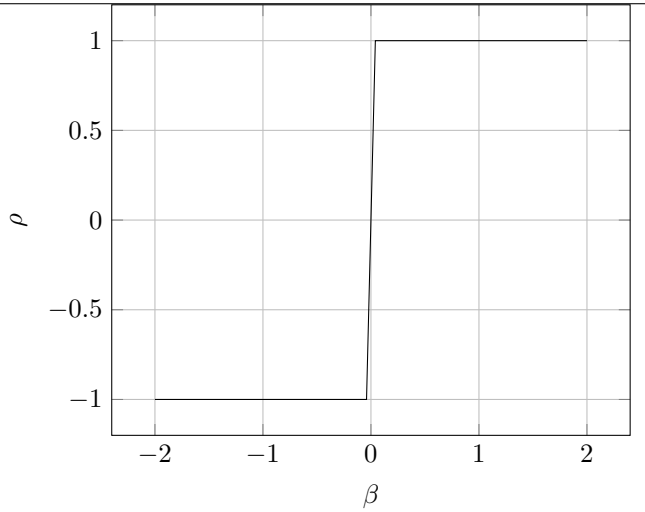
Меры связи между перменными

$$\rho = \frac{\operatorname{cov}(y, x \mid \hat{s})}{\sigma_x \sigma_y}$$

$$y = \alpha + \beta x + \varepsilon$$

$$\beta = \frac{\operatorname{cov}(y, x)}{\sigma_x^2}$$



$$y = \alpha + \beta x$$

корреляция Спирмена

$$p(n_i^a, n_i^b) = 1 - \frac{\sum_{i=1}^{m} d_i^2}{m(m-1)}, \quad d_i = n_i^a - b_i^b$$

$$S = (y_1, x_1), \ldots, (y_m, x_m)$$
$$S = S_1, \ldots, S_m$$

$$S_{j'}, S_{j''} \text{ – согласованы } \Leftrightarrow$$
$$x_{j'} > x_{j''} \cap y_{j'} > y_{j''}$$
$$x_{j'} < x_{j''} \cap y_{j'} < y_{j''}$$

$$\gamma = \frac{N_{\text{с}} - N_{\text{нс}}}{N_{\text{с}} + N_{\text{нс}}}; \quad f = \frac{N_{\text{с}} - N_{\text{нс}}}{\left(\frac{m(m-1)}{2}\right)}$$

не учитываются равенства, Кэмпбелл

Критерий сравнения бинарных переменных

$$\begin{array}{ccc} & A & B \\ F & a & b \\ M & c & d \end{array}$$

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(b+d)(a+c)}} = \frac{\chi^2}{m}$$

$$\chi^2 = \sum \frac{(mp_i - m_i)^2}{p_i} = [(\nu_1 - \nu_0)^2 m_1 + (\nu_2 - \nu_0)^2 m_2] \frac{1}{\nu_0(1-\nu_0)}$$

$$\nu_1 = \frac{\#\text{AF}}{\#\text{F}}; \quad \nu_0 = \frac{\#\text{A}}{\#\text{A} + \#\text{B}}; \quad m_1 = \#\text{F}; \quad \nu_2 = \frac{\#\text{AM}}{\#\text{M}}; \quad m_2 = \#\text{M}$$

Гудман, Крускал

$$y_1, \dots, \dots, \dots, y_m$$

разнот-авыолаывфлода вфы $y - y \mid x$ — мера Крускала
$x - x \mid y$

$$y \mid x_1, \dots, x_n$$

оптимальна

$$\mathbb{E}(y - R)^2 = \min_R$$

$$\Leftrightarrow$$

$$\mathbb{E}(Y - R)^2 \le \mathbb{E}(Y - \alpha - \beta R)^2 =$$

$$\left[\begin{cases} \frac{\partial \mathbb{E}(\alpha,\beta)}{\partial \alpha} = 0 \Big|_{\alpha=0,\beta=1} \\ \frac{\partial \mathbb{E}(\alpha,\beta)}{\partial \beta} = 0 \Big|_{\alpha=0,\beta=1} \end{cases} \right] =$$

$$\mathbb{E}(Y^2 + \alpha^2 + \beta^2 R^2 - 2\alpha Y - 2\beta RY + 2\alpha\beta R) \ge$$
$$\mathbb{E}(Y^2 + R^2 - 2YR)$$

$$\mathbb{E}(\alpha^2 + (\beta^2 - 1)R^2 - 2\alpha Y - 2(\beta - 1)RY + 2\alpha\beta R) \geq 0$$
$$\mathbb{E}(-2Y(\alpha + \beta - 1) + \alpha^2 + (\beta^2 - 1)R^2 + 2\alpha\beta R) \geq 0$$
$$\mathbb{E} - 2Y + 2\alpha + 2\beta R = 2\alpha + 2\mathbb{E}(\beta R - Y) = 0$$
$$\Rightarrow \mathbb{E}Y = \beta\mathbb{E}R$$
$$\mathbb{E} - 2Y + 2\beta R^2 + 2\alpha R \Rightarrow \begin{cases} \mathbb{E} - Y + R^2 + R \\ \mathbb{E}R^2 = \mathbb{E}YR \end{cases}$$
$$\rho = \sqrt{\frac{D(R)}{D(Y)}}$$

$$\mathbb{E}(Y - R)^2 = \mathbb{E}Y^2 - 2\mathbb{E}(YR) + \mathbb{E}R^2 =$$
$$\mathbb{E}Y^2 - \mathbb{E}R^2$$
$$\frac{\mathbb{E}(Y - R)^2}{\mathbb{D}Y} = 1 - \frac{\mathbb{E}R^2}{\mathbb{E}Y^2} = 1 - \rho^2$$
$$\rho^2 = 1 - \frac{\mathbb{E}(Y - R)^2}{\mathbb{D}Y}$$

