta\xi_2} = 0.5 \qquad \rho_{\eta\xi_1} = 0.82 \qquad \rho_{\xi_1\xi_2} = -0.78$$

14.7.3 Методы регуляризации

ridge-регрессия (гребневая) - борется с мультиколлинеарностью и переобучением (штрафная функция $\beta_1^2 + \beta_2^2 + \cdots + \beta_p^2$)

$$e^T e \to \min$$
$$(Y - \Psi\tilde{\beta})^T(Y - \Psi\tilde{\beta}) + \alpha\tilde{\beta}^T\tilde{\beta} \to \min \qquad (*)$$
$$Y^TY - \tilde{\beta}^T\Psi^TY - Y^T\Psi\tilde{\beta} + \underbrace{\tilde{\beta}^T\Psi^T\Psi\tilde{\beta} + \alpha\tilde{\beta}^T\tilde{\beta}}_{\tilde{\beta}^T(F+\alpha E)\tilde{\beta}}$$

Схоже с поиском условного максимума $L = f + \lambda g \to \min$

$$e^T e \to \min \qquad \tilde{\beta}^T\tilde{\beta} = t$$

α - через CVSS

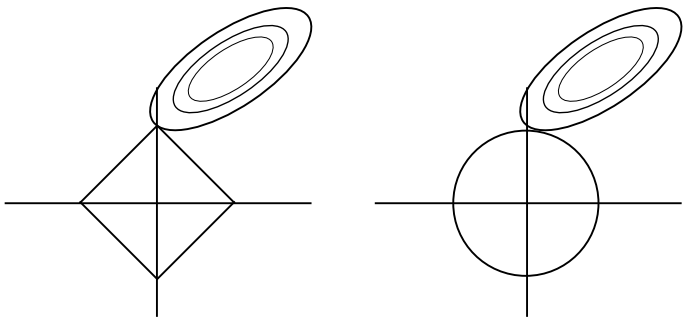
$$\alpha_0 \le \alpha \le \alpha_1$$

α_i - один элемент выборки выбрасываем, по остальным ищем $\tilde{\beta}$ по $(*)$

$\alpha_i \to \tilde{\beta} \to CVSS_1 = (y_1 - \tilde{y}_1)^2 \to CVSS = \sum CVSS_i$

Выбираем α при котором CVSS минимален

lasso-регрессия - убирает незначиммые факторы (штрафная функция $|\beta_1| + |\beta_2| + \cdots + |\beta_p|$)



lasso ridge

У lasso острый кочик поэтому вероятность того что $(*)$ его коснётся больше, чем у ridge

---

Elastic net

$$e^T e + \alpha_1 \tilde{\beta}^T \tilde{\beta} + \alpha_2(|\tilde{\beta}_1| + \cdots + |\tilde{\beta}_p|)$$

Bridge

$$e^T e + \alpha(|\tilde{\beta}_1|^q + \cdots + |\tilde{\beta}_p|^q)$$

14.7.4 Статист. анализ при наруш. усл. регрессии

Непараметрический бутстрап - метод доверительных интервалов

а) Проверка значимости β_i

$$H_0 : \beta_i = 0 \qquad H_1 : \beta_i \neq 0$$
$$h = \beta_i \qquad \tilde{h} = \tilde{\beta}_i \qquad \Delta = \tilde{h} - h$$

Подвыборка объёмом n $\rightarrow$ уравнение линейной регрессии $\tilde{\beta}_i^* \rightarrow \Delta_i^* = \tilde{\beta}_i^* - \tilde{\beta}_i$, повторяем 1000 раз, строим вариационный ряд $\Delta^*_{(1)} \dots \Delta^*_{(1000)}$

$$K_1 = [\frac{1-\beta}{2}N] \qquad K_2 = [\frac{1+\beta}{2}N]$$
$$P(\Delta^*_{K_1} < \underbrace{\tilde{h}^*}_{\tilde{h}} - \underbrace{\tilde{h}}_{h} < \Delta^*_{K_2}) \approx \beta$$
$$I = (\tilde{h} - \Delta^*_{K_2} < h < \tilde{h} - \Delta^*_{K_1})$$

