Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Štatistická inferencia I

Zadania príkladov a domácich úloh a niektoré riešenia

Stanislav Katina

katina@math.muni.cz

14. decembra 2014

Obsah

1 Model rozdelenia pravdepodobnosti a štatistický model

Príklad 1 (porovnanie dvoch typov modelov) *Model rozdelenia pravdepodobnosti je modelom náhodnej premennej X, napr. model rozdelenia pravdepodobnosti náhodnej premennej X šírka dolnej čeľuste alebo (2) model rozdelenia pravdepodobnosti náhodnej premennej X hrúbka kožných rias u dospelých zdravých žien. Štatistický model je modelom náhodnej premennej $Y|X$ (Y kauzálne závisí na X), napr. (1) model závislosti náhodnej premennej Y šírka dolnej čeľuste na premennej X pohlavie alebo (2) model závislosti náhodnej premennej Y hrúbka kožných rias u dospelých zdravých žien na premennej X BMI. Všimnime si, že náhodné premenné označujeme X alebo Y podľa toho, aký model ich charakterizuje.* pred

Príklad 2 (jednoduchý náhodný výber) *V jednoduchom náhodnom výbere s rozsahom n z populácie s konečným rozsahom N má každý prvok rovnakú pravdepodobnosť vybratia. Ak vyberáme bez vrátenia, hovoríme o* ***jednoduchom náhodnom výbere bez vrátenia***[1] *(Dalgaard 2008). Ak vyberáme s vrátením, hovoríme o* ***jednoduchom náhodnom výbere s vrátením***[2]. *Majme množinu $\mathcal{M}$ s $N = 10$ prvkami a chceme z nej vybrať $n = 3$ prvkov (a) bez vrátenia a (b) s vrátením. Koľko máme možností? Ako vyzerá jedna takáto možnosť, ak ide o množinu $\mathcal{M} = \{1, 2, \ldots, 10\}$. Zopakujte to isté pre $N = 100$, $n = 30$ a množinu $\mathcal{M} = \{1, 2, \ldots, 100\}$.* cvič

Riešenie aj v R

(a) Spolu máme $\binom{N}{n}$ možných náhodných výberov. Ak $N = 10$ a $n = 3$, potom kombinačné číslo $\binom{N}{n} = \frac{N!}{(N-n)!n!} = \binom{10}{3} = 120$ možností. Ak $N = 100$ a $n = 30$, potom $\binom{N}{n} = \binom{100}{30} = 2.937234 \times 10^{25}$ možností.

```
choose(10,3) # pocet vsetkych moznych vyberov bez vratenia
choose(100,30)
library(utils)
combn(10,3) # pocet vsetkych moznych vyberov bez vratenia
combn(100,30)
sample(x=1:10,size=3,replace = FALSE) # jednoduchy nahodny vyber bez vratenia
sample(x=1:100,size=30,replace = FALSE)
```

(b) Spolu máme $\binom{N+n-1}{n}$ možných náhodných výberov. Ak $N = 10$ a $n = 3$, potom $\binom{N+n-1}{n} = \frac{(N+n-1)!}{(N-1)!n!} = \binom{10+3-1}{3} = 220$ možností. Ak $N = 100$ a $n = 30$, potom $\binom{N+n-1}{n} = \binom{100+30-1}{30} = 2.009491 \times 10^{29}$ možností.

```
choose(10+3-1,3) # pocet vsetkych moznych vyberov s vratenim
choose(100+30-1,30)
library(utils)
combn(10+3-1,3) # pocet vsetkych moznych vyberov s vratenim
combn(100+30-1,30)
sample(x=1:10,size=3,replace = TRUE) # jednoduchy nahodny vyber s vratenim
sample(x=1:100,size=30,replace = TRUE)
```

Príklad 3 (jednoduchý náhodný výber) *Nech je skupina ľudí označená identifikačnými číslami (ID) od 1 do 30. Vyberte (a) náhodne 5 ľudí z 30 bez návratu, (b) náhodne 5 ľudí z 30 s návratom a nakoniec (c) náhodne 5 ľudí z 30 bez návratu, kde ľudia s ID od 28 do 30 majú pravdepodobnosť vybratia 4× väčšiu ako ľudia s ID od 1 do 27.* cvič.

[1] Kombinácie bez opakovania n-tej triedy z N prvkov množiny $\mathcal{