Точечные оценки
несмещенность:

$$\mathbb{E}(\hat{\theta}(S)) = \theta$$

состоятельность:

$$\hat{\theta} \to_{m\to\infty} \theta$$
$$\forall\varepsilon > 0 \lim_{m\to\infty} P\{|\hat{\theta}_m(\hat{S}_m) - \theta| > \varepsilon\} = 0$$

эффективность (оптимальность):

$$\mathbb{E}(\hat{\theta} - \theta)^2 \leq \mathbb{E}(\hat{\theta}' - \theta)^2 \forall\hat{\theta}'$$
$$\mathbb{E}(\hat{\theta} - \theta)^2 \geq \frac{1}{mI(\theta)}, I(\theta) = \mathbb{E}\left(\frac{\partial l}{\partial\theta}\right)^2$$

$$L(s|\overline{x}, \overline{\theta}) = \prod_{\overline{x}_j \in \overline{s}} f(x_j \mid \overline{\theta})$$
$$s_m(\hat{\theta}_m - \theta) \to \mathcal{N}(0, I(\theta))$$
$$\int L(\theta, \hat{\theta}) f(\theta) d\theta \to \min_{\hat{\theta}} \Rightarrow$$
$$\hat{\theta} = \int \theta f(\theta \mid \overline{s}) d\theta$$

Метод моментов

$$
\begin{aligned}
U &\quad f(U, \theta_1, \ldots, \theta_k) \\
\hat{M}_1: &\quad \mathbb{E}(U) = g_1(\theta_1, \ldots, \theta_k) \\
\hat{M}_2: &\quad \mathbb{E}(U - \mathbb{E}U)^2 = g_2(\theta_1, \ldots, \theta_k) \\
\hat{M}_3: &\quad \mathbb{E}(U - \mathbb{E}U)^3 = g_3(\theta_1, \ldots, \theta_k)
\end{aligned}
$$

$$
\begin{aligned}
\hat{M}_1 &= g_1(\hat{\theta}_1, \ldots, \hat{\theta}_k) \\
\hat{M}_2 &= g_2(\hat{\theta}_1, \ldots, \hat{\theta}_k) \\
\hat{M}_3 &= g_3(\hat{\theta}_1, \ldots, \hat{\theta}_k)
\end{aligned}
$$

Распределение Парето, возникло из решения уравнения Колмогорова, в котором зашита следующая идея:

//прирост успеха к накопленному успеху//

кол-во богатства, кол-во ученых по числу публикаций, количество белков по доменам в организме

$$
\begin{gathered}
P(x \mid k, x_m) = \frac{k x_m^u}{x^{u+1}} [x_{\geq} x_m] \\
\mathbb{E}U = \frac{k}{k-1} x_m, \quad \mathbb{E}U^2 = \frac{k}{k-2} x_m^2 \\
\mathbb{D}U = \frac{k x_m^2}{(k-2)(k-1)^2}
\end{gathered}
$$

Пример:

$$
\begin{gathered}
35 \cdot 10^3 = \frac{k}{k-1} x_m \\
4 \cdot 10^8 = \frac{k x_m^2}{(k-2)(k-1)^2} \\
k(k-2) = \frac{400}{35} \\
k^2 - 2k - \frac{400}{35} = 0 \\
k \approx 3 \\
x_m = \frac{2}{3} \cdot 35 \cdot 10^3 \approx 23 \cdot 10^3
\end{gathered}
$$

МНК:

$$\prod_{j=1} f(x_j \mid \theta_1, \ldots, \theta_k) : y_j = x + \sum_{i=1}^{n} \beta_j x_{ji} + \varepsilon_j$$

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

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - \alpha - \sum_{i=1}^{n} \beta_i x_{ji})^2/2\sigma^2}$$

$$\hat{\beta} = (\hat{x}^T \hat{x})^{-1} \hat{x}^T; \quad y = \hat{C}_1 y_1 + \cdots + \hat{C}_m y_m$$

$$\mathbb{E}\hat{\beta}_i = \beta_i$$

$$\forall \overline{\gamma} \in \mathbb{R}^{n+1} : \mathbb{E}(\overline{\gamma} + \hat{\beta} - \overline{\gamma}^T \beta)^2 \leq \mathbb{E}(\gamma^T \hat{\beta}' - \gamma^T \beta)^2$$

$$\Leftrightarrow \sum_{\hat{\beta}} - \sum_{\hat{\beta}'} \geq 0$$

теорема Гаусса-Маркова

1. гомоскедактичность
2. нез. ξ
3. $\exists$ однозн. р.
4. $\exists$ реш.

$\Rightarrow$ МНК $\to$ BLUE

Дов. интервал

$$\theta \in [\theta_l(\hat{s}), \theta_u(\hat{s})]$$

Распределение Стьюдента

$$\sim T$$

$$\frac{U}{\sqrt{\frac{1}{n}V}}, \quad U \sim N, \quad V = \sum_i U_i^2 \sim \chi^2$$

$$z = \frac{\overline{x} - \hat{\mu}}{\sqrt{\frac{D}{m}}} \Rightarrow z \sim T$$

$$D = \sum_i (x_i - \overline{x})^2 \sim \chi^2$$

$$t_{m-1;\alpha/2} \leq z \leq t_{m-1;1-\alpha/2}$$

$$t_{m-1;\alpha/2} \leq \frac{\hat{x} - \mu}{\sqrt{D/m}} \leq t_{m-1;1-\alpha/2}$$

$$x - \sqrt{\frac{D}{m}} t_{m-1;1-\alpha/2} \leq \mu \leq x - \sqrt{\frac{D}{m}} t_{m-1;\alpha/2}$$

$$x - c \leq \mu \leq x + c$$

Доверительный интервал Клоппера-Пирсона
α — уровень значимости

$$P_u :$$

$$\sum_{i=0}^{k} C_n^i p_u^i (1 - p_u)^{n-i} = \frac{\alpha}{2}$$

$$P_l :$$

$$\sum_{i=k}^{n} C_n^i p_l^i (1 - p_l)^{n-i} = \frac{\alpha}{2}$$

$$\text{Acc} \sim \mathcal{N}\left(p, \frac{p(1-p)}{m}\right)$$

$$Z_{N,\alpha/2} \leq \frac{\nu - p}{\sqrt{\frac{p(1-p)}{m}}} \leq Z_{N,1-\alpha/2}$$

