

Grafické ověřování a testování vybraných modelů

1 Grafické ověřování empirického rozdělení

Při grafické analýze empirického rozdělení vycházíme z empirické distribuční funkce $F_n(x)$ příslušné k náhodnému výběru $X_1, X_2, \dots, X_n$ a sledujeme, do jaké míry se $F_n(x)$ shoduje s distribuční funkcí $F(x)$, z níž náhodný výběr pochází. Protože shodu $F_n(x)$ s $F(x)$ je často obtížné z jejich grafu posoudit, zavádí se vhodná transformace empirické distribuční funkce, která vede k lineárnímu průběhu této transformované funkce na hodnotách náhodného výběru.

Jestliže náhodná veličina $X \sim N(\mu, \sigma^2)$, pak lze její distribuční funkci napsat ve tvaru

$$F(x) = P(X \leq x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad x \in R, \quad (1)$$

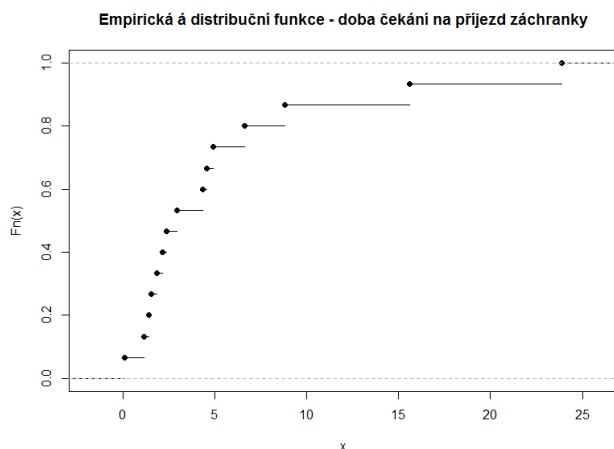
kde Φ je distribuční funkce normálního rozdělení $N(0, 1)$. Aplikujeme-li na vztah (1) funkci Φ^{-1} , která je inverzní funkcí k Φ , dostaneme použitím kvantilů $u_\gamma = \Phi^{-1}(\gamma)$ vztah

$$u_{F(x)} = \Phi^{-1}(F(x)) = \frac{x - \mu}{\sigma},$$

který vyjadřuje kvantil $u_{F(x)}$ jako lineární funkci na x . Vyneseme-li do grafu body $(x, u_{F_n(x)})$ pro hodnoty $x = x_{(1)}, \dots, x_{(n)}$, budou tyto body až na náhodné odchylky seskupeny kolem přímky, za předpokladu, že teoretická distribuční funkce náhodného výběru je distribuční funkcí normálního rozdělení $N(\mu, \sigma^2)$. Z grafu lze také odečíst, že pro $u_{F(x)} = 0$ je $x = \mu$ a dále, že pro $u_{F(t_1)} = 1$ je $t_1 = \mu + \sigma$ a pro $u_{F(t_2)} = -1$ je $t_2 = \mu - \sigma$. Odtud plyne, že σ lze odhadnout veličinou $\hat{\sigma} = (t_1 - t_2)/2$, kde t_1 , respektive t_2 je bod, ve kterém přímka proložená body $(x_{(i)}, u_{F_n(x_{(i)})})$ protíná rovnoběžku s osou x procházející bodem $(0, 1)$ respektive $(0, -1)$.

Příklad

Bylo zaznamenáno $n = 15$ dob čekání na příjezd vozu rychlé záchranné pomoci k silniční nehodě: 0,09; 1,15; 1,41; 1,55; 1,83; 2,15; 2,35; 2,96; 4,34; 4,58; 4,89; 6,06; 8,08; 15,06; 23,09. Uvedené hodnoty byly měřeny v minutách od okamžiku telefonického oznámení nehody na linku tísňového volání do okamžiku příjezdu vozu. Ověřte, zda lze data pokládat za náhodný výběr z normálního rozdělení.

