Metody pro ověřování normality

- obecné metody pro ověřování shody teoretického a empirického rozdělení (Vyhovují data naší představě; našemu modelu?)

1.) Grafické metody

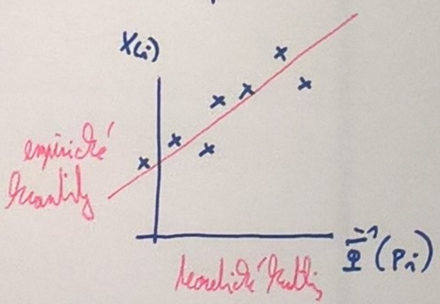
a) histogram
b) jádrový odhad hustoty
} do jednoho grafu + graf teoretické hustoty s odhadnutými parametry normálního rozdělení

c) boxplot

d) Q-Q plot = porovnává teoretické a empirické kvantily

idea: uspořádám pozorování $x_{(1)} \le x_{(2)} \le \dots \le x_{(n)}$, kde $x_{(i)}$ je $p_i = \frac{i-\beta}{n+1-2\beta}$ $(0 \le \beta < 1)$ – kvantil.

v praxi: $\beta = 0{,}5 \dots p_i = \frac{i-0{,}5}{n}$

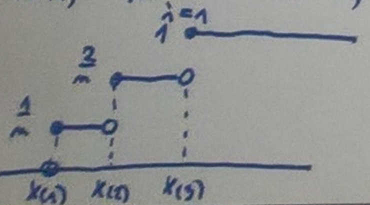


$\beta = 0{,}3175$ NP plot (normal-probability plot)

Q-Q plot je graf $[\Phi^{-1}(p_i), x_{(i)}]$ pro $i = 1, \dots, n$.

e) P-P plot (percent-percent, probability-probability) = porovnává empirickou a teoretickou distribuční funkci

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \le x\}$$



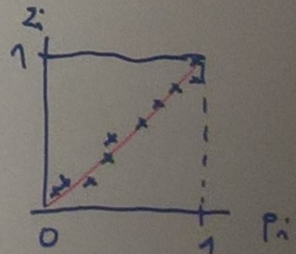
$p_i = \frac{i}{n+1}$, $i = 1, \dots, n$

$$Z_i = \Phi\left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\right)$$

$\hat{\mu} = \bar{X}$

$\hat{\sigma} = S$

P-P plot je graf $[p_i, Z_i]$ pro $i = 1, \dots, n$.



2.) Statistické testy

$X_1, \dots, X_n$ je náh. výběr z rozdělení s distr. fcí F

H_0: F je distribuční funkce normálního rozdělení (μ a σ^2 jsou rušivé parametry)

H_1: F není . . . . .

a) testy založené na momentech

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\}^{3/2}} \quad \dots \text{ výběrová šikmost}$$

$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\}^{2}} \quad \dots \text{ výběrová špičatost}$$

Tvrzení

Nechť $X_1, \dots, X_n$ je náh. výběr z normálního rozdělení. Pak veličiny γ_1 a γ_2 jsou asymptoticky normální a navíc asymptoticky nekorelované. Dále platí

$$E\gamma_1 = 0, \quad E\gamma_2 = \frac{3(n-1)}{n+1}, \quad D(\gamma_1) = \frac{6(n-2)}{(n+1)(n+3)} \quad \text{a} \quad D(\gamma_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

Pozn.

$$\frac{\gamma_1 - E\gamma_1}{\sqrt{D(\gamma_1)}} \overset{H_0}{\approx} N(0,1) \quad \text{a} \quad \frac{\gamma_2 - E\gamma_2}{\sqrt{D(\gamma_2)}} \overset{H_0}{\approx} N(0,1).$$

testová statistika založená na šikmosti

testová statistika založená na špičatosti

Aproximace využitelná pro n velké ($n \geq 200$, resp. $n \geq 500$).