Bootstrap
//Bagging//

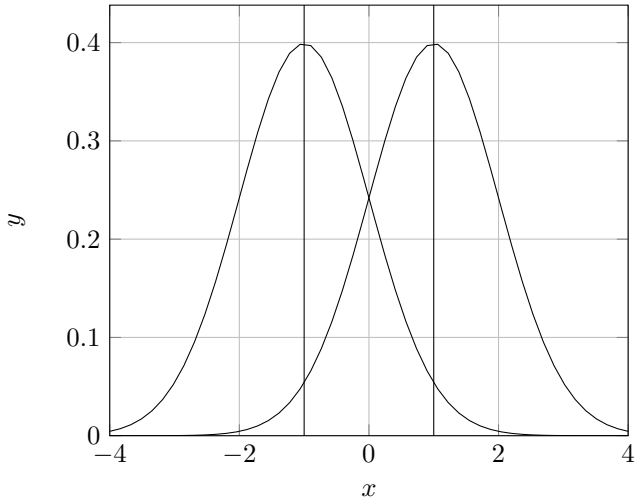
Статистические тесты

- Выборка; статистика критерия;
- (Нейманн-Пирсон) 2 гипотезы: h_0, h_1;

  Выбор на основе статистического критерия:

$$T(\tilde{s}) > \delta$$

$$T > \delta$$



α — ошибка 1-го рода (1-специфичность) $p(pred = h_1 \mid h_0)$ β — ошибка 2-го рода $p(pred = h_0 \mid h_1)$

$1 - \beta$ — мощность критерия

несмещенный критерий:

$(1 - \beta > \alpha)$ (wtf????)

состоятельный:

при фиксированном α, $\delta \to \infty$ (объем выбоки), мощность $\to 1$

лемма Неймана-Пирсона: при фиксированном α для максимальной мощности надо использовать $T = \frac{p(\hat{S}|h_1)}{p(\hat{S}|h_0)}$.

$$P[T(\tilde{w}) \geq T(\hat{S})]$$

„от противного”

//альтернативной гипотезы — нет//

двусторонние/односторонние критерии

Тест Стьюдента для корреляции Пирсона:

$$U_0, U_1, \ldots, U_n$$

$$V = \sum_{i=1}^{n} U_i \sim \chi^2 \text{ — с числом степеней свободы } n$$

$$\frac{u_0}{\sqrt{\frac{1}{n}V}} \sim \tau \text{ — с числом степеней свободы n}$$

$$X_1, X_2$$

$$\tilde{S} = \{(x_{11}, x_{12}), \ldots, (x_{m1}, x_{m2})\}$$

$$T(\omega) = \rho(\omega)\sqrt{\frac{m-2}{1-\rho^2(\omega)}}$$

Статистика $T(\omega)$ распределена по Стьюденту с $m-2$ степенями свободы.

одностороннее p-value: $p = 1 - \mathbb{F}^{\text{st}}_{m-2}[T(\tilde{S})]$

Гипотеза на нормальность

Если два распределения нормальны, то статистика $T(\omega) = 1 + \log\frac{1+\hat{\rho}(\omega)}{1-\hat{\rho}(\omega)}$ хорошо приближается нормальным распределением

(в нулевую гипотезу входит ρ_0 — коэффициент корреляции)

$$\frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0} \text{ — Z-преобразование Фишера}$$

$$\text{Дисперсия} — \sqrt{\frac{2}{m-3}}$$

$$z_0 = \frac{1}{2}\sqrt{\cdots}$$

$$\{x_1, \ldots, x_{m_1}\}$$
$$\{x_{m_1+1}, \ldots, x_{m_2}\}$$

$$\hat{\sigma}_1 = \sqrt{\frac{1}{m_1 - 1}\sum_{i=1}^{m_1}(x_i - \mu_1)^2}$$

$$\hat{\sigma}_2 = \sqrt{\frac{1}{m_2 - m_1 - 1}\sum_{i=m_1+1}^{m_2}(x_i - \mu_2)^2}$$

$$\hat{\mu}_1 = \frac{1}{m_1}\sum_{i=1}^{m_1} x_i$$

$$\hat{\mu}_2 = \frac{1}{m_2 - m_1}\sum_{i=m_1+1}^{m_2} x_i$$

$$\hat{\sigma}_{12} = \frac{(m_1 - 1)\hat{\sigma}_1 + (m_2 - m_1 - 1)\hat{\sigma}_2}{m_2 - 2}$$

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2 - \delta}{\hat{\sigma}_{12}\sqrt{\frac{1}{m_1} + \frac{1}{m_2 - m_1}}} \text{ — по } z$$

T — по Стьюденту с $\text{df} = m_2 - 2$

Критерий. p-value: $p = 1 - \mathbb{F}^{\text{st}}_{m-2}[T(\tilde{S}_1, \tilde{S}_2)]$

$$\nu = \frac{\left(\frac{\hat{\sigma}_1^2}{m_1} + \frac{\hat{\sigma}_2^2}{m_2 - m_1}\right)}{\frac{\hat{\sigma}_1^4}{m_1^2(m_1-1)} + \frac{\hat{\sigma}_2^4}{(m_2-m_1)^2(m_2-m_1-1)}}$$

Значимость регрессионных коэффициентов.

$$y_1 = \beta_0 + \sum_{i=1}^{n} \beta_i x_{1i} + \varepsilon_1$$
$$y_m = \beta_0 + \sum_{i=1}^{n} \beta_i x_{mi} + \varepsilon_m$$

$$\beta_i \qquad \beta_i'$$

гомоскедактичность

$$\hat{\sigma}^2(\beta_i) = \frac{\hat{\sigma}^2_{\text{err}}}{\sum_{j=1}^{m} (x_{ji} - \overline{x}_i)^2}$$
$$\hat{\sigma}^2_{\text{err}} = \frac{\sum (y_j - \hat{y}_j)^2}{m - n - 1}$$