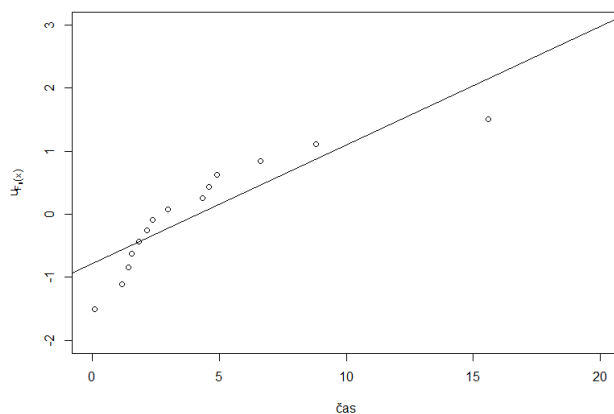


Proložená přímka má rovnici

$$y = -0,79570 + 0,19716x.$$

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

	$x_{(i)}$	$F_n(x_{(i)})$	$u_{F_n(x_{(i)})}$
1	0,09	0,07	-1,50
2	1,15	0,13	-1,11
3	1,41	0,20	-0,84
4	1,55	0,27	-0,62
5	1,83	0,33	-0,43
6	2,15	0,40	-0,25
7	2,35	0,47	-0,08
8	2,96	0,53	0,08
9	4,34	0,60	0,25
10	4,58	0,67	0,43
11	4,89	0,73	0,62
12	6,06	0,80	0,84
13	8,08	0,87	1,11
14	15,06	0,93	1,50
15	23,09	1,00	-



Distribuční funkce exponenciálního rozdělení $Ex(\lambda)$ je $F(x, \lambda) = 1 - e^{-\lambda x}$ pro $x \geq 0$ a jinak je rovna nule. Pro $x \geq 0$ lze ze vztahu pro distribuční funkci stanovit λx . Dostaneme

$$\lambda x = \ln(1 - F(x))^{-1} = \ln \frac{1}{1 - F(x)}.$$

To znamená, že v případě, když náhodný výběr $X_1, \dots, X_n$ je exponenciálního rozdělení s parametrem λ , pak body $(x_{(i)}, \ln \frac{1}{1 - F(x_{(i)})})$ pro $i = 1, 2, \dots, n$ vykazují lineární průběh. Po dosazení $\frac{i}{n}$ za $F(x_{(i)})$ dostaneme, že body $(x_{(i)}, \ln \frac{n}{n-i})$ leží pro výběr z exponenciálního rozdělení přibližně na přímce. Z grafu této přímky lze také přibližně odhadnout λ . Je-li $x = \frac{1}{\lambda}$, platí $1 = \lambda x = \ln \frac{1}{1 - F(x)}$, a tedy x -ovou souřadnici x_0 průsečíku přímky proložené body $(x_{(i)}, \ln \frac{1}{1 - F(x_{(i)})})$ s rovnoběžkou s osou x vedenou bodem $(0, 1)$ lze považovat za odhad $1/\lambda$. Tedy odtud λ je potom přibližně rovno $1/x_0$.

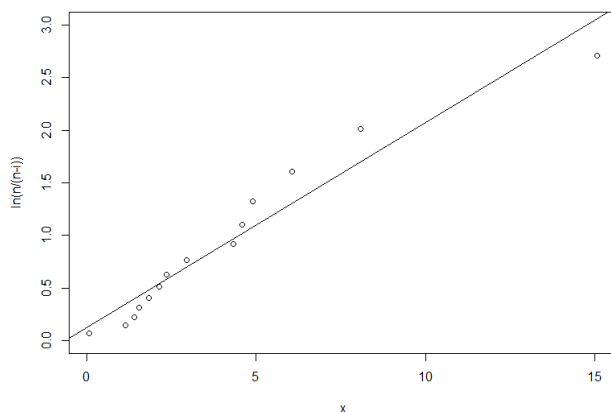
Příklad

Grafickou metodou ověřte, zda data z předchozího příkladu lze považovat za náhodný výběr z exponenciálního rozdělení $Ex(\lambda)$. Z grafu odhadněte parametr λ .