Jarqueho-Berův test – využívá šikmosti i špičatosti zároveň

$$JB = \frac{n}{6}\left(\gamma_1^2 + \frac{(\gamma_2 - 3)^2}{4}\right) \overset{H_0}{\approx} \chi^2_2.$$

Použitelné pro $n \geq 200$.

b) regresní testy

Shapirův-Wilkův test

- je založen na porovnání 2 odhadů σ^2 – výběrového rozptylu a nejlepšího odhadu σ^2 získaného metodou nejmenších čtverců

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \quad \text{kde } (a_1, \dots, a_n)^T = \frac{\mathbf{m}^T \mathbf{V}^{-1}}{(\mathbf{m}^T \mathbf{V}^{-1} \mathbf{V}^{-1} \mathbf{m})^{\frac{1}{2}}}, \quad \text{kde } \mathbf{m} = (m_1, \dots, m_n)^T, \quad \mathbf{V} = (v_{ij})_{i,j=1}^{n}$$

$m_i = E Y_{(i)}$, kde $Y_1, \dots, Y_n$ je náh. výběr z $N(0,1)$.

$v_{ij} = C(Y_{(i)}, Y_{(j)})$

$Y_{(1)} \le Y_{(2)} \le \dots \le Y_{(n)}$ je usp. náh. výběr

$X_1, \dots, X_n$ náh. výběr z $N(\mu, \sigma^2)$

$+\sigma m_i - \sigma m_i$

$X_{(i)} = \mu + \sigma \cdot Y_{(i)}$ pro $i = 1, \dots, n$

$E X_{(i)} = \mu + \sigma m_i$

$X_{(i)} = \mu + \sigma m_i + e_i$, kde $e_i = \sigma (Y_{(i)} - m_i)$ jsou chyby modelu s varianční maticí $\sigma^2 \cdot \mathbf{V}$.

$$\mathbf{X}_{(\cdot)} = \begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{pmatrix}$$

odhadneme σ metodou nejmenších čtverců $\hat{\sigma} = \dfrac{\mathbf{m}^T \mathbf{V}^{-1} \mathbf{X}_{(\cdot)}}{\mathbf{m}^T \mathbf{V}^{-1} \mathbf{m}}$.

$$W = \frac{(\mathbf{m}^T \mathbf{V}^{-1} \mathbf{m})^2}{\mathbf{m}^T \mathbf{V}^{-1} \mathbf{V}^{-1} \mathbf{m}} \cdot \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Pozn.

- prahy V jsou známé pro $n \le 20$, pro $n > 20$ se používají aproximace.
- $W \le 1$. Pro alternativu svědčí malé hodnoty W. Rozdělení W za platnosti H_0 je tabelováno.
- test se hodí pro malé rozsahy výběrů ($n \le 50$).

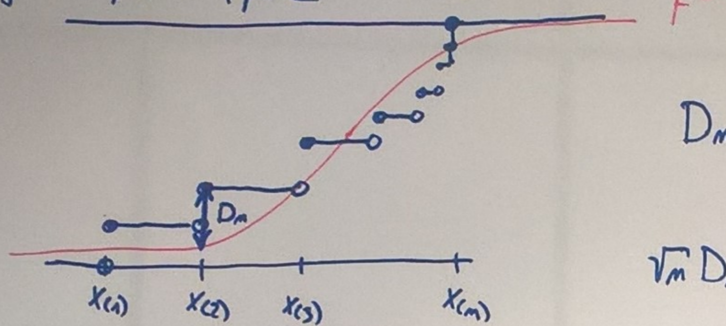
c) Testy založené na empirické distribuční funkci

Kolmogorovův - Smirnovův test

$H_0^*: F = F^*$, kde F^* je distribuční funkce $N(\mu, \sigma^2)$, μ a $\sigma^2 > 0$ známé.

$H_1^*: F \neq F^*$

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}$$



$$D_n = \max_{x \in \mathbb{R}} \left\{ \left| \hat{F}_n(x) - F^*(x) \right| \right\} = \max_{i=1,\dots,n} \left\{ \left| \frac{i}{n} - \Phi\left(\frac{X_{(i)} - \mu}{\sigma} \right) \right| \right\}$$

$\sqrt{n}\, D_n \overset{H_0^*}{\sim} \sup_{t \in [0,1]} |B(t)|$, kde $B(t)$ je Brownův most $C(0,1)$.

Rozdělení n.v. $Y = \sup_{t \in [0,1]} |B(t)|$ je známé, ale nedá se vyjádřit v uzavřené formě.

$$F_Y(y) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2 y^2}, \quad y > 0$$

$$\doteq 1 - 2 e^{-2y^2}$$

$(1-\alpha)$-kvantil je přibližně $\sqrt{\frac{1}{2} \log \frac{2}{\alpha}}$, spec. pro $\alpha = 0{,}05$ je 95%-kvantil 1,36.

Test lze použít jen pro μ a σ^2 známé.