Статистика критерия:

$$T_i = \frac{\hat{\beta}_i - \beta_i'}{\hat{\sigma}(\beta_i)} \sim \mathcal{T}, \quad \text{df} = m - n - 1$$

//β_i' часто устанавливают равным нулю//
R-тест
R-распределение:

$$V = \frac{U_1 d_2}{U_2 d_1}$$
$$U_1 \sim \chi^2, U_2 \sim \chi^2$$
$$\text{df} = d_1, \text{df} = d_2$$

$$\mathrm{SSM} = \sum_{j=1}^{m} (y_j - \overline{y})^2$$

$$\mathrm{SSR} = \sum_{j=1}^{m} (\hat{y}_j - y_j)^2$$

$$\mathrm{MSM} = \frac{\mathrm{MSS}}{n}$$

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{m-n-1}$$

Критерий односторонний $\frac{\mathrm{MSM}}{\mathrm{MSR}} \sim F(n, m-n-1)$ (распределение Фишера)

β_i $\beta_0 = \beta_1 = \cdots = \beta_n = 0$

Дисперсионный анализ

X, Y

принимает I значений:

$$\mu_1, \ldots, \mu_I$$
$$\mu_i = \mathbb{E}Y \mid x = i$$

субпопуляции

предположение: внутри субпопуляций данные нормально распределены, причем стандартные отклонения одинаковые

$$J_1, \ldots, J_I$$

J_i — (количество объектов в субпопуляции)

$$\mu = \frac{1}{m}\sum_{i=1}^{I} J_i\mu_i; \quad m = \sum_{i=1}^{I} J_i$$

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \mu - \alpha_i)^2$$

$$\sum J_i\alpha_i = 0$$

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{I}\sum_{j=1}^{J_i} y_{ij}$$

$$\alpha_i = \frac{1}{J_i}\sum_{i=1}^{J_i} y_{ij} - \hat{\mu}$$

$$\mathrm{SSB} = \sum_{i=1}^{I} J_i\alpha_i^2$$

$$\mathrm{SSR} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \alpha_i - \mu)^2$$

$$\frac{\mathrm{SSB}(m-1)}{\mathrm{SSR}(I-1)} \sim F(I-1, m-I)$$

χ^2, критерий χ^2

$$X$$

$$\tilde{S} = \{x_1, \ldots, x_m\}$$

$$T_{\chi^2} = \sum_{i=1}^{k}\frac{(mp_i - m_i)^2}{mp_i} \qquad \begin{cases} mp_i > 10 \\ m \to \infty \\ T_{\chi^2} \approx \chi^2(k-1) \end{cases}$$

G-тест

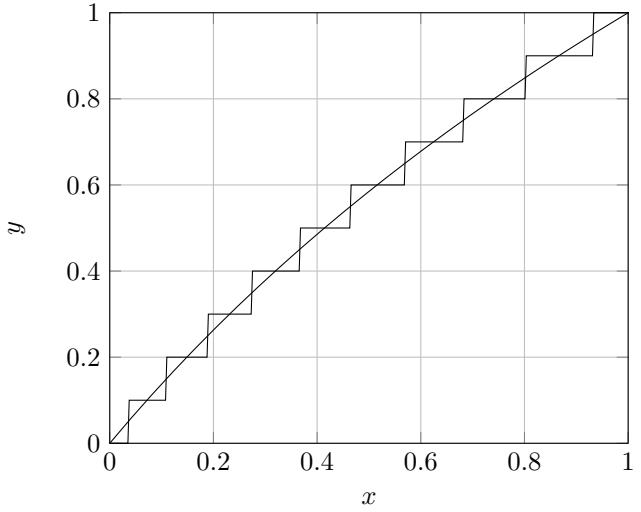
$$T_\sigma = 2\sum_{i=1}^{k} m_i \log\frac{m_i}{mp_i} \text{ (критерий максимального правдоподобия)}$$

короче, это круче предыдущего, тк это практически KL, а предыдущий это его приближение до квадратичного члена.

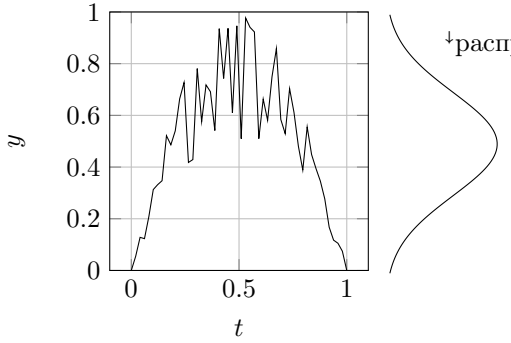
критерий Колмогорова-Смирнова
статистика Колмогорова:

$$D_n = \sup_x |F_n(x) - F(x)|$$

$$\Pr\{D_m > \varepsilon\} \to_{m\to\infty} 0$$



Броуновский мост



↓распределение броуновского моста

$$K = \sup_{t\in[0,T]} |B(t)|$$

$$\sum_k (t)^k e^{-2k^2x^2}$$

$mD_m \to K$ при $m \to \infty$

коррекция Фавыолрфа-Левенфорса, если делаем что-то интересное дважды по тем же данным