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Podobným postupem jako v předchozím příkladě zde byly vypočteny hodnoty $\ln \frac{n}{n-i}$ pro $n = 15$ a $i = 1, \dots, 14$. Na následujícím obrázku jsou zobrazeny body $(x_{(i)}, \ln \frac{n}{n-i})$, $i = 1, \dots, 14$, a dále potom jimi proložená přímka

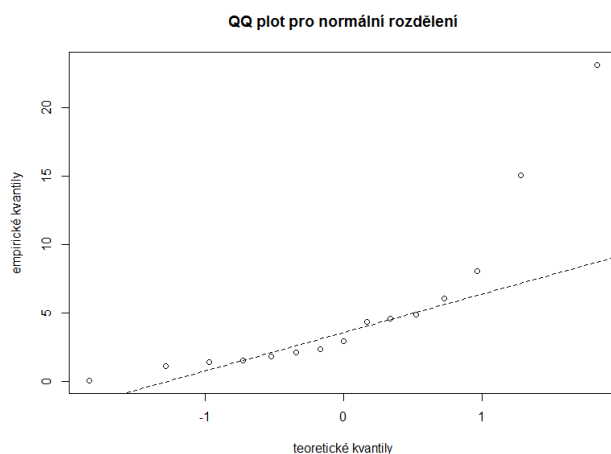
$$y = 0,12274 + 0,19475x.$$



Body vykazují přibližně lineární průběh, tedy lze pokládat náhodný výběr za hodnoty exponenciálního rozdělení. Parametr λ lze potom odhadnout z rovnice přímky tak, že $\hat{\lambda} = \frac{1}{x}$ pro $y = 1$, tedy

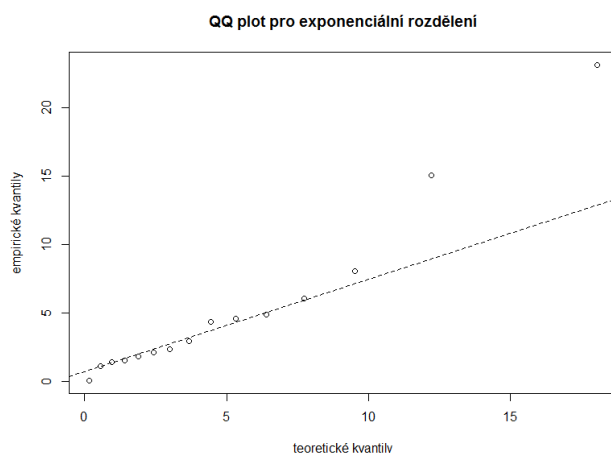
$$\hat{\lambda} = \frac{0,19475}{1 - 0,12274} \doteq 0,22199.$$

Při grafickém ověřování rozdělení se někdy vychází z grafů, které se nazývají QQ plot (z anglického quantile-quantile plot). Tento graf se získá tak, že se na vertikální ose y vynášejí kvantily empirické distribuční funkce $F_n(x)$ a na horizontální osu x se vynášejí kvantily hypotetické distribuční funkce $F(x)$. V programu R QQ plot získáme příkazem `qqnorm` pro normální rozdělení a `qqplot` libovolné rozdělení.



Obr. 1: QQ plot pro ověření normálního rozdělení pro dobu čekání na příjezd vozu rychlé záchranné pomoci

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



Obr. 2: QQ plot pro ověření exponenciálního rozdělení pro dobu čekání na příjezd vozu rychlé záchranné pomoci

2 Testy dobré shody

2.1 χ^2 -test dobré shody

Univerzálním testem k ověřování, zda náhodný výběr pochází z nějakého diskrétního či spojitého rozdělení, je tzv. χ^2 -test dobré shody.

Hodnoty náhodného výběru $x_1, x_2, \dots, x_n$ roztrídíme do k disjunktních tříd, přičemž $n_j, j = 1, 2, \dots, k$, je četnost j -té třídy, resp. j -té obměny a $\hat{\pi}_j$ je hypotetická pravděpodobnost, resp. její odhad, že náhodná veličina X nabude hodnoty z j -té třídy, resp. j -té obměny, počítaná za podmínky, že X má předpokládané rozdělení. Východiskem pro konstrukci testového kritéria je porovnání relativní četnosti n_j/n s hypotetickou pravděpodobností $\hat{\pi}_j$, resp. s jejím odhadem. H_0 : náhodná veličina X má rozdělení daného typu $\rightarrow H_1$: náhodná veličina X nemá rozdělení daného typu. Potom testové kritérium

$$\chi^2 = \sum_{j=1}^k \frac{(n_j - n\hat{\pi}_j)^2}{n\hat{\pi}_j}$$

má za předpokladu platnosti hypotézy H pro velké n (asymptoticky) Pearsonovo χ^2 rozdělení se stupni volnosti $\nu = k - c - 1$, kde c je počet odhadovaných parametrů ověřovaného rozdělení. Kritický obor je

$$W_\alpha = \{\chi^2; \chi^2 \geq \chi_{1-\alpha}^2(\nu)\},$$

kde $\chi_{1-\alpha}^2(\nu)$ je kvantil Pearsonova rozdělení.