Если $0 \ni I$ (I накрывает 0) $\Rightarrow$ коэф. не явл. значимым, иначе значим

---

b) Сравнение регрессий

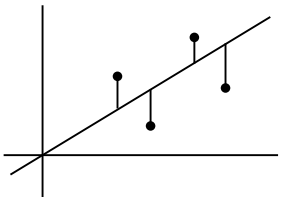
$$H_0 : R_1^2 = R_0^2 \qquad H_1 : R_1^2 > R_0^2$$
$$h = R_1^2 - R_0^2 \qquad \tilde{h} = \tilde{R}_1^2 - \tilde{R}_0^2 \qquad \Delta = \tilde{h} - h$$
$$H_0 : h = 0 \qquad H_1 : h > 0$$
$$\Delta^*_{(1)} \dots \Delta^*_{(1000)} \qquad K = [(1-\beta)N]$$
$$P(\Delta^*_{(K)} < \underbrace{\tilde{h}^* - \tilde{h}}_{\tilde{h}-h} < \Delta^*_{1000}) \approx \beta$$
$$I = (\tilde{h} - \Delta^*_{1000}, \tilde{h} - \Delta^*_{(K)})$$

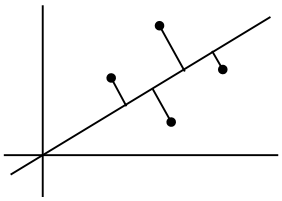
Если $0 \ni I$ различие незначимо, иначе регрессии значимо различаются

14.7.5 Разные типы регрессии

$\eta = f(x) + \varepsilon(x)$ работает когда $\sigma_\xi << \sigma_\eta$

$\eta = f(\xi) + \varepsilon(\xi)$ когда $\sigma_\xi \sim \sigma_\eta$





Бонус медианная регрессия Тейла-Сена

Пример (Дисперсионный анализ)**.** Три группы - разные методы, $n_1 = 7$, $n_2 = 5$, $n_3 = 3$, $\alpha = 0.05$

1 гр	1	3	2	1	0	2	1
2 гр	2	3	2	1	4		
3 гр	4	5	3				

Индикаторные переменные $x_1x_2x_3$, $x_i = 0, 1$

$$\eta = \beta_1 x_2 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

$$\Psi = \begin{pmatrix} 1 & 0 & 0 \\ & \cdots \times 5 & \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ & \cdots \times 3 & \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \qquad Y = \begin{pmatrix} 1 \\ \dots \\ 1 \\ 2 \\ \dots \\ 4 \\ 4 \\ 5 \\ 3 \end{pmatrix}$$

$$F = \Psi^T\Psi = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{pmatrix} \qquad \begin{pmatrix} \frac{1}{7} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \qquad \tilde{\beta} = F^{-1}\Psi^T Y = \begin{pmatrix} 1.43 \\ 2.4 \\ 4 \end{pmatrix}$$

Все коэфф значимы (по p-value)

$$RSS = 12.91 \qquad TSS = 26.39$$

$$R^2 = \frac{TSS - RSS}{TSS} = 0.52$$

$$H_0 : \beta_3 = \beta_2 \qquad H_1 : \beta_3 \neq \beta_2$$

$$p = 3 \qquad n = 15$$

$$\Delta = \frac{\beta_2 - \beta_3}{\sqrt{RSS(F_{22}^{-1} + F_{33}^{-1})}}\sqrt{n-p} \sim t(n-p)$$

$$\tilde{\Delta} = -2.112 \qquad \text{p-value} = P(|\Delta| \geq |\tilde{\Delta}| \Big| H_0) = 2\int_{2.112}^{\infty} q(t)dt = 0.057$$

Методы 2 и 3 одинаковы

Без учёта множественности проверки p-value$_{23} = 0.057$,p-value$_{12} = 0.1357$,p-value$_{13} = 0.0037$,

Холм-Бонферрони: $m = 3$

-	+	+
0.0037	0.057	0.135
α/m	$\alpha/(m-1)$	α
0.0167	0.025	0.05