Про связь G-теста и критерия χ^2

$$\frac{mp_i}{m_i} = 1 + \delta_i; m_i = \frac{mp_i}{1+\delta_i}$$

$$\sum_{i=1}^{k} m_i \log \frac{m_i}{mp_i} = \sum_{i=1}^{k} \frac{mp_i}{1+\delta_i} \log(1+\delta_i) =$$

$$\sum_{i=1}^{k} \frac{mp_i}{1+\delta_i}(\delta_i - \frac{\delta_i^2}{2} + \dots) =$$

$$= \left[\sum m_i \frac{mp_i - m_i}{m_i} = \sum_{i=1}^{k} mp_i - m_i = 0 \right] =$$

$$= \sum_{i=1}^{k} -\frac{mp_i}{1+\delta_i} \left(\frac{\delta_i^2}{2} \right) == \sum_{i=1}^{k} -\frac{m_i}{1} \left(\frac{(mp_i - m_i)^2}{2m_i^2} \right) =$$

$$\sum_{i=1}^{k} -\frac{(mp_i - m_i)^2}{2m_i} = -\sum \frac{(mp_i - m_i)^2(1+\delta_i)}{2mp_i}$$

при $\delta_i = 0$ получается $-\frac{1}{2} \sum \frac{(mp_i - m_i)^2}{mp_i}$

K/C

$$S_1 = (x_1^{(1)}, \dots, x_m^{(1)})$$
$$S_2 = (x_2^{(2)}, \dots, x_m^{(2)})$$
одинаковы ли?

$$D = \sup_x |F_1(x) - F_2(x)|$$

$$\sqrt{\frac{m_1 m_2}{m_1 + m_2}} D \to_{m_1, m_2 \to \infty} K$$

$$P\left(\sqrt{\frac{m_1 m_2}{m_1 + m_2}} D \le x \right) = P(K \le X) \text{ при } m_1, m_2 \to \infty$$

Критерий Мана-Уитни

Предположений о распределении не требуется

Почему ранги? Устойчивы к выбросам, очень низкие p-values.

$$S_1 = (x_1^{(1)}, \dots, x_m^{(1)})$$
$$S_2 = (x_2^{(2)}, \dots, x_m^{(2)})$$

$$R_1^1 = \sum_{i=1}^{m} r(x_i^{(1)}); R_2^1 = \sum_{i=1}^{m} r(x_i^{(2)})$$

r – ранг в объединенной выборке

$$U_1 = R_1 - \frac{m_1(m_1+1)}{2}; U_2 = R_2 - \frac{m_2(m_2+1)}{2}$$
$$U_1 + U_2 = m_1 m_2$$
$$T = \min(U_1, U_2)$$
$$\mu = \frac{m_1 m_2}{2}; \sigma = \sqrt{\frac{m_1 m_2 (m_1 + m_2 + 1)}{12}}$$

ties: один и тот же/ как бы среднее

Точный тест Фишера

Задача анализа таблицы собряженности (X – pred, Y – true)

	Y_1	Y_2
X_1	a	b
X_2	c	d

$$P(a,b,c,d) = \frac{C_{a+c}^a C_{b+d}^d}{C_m^{a+d}}$$

экстремальное значение – max от mean

1	8
7	6

$$T = \sum_{k=1}^{2} \sum_{k=1}^{2} \frac{m_{kk'} - m p_{kk'}}{m p_{kk'}}$$
$$p_{kk'} = p(x = x_k) p(y = y_k)$$
$$\Phi^2 m = \frac{(ad - bc)^2 m}{(a+b)(c+d)(b+d)(a+c)}$$

20	5
5	10

Тест А.-В.

+++--++---++--+++-+-

Runs-тест

Серия

n_+ – число + в последовательности

n_- – число - в последовательности

k – число серий

$$T = \frac{R - \hat{R}}{\sigma_R}; \hat{R} = \frac{2n_+n}{n_+ + n} + 1$$

$$\sigma_R^2 = \frac{(2n_+n)(2n_+n - n_+ - n_-)}{(n_+ + n)(N_+ + n_- - 1)}$$

T распределена нормально ($n_+ > 10$, $n_- > 10$)

Уилкоксон Wilcoxson

$$\left\{\left(x_1^{(1)}, x_1^{(2)}\right), \ldots, \left(x_m^{(1)}, x_m^{(2)}\right)\right\}$$

$$P_j = |x_j^{(1)} - x_j^{(2)}|$$

$$T = \sum_{j=1}^{m} R_j \text{sgn}(x_j^{(2)} - x_j^{(1)})$$

$\frac{T}{\sigma}$ стремится к нормальному распределению, где $\sigma = \sqrt{\frac{m(m+1)(2m+1)}{6}}$ – ранговый критерий Пирсона.

Где-то тут должны были быть перестановочные тесты и, видимо, введение во множественное тестирование, которые я пропустил.

Точное p-значение

FWER $= P(V \geq 1)$ — приближение Бонферрони

у p-значения вероятность быть < 0.01 равна 0.01

что делать, если много гипотез?

неравенство Буля

$\Rightarrow$ p-значения надо умножать на n ; (соотвественно, уровень значимости делить на n)

$$\text{PFER} = \mathbb{E}(V) = \mathbb{E}[\sum_{i=1}^{n} I(H_0^i)] =$$

$$\sum_{i=1}^{n} \mathbb{E} I(H_0^i) = \sum_{i=1}^{n} \alpha = n\alpha \text{— поправка Бонферрони}$$

$$\frac{1}{N}\text{PFER} = \alpha = \text{PCER} \leq \text{FWER} \leq \alpha n = \text{PFER} \text{ — неравенство Буля}$$

FPR — доля ошибок первого рода среди отвергнутых нулевых гипотез. короче, он точнее описывает ситуацию

например, у нас есть пациенты и мед. препарат. если на 30 из 100 пациентов он подействовал, то PFER бесполезен.

$$100 \cdot 0.01 = 1$$

$$\text{FDR} = \frac{1}{30} \text{ — очень маленький}$$

Перестановочный тест

Пусть y не зависит от $x \Rightarrow y$ можно перестанавливать.

сколько отвержено?

сравниваем с реальным

одношаговая процедура

короче, поправка Бонферрони одношаговая. но можно лучше.