Při praktickém provádění testu se požaduje, aby ve všech třídách byly teoretické četnosti větší než 5, tj.

$$n\hat{\pi}_j > 5, \quad j = 1, 2, \dots, k.$$

Není-li tato podmínka splněna, přistupujeme ke slučování tříd.

2.2 Kolmogorov-Smirnovův test

Budeme předpokládat, že $X_1, X_2, \dots, X_n$ je náhodný výběr z rozdělení se spojitou distribuční funkcí, která nezávisí na neznámých parametrech. Chceme testovat nulovou hypotézu, že tato distribuční funkce je rovna dané distribuční funkci F . Je-li $F_n(x)$ výběrová distribuční funkce náhodného výběru $X_1, \dots, X_n$, pak při Kolmogorov-Smirnovově testu se používá statistika

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

Protože distribuční funkce F je neklesající a F_n je po částech konstantní funkce, která má skoky v bodech $X_{(1)}, X_{(2)}, \dots, X_{(n)}$, lze statistiku D přepsat do tvaru, který je výhodnější pro její výpočet. Dostaneme

$$D_n = \max_{1 \leq i \leq n} \{ \max [|F(X_{(i)}) - F_n(X_{(i)})|, |F(X_{(i)}) - F_n(X_{(i-1)})|] \},$$

kde $F(X_{(i)})$ je hodnota hypotetické distribuční funkce F v bodě $X_{(i)}$ a klademe $F_n(X_{(0)}) = 0$.

Nulovou hypotézu zamítneme na hladině významnosti α , když $D_n > D_\alpha(n)$ (kritické hodnoty jsou tabelovány). Je-li n velké, lze na základě aproximace $P(D_n > v) \doteq 2e^{-2nv^2}$, která dává uspokojivé výsledky pro $n > 35$, zamítnout hypotézu H_0 na hladině významnosti α , když pro hodnotu v statistiky D_n , platí $2e^{-2nv^2} < \alpha$. Odtud dostaneme pro asymptotické kritické hodnoty $D_\alpha(n)$ vztah

$$D_\alpha(n) \doteq \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}.$$

Příklady k procvičení

1. Náhodná veličina X popisuje dobu do poruchy vybraného zařízení. Generujte náhodné výběry z rozdělení

- normálního $N(0, 1)$
- normálního $N(10, 16)$
- exponenciálního $Ex(5)$
- Weibullova $W(2, 1)$
- gamma $\Gamma(2, 1)$
- logaritmicko-normálního rozdělení $LN(2, 4)$

postupně rozsahu $n = 5, 10, 30, 50, 100$. Graficky znázorněte výběrovou distribuční funkci, histogram, Q-Q plot a testujte hypotézu, že výběr pochází z hypotetického rozdělení, které vhodně vyberete. Vyberte také rozdělení z něhož je výběr a také rozdělení, které neodpovídá vybrané simulaci. Úlohu řešte také pro reálná data, která vyberete s ohledem na vaši studijní specializaci. Pro řešení využijte podle vlastní volby software STATISTICA, R nebo MATLAB.

2. Náhodná veličina X popisuje počet poruch vybraného zařízení za jednotku času. Generujte náhodné výběry z rozdělení

- binomického rozdělení $Bi(10; 0,1)$



evropský
sociální
fond v ČR



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



UNIVERZITA
OBRANY

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

b) Poissonova rozdělení $Po(2)$

c) negativně binomického rozdělení $NB(2; 0,5)$

postupně rozsahu $n = 5, 10, 30, 50, 100$. Graficky znázorněte výběrovou distribuční funkci, histogram, Q-Q plot a testujte hypotézu, že výběr pochází z hypotetického rozdělení, které vhodně vyberete. Vyberte také rozdělení z něhož je výběr a také rozdělení, které neodpovídá vybrané simulaci. Úlohu řešte také pro reálná data, která vyberete s ohledem na vaši studijní specializaci. Pro řešení využijte podle vlastní volby software STATISTICA, R nebo MATLAB.