Lillieforsova modifikace K-S testu

- zde již testujeme původní H_0 (F je normální s neznámými parametry)

μ a σ^2 odhadneme z dat $\Rightarrow \hat{\mu}, \widehat{\sigma^2}$

$$D_n^* = \max_{i=1,\dots,n} \left\{ \left| \frac{i}{n} - \Phi\left(\frac{X_{(i)} - \hat{\mu}}{\hat{\sigma}} \right) \right| \right\}$$

. Testová statistika $\sqrt{n}\, D_n^*$ již nemá za H_0 výše uvedené rozdělení $\Rightarrow$ nutno upravit „kvantily".
pro $n \ge 30$ a $\alpha = 0{,}05$ se používá 0,886.

další třída testů: $n\cdot\int_{-\infty}^{\infty}\left(\hat F_n(x)-F^*(x)\right)^2 \underbrace{\psi(F^*(x))}_{\text{váhová funkce}} f^*(x)\,dx$

Cramérův-von Misesův test

$\psi(y)=1$

$$CvM = n\int_{-\infty}^{\infty}\left(\hat F_n(x)-F^*(x)\right)^2 f^*(x)\,dx = \frac{1}{12n} + \sum_{i=1}^{n}\left(p_{(i)} - \frac{2i-1}{2n}\right)^2$$, kde $p_{(i)} = \Phi\left(\frac{x_{(i)}-\hat\mu}{\hat\sigma}\right)$.

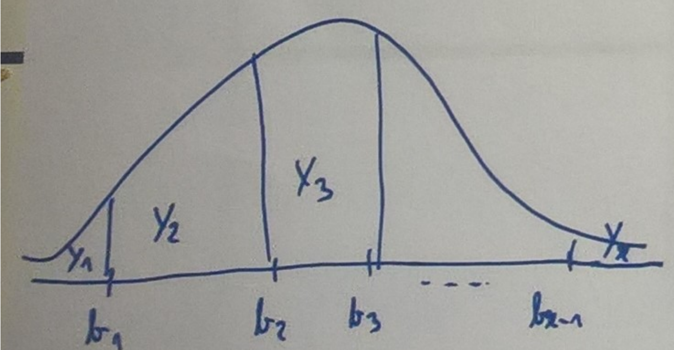
Andersonův-Darlingův test

$\psi(y) = \frac{1}{y(1-y)}$, $0<y<1$ ... dává větší váhu pozorováním na chvostech

$$AD = n\int_{-\infty}^{\infty}\frac{\left(\hat F_n(x)-F^*(x)\right)^2}{F^*(x)\left(1-F^*(x)\right)} f^*(x)\,dx = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\log p_{(i)} + \log\left(1-p_{(n-i+1)}\right)\right).$$

d) testy dobré shody

Pearsonův χ^2-test dobré shody



označme Y_i = počet pozorování, které padnou do intervalu $(b_{i-1}, b_i]$ pro $i=1,\dots,k$

$b_0=-\infty$, $b_k=\infty$

$p_i = P(\text{dané pozorování padne do intervalu } (b_{i-1}, b_i]) = p_i(\mu,\sigma) = P(X_j\in(b_{i-1}, b_i]) = \int_{b_{i-1}}^{b_i} f(x,\mu,\sigma^2)\,dx$.

np_i = očekávaný počet pozorování, které padnou do intervalu $(b_{i-1}, b_i]$

$\|$

$np_i(\mu,\sigma)$

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(Y_i - n p_i(\mu, \sigma)\right)^2}{n p_i(\mu, \sigma)} \overset{H_0^*}{\approx} \chi^2_{k-1}$$

– Testujeme H_0^*, tedy μ a σ^2 známé.

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(Y_i - n p_i(\hat{\mu}, \hat{\sigma})\right)^2}{n p_i(\hat{\mu}, \hat{\sigma})} \overset{H_0}{\approx} \chi^2_{k-3}$$

– Testujeme H_0, tedy μ a σ^2 rušivé parametry.

Poznámka

Jak volit intervaly $(b_{i-1}, b_i]$?

$$p_i(\hat{\mu}, \hat{\sigma}) = \frac{1}{k}, \quad \forall i = 1, \dots, k$$

$$k \doteq 2 \cdot n^{\frac{2}{5}}$$

$$\doteq 15 \left(\frac{n}{100}\right)^{\frac{2}{5}}$$