процедуры пошагового спуска

Бонферрони-Холма:

$$H_0^{(1)}, \dots, H_0^{(n)}$$

$$p^{(1)}, \dots, p^{(n)}$$

$$\alpha$$

$$\text{найдем } h : p^h > \frac{\alpha}{n+1-h}$$

Все гипотезы $H_0^{(1)}, \dots, H_0^{(h-1)}$ отвергаются.

Доказательство:

Пусть h — первая верная, но отвергнута ошибочно.

Пусть m_0 — общее количество верных нулевых гипотез.

Тогда

$$h - 1 \le m - m_0$$

$$\frac{1}{m-h+1} \le \frac{1}{m_0}$$

$$p(h) \le \frac{\alpha}{m-h+1} \le [\text{по построению самой процедуры}]$$

$$\le \frac{\alpha}{m_0} \text{ по Бонферрони}$$

■

короче, вот от самого грубого до самого точного метода из рассмотренных:

- Бонферрони
- Бонферрони-Холма
- Перестановочное тестирование

Временные ряды

$$x_1, \dots, x_T$$

Каждый временной ряд — реализация случайного процесса.
Каждый процесс:

$$X_t, t = \{\dots, -2, -1, 0, 1, 2, \dots\}$$

Пусть есть фиксированный интервал $t_1, \dots, t_n$
Процесс характеризуется совместным распределением:

$$f(x_{t_1}, \dots, x_{t_n})$$

Стационарный процесс (строго стационарный):

$$f(x_{t_1+A}, \dots, x_{t_n+A}) = f(x_{t_1}, \dots, x_{t_n})$$
$$\mathbb{E}(x_t) \text{ от } t \text{ независимо}$$
$$\mathbb{D}(x_t) \text{ от } t \text{ независимо}$$

$$\operatorname{cov}(x_{t_1}, x_{t_2}) = \int \cdots \int (x_{t_1} - \mu)(x_{t_2} - \mu) f(x_{t_1}, x_{t_2}, \dots, x_{t_n}) dx_{t_1} \dots dx_{t_n} =$$
$$\operatorname{cov}(x_{t_1+\Delta}, x_{t_2+\Delta}) = \int \cdots \int (x_{t_1+\Delta} - \mu)(x_{t_2+\Delta} - \mu) f(x_{t_1+\Delta}, x_{t_2+\Delta}, \dots, x_{t_n}) dx_{t_1} \dots$$

Стационарность слабая (стационарность в широком смысле):

- независимость $\mathbb{E}$, $\mathbb{D}$ от t;
- автоковариационная функция зависит только расстояния между точками

Белый шум: в $\forall t \;\; \mathbb{E} = 0, \;\; \mathbb{D} = \sigma^2$
Между x_{t_1} и x_{t_2} зависимости нет (если $t_1 \neq t_2$)
Гауссовый шум — белый, причем нормальный в каждой точке.
Красный шум — процесс случайного блуждания

$$x_t = x_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim \text{белый шум}$$
$$\mathbb{E}x_t = 0 \quad \mathbb{D}x_t = t\mathbb{D}\varepsilon$$

ряд с трендом, ряд с сезонами
теорема Вольда

Любой стационарный процесс представим в следующем виде:

$$x_t = \mu + \sum_{i=0}^{\infty} b_j \varepsilon_{t-j}, \quad \sum_{i=0}^{\infty} |b_i| < \infty$$

процесс скользящего среднего:

$$x_t = \mu + \sum_{j=1}^{q} b_j \varepsilon_{t-j}$$

$$\mathbb{E}x_t = \mu, \quad \mathbb{D}x_t = \sum_{j=1}^{q} b_j^2 \sigma^2 \text{ — не зависит от } t$$

автоковариация зависит только от расстояния:

$$\operatorname{cov}(x_{t+\tau}, x_t) = \sigma^2 \sum_{j=1}^{q-\tau} b_j b_{j-\tau}$$

линейная авторегрессия

$$x_t = \sum_{j=1}^{p} a_j x_{t-j} + \varepsilon_t$$

прогнозирование временных рядов
сезонные колебания — можно учитывать
сезонность, тренд, стохастическая случайность
ARMA — для стационарных рядом
ARIMA — не обязательно только для стационарных рядов

$$y_t = x_t - x_{t-1}$$

Тест на стационарность Дики-Фуллера

$$x_{t+1} - x_t = \gamma x_{t-1} + \varepsilon_t$$
$$x_t = \rho x_{t-1} + \varepsilon_t$$

$$x_t = \sum_{j=1}^{p} a_j x_{t-j} + \sum_{i=1}^{q} b_j \varepsilon_j$$

ACF: $Y(t) = P(y_t, y_{t-i})$ — коэффициент корреляции

PACF: $\gamma_p(t) = P(y_t - \sum_{j=1}^{i-1} \Phi_j y_{t-j}, y_{t-i})$ — показывает связь только между y_t и y_{t-i} без влияния промежуточных (где $\Phi = \arg\min_\Phi \sum_j (y_t - \Phi_j y_{t-j})^2$)

Переходим к виртуальным переменным, по которым можно найти a

$$z_1 = x_t$$
$$x_2 = x_{t-1} - b_1 z_1$$
$$z_k = x_k - \sum_{j=1}^{q-1} b_j z_j$$

$$z_t = \sum_i a_i z_{t-i} + \varepsilon_t$$

$$\overline{a} = \arg\min_a \sum_{t=1}^{T} \left(z_t - \sum_i a_t z_{t-i} \right)^2$$

Box-Jenkins
Только если стационарен!

$$Lx_t := x_{t-1}, \quad L^2 x_t = x_{t-2}$$

MA: $x_t = (1 + b_1 L + \cdots + b_q L^q)\varepsilon_t$
AR: $x_t = (a_1 L + \cdots + a_p L^p)x_t + \varepsilon_t$
ARMA: $(1 - a_1 L - \cdots - a_p L^p)x_t = (1 + b_1 L + \cdots + b_q L^q)\varepsilon_t$

$$x_t = \frac{1 + b_1 L + \cdots + b_q L^q}{1 - a_1 L - \cdots - a_p L^p}\varepsilon_t$$

экзамен 22 числа в 10.

временные ряды.

стационарность — совместная плотность не меняется со сдвигом по времени. проверить сложно, используют стационарность ковариции. (функция автокорелляции)

белый шум — мат. ожидание нулевое, любые две точки независимые. не путать с гауссовским (нормальное распределение в каждой точке)

процесс случайного блуждание (красный шум) — приращение ряда является белым шумом. мат. ожидание ноль, дисперсия растет линейно пропорционально дисперсии приращения

процесс с трендом, процесс с сезонностью — не стационарные

теорема Вольда — представление через сумму шумов

процесс скользящего среднего — мат. ожидание и сумма шумов (обрезанный ряд т. Вольда), стационарный:

$$X_t - \mu = \sum_{i=1}^{q} b_i \varepsilon_{t-i} + \varepsilon_t$$

процесс авторегрессии — выражается через значения в предыдущие моменты времени:

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t$$

ARMA (auto-regressive moving average)

Не может быть превращен в стационарный вычитанием линейного тренда:

$$X_t = a + X_{t-1} + \varepsilon_t$$

дрифт. потому что каждый раз добавляется a.

Как из случайного блуждания получить стационарный ряд? Разность соседних точек:

$$Y_t = X_t - X_{t-1}$$

Интегрированный порядка 1 — если можно получить стац. вычитанием соседних.

К Y_t тоже можно применять эту операцию. Пусть после одного раза не получилось, давайте еще раз.

Интегрированный порядка k (если после k применений появляется стационарность):

$$X_t \sim I(k)$$

Модель ARMA работает только для стационарного ряда. Поэтому для того, чтобы использовать процедуру ARMA, надо превратить исходный ряд в стационарный. Как? Взятием разности.

Модель ARMA, примененная к полученному ряду, называется ARIMA.

Есть следующий вопрос по ARMA.

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \sum_{i=1}^{q} b_t \varepsilon_{t-i} + \varepsilon_t$$

Для того чтобы задать его, нужно знать a и b. Можно применить МНК. Мы знаем X_t. Но мы должны получить коэффициенты b

Была предложенна схема перехода к дополнительным переменным Z, которые выражаются через X:

$$Z_1 = X_1; Z_2 = X_2 - b_1 Z_1; \ldots; Z_p = X_p - \sum_{i=1}^{p} b_i Z_{p-i}$$

$$Z_t = \sum_{i=1} a_i Z_{t-1} + \varepsilon_t$$

Здесь МНК уже известен. Справедливость доказана через операторы сдвига.

$$Q = \sum_{t=1}^{T} \left(Z_t - \sum_{i=1}^{p} a_i Z_{t-i} \right)^2$$

Как получить b? Перебираем на какой-то сетке, на котором значение Q минимально.

Мы не знаем еще p и q. Смотрим на коэффициент корреляции. Он должен убывать по времени. Смотрим, когда эта автокорреляция занулится.

Кроме ак используется частная автокорелляция. То есть мы должны при вычислении корреляции между переменными вычесть влияние промежуточных.

$$X_{t-\mu} \qquad X_t$$

$$Z_t = X_t - \sum_{i=1}^{M-1} \Phi X_{t-i}$$

$$\rho\left[Z_t, X_{t-\mu}\right] \text{ — частная автокорреляция}$$

(не уверен, но для a, b — обычная ак, для p, q — частная ак)

Про операторы сдвига.

$$LX_t = X_{t-1}$$

Авторегрессия:

$$X_t = \sum_{i=1}^{p} a_i L^i X_t + \varepsilon_t$$

$$\left(X - \sum_{i=1}^{p} L^i X_i \right) = \varepsilon_t$$

$$\left(1 - \sum_{i=1}^{t} L^i \right) X_i = \varepsilon_t$$

Процесс взятия K-разности: $(1-L)^K X_t$

$$A(L)(1-\rho L)^K X_t = \varepsilon_t$$

Возможность такого представления означает, что, воздействуя так, мы получаем стационарный процесс

Процесс с единичным корнем — если с помощью разностей можно получить стационарный ($\rho = 1$; н-р, процесс случайного блуждания). Процесс плохой, плохо предсказуемый и т.д.

Наличие корня ρ:

$$X_t = \rho X_{t-1} + \varepsilon_t$$
$$(1 - \rho L)X_t = \varepsilon_t$$

Есть в эконометрике концепция. Пусть есть несколько переменных. Очень часто возникают корреляции между этими переменными, который 1) статистически достоверны, 2) никак не объяснимы логически. Ложные регрессии. Встречаются значительно чаще, чем уровень значимости. Были попытки объяснить это какими-то скрытими переменными, влияниями. Но если мы будем моделировать случайные процессы, мы можем так сделать. Получается, что коэффициент корреляции 0.8 для ряда длины 15 раз в 20 случаев. Это характерно для стохастически нестационарных временных рядов. Замечено уже Ньювеллом. В 70-е годы утвердилось. Получается, стандартная статистика совершенно неприменима. Ее можно применять только когда ряды стационарны.

Как бороться с процессом случайного блуждания? Нужен инструмент.

Концепция коинтеграции. Опубликовано в 1981. Потребовали, чтобы остатки были стационарными.

Для того чтобы сказать, что процессы связаны, нужно показать, что остатки были не только маленькие, но и стационарными:

$$X_t^1 = \alpha X_t^2 + z_i$$

Напрямую тест Дики-Фуллера использовать нельзя, хоты бы потому, что мы уже подбираем коэффициенты через МК. Надо его корректировать, а то он слишком оптимистично. Это сделано через моделирование Монте-Карло.

Что такое коинтеграция на наборе переменных?

$$\exists a_i : \sum_{t=1}^{K} a_t X^t \text{ — стационарен}$$

Постоянно подвергается критике.

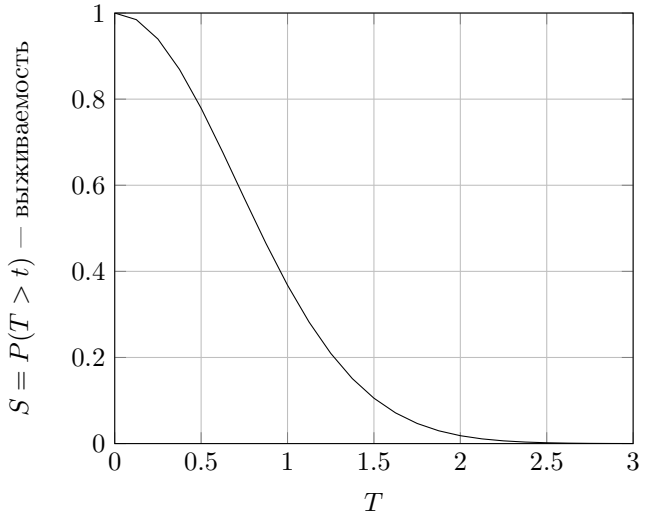
Когда нужен? Когда изучаем процессы с единичными корнями или процессы, которые сводятся к стационарным взятием разности.

Анализ выживаемости.

Используется в онкологии, в анализе занятости.

Нужно предсказать, что произойдет с объектом.

Точный момент времени — не предскажем. Наиболее точную картину дает кривая выживаемости.



Метод Каплана-Майера. Разбивается интервал времени на маленькие интервальчики.

$$\prod_{j=1}^{J} \frac{m_j - a_j}{m_j}$$

Цензурирование выборки данных (известно, что через 5 месяцев жив; известны только нижние границы)

$$\tilde{S} = \{(\alpha_j, t_j, x) | j = 1, \ldots, n\}$$

Цель — восстановить $S(t, \overline{X})$

Модель Кокса. Основывается на модели

$$\lambda(t) = \lim_{\Delta \to 0} \frac{P(T \leq t + \Delta | T > t)}{\Delta}$$

Из определения условной вероятности:

$$\lambda(t) = -\frac{dS}{S}$$

Что отсюда следует?

$$\log S = \int \lambda(t) dt$$

$$S = e^{-\int \lambda(t) dt}$$

$$\lambda(t, X) = \lambda_0(t) \lambda(X)$$

Интересная модель:

$$\lambda(t, X) = \lambda_0(t)e^{-\beta X}$$

Отсюда:

$$S(t, X) = (S_0(t))^{e^{-\beta x}}$$
$$S_0(t) = e^{-\int \lambda(t)Adt}$$

Максим

Чем занимается наука? Мы проверяем, если есть зависимость между целевой и .. переменной.

$$Y <-> X$$
$$H_0 : X \perp Y$$
$T(X, Y)$ — чем больше, тем больше зависимость

Генерируем новую выборку:

$$(X, Y) \to (X, \tilde{Y}), \tilde{Y} = o(Y)$$

Перестановочный тест:

$$\{T(X, \tilde{Y})|\forall \tilde{Y}\}$$

Что нужно сделать, чтобы найти вероятность того, что X и Y независимы?

$$T = T(X, Y)\{\tilde{T} = T(X, \tilde{Y})\}$$
$$\frac{\sum_{\tilde{T}} [T \le \tilde{T}]}{|\{\tilde{T}\}|} = p^*$$

Что делать, если у нас много X?

$$Y <-> X_1, \ldots, X_d$$
$$T_1(X_1, Y), \ldots, T_d(X_d, Y)$$
комплексная нулевая гипотеза
$$H_0 = \cap_{i=1}^{d} H_{i0}, H_{i0} = X_i \perp Y$$

$$p^* \approx p = P(T \le \tilde{T} \mid H_0)$$

Подставляем в combining function ψ

$$p_1^*, \dots, p_d^* \mapsto \psi(p_1^*, \dots, p_d^*)$$
$$\psi : [0,1]^d \to \mathbb{R}$$
$$1)\psi(\dots, \hat{p}_i, \dots) \geq \psi(\dots, \overline{p}_i, \dots), \hat{p}_i \leq \overline{p}_i$$
$$2)\exists\overline{\psi} \leq +\infty : \psi(\dots, p_i, \dots) \to \overline{\psi}$$
$$3)\forall\alpha > 0 \exists\psi_\alpha < \overline{\phi} : p(\phi > \phi_\alpha \mid H_0) = \alpha$$

Мы встроили перестановочный тест в перестановочный тест, чтобы можно было делать перестановочный тест пока мы делаем перестановочный тест:

$$p^* = \frac{\sum_{\tilde{\psi}} [\psi \leq \tilde{\psi}]}{|\{\tilde{\psi}\}|}$$

Значимый: $p(T \geq T_\alpha | H_1) \geq \alpha$, если $P(T \geq T_\alpha | H_0) = \alpha$
Несмещенный: $\forall z \in \mathbb{R} : P(T \leq z \mid H_0) \geq P(T \leq z \mid H_1)$
Состоятельный: $P(T \geq T_\alpha \mid H_1) \to_{n\to\infty} 1, \forall\alpha > 0$

Если три свойства выполнены, то это NPC (non-parametric combination)

Первый вариант: функция Фишера

$$\psi(\vec{p}) = -2\sum_i \log p_i$$

Второй вариант: функция Липтака

$$\psi(\vec{p}) = \sum_i F^{-1}(1 - p_i)$$

если совпадает с сигмоидой, то

$$F(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$$
$$F^{-1}(x) = -\log\left(x^{-1} - 1\right) = \log\left(\frac{x}{1-x}\right)$$
$$\psi(\vec{p}) = \sum_i \log\frac{1 - p_i}{p_i}$$

Третий вариант: функция Типпита

$$\psi(\vec{p}) = \max_i\{1 - p_i\}$$

Используют и менее обоснованные вещи. Например, $\hat{\psi}(\vec{p}) = -\frac{1}{d}\sum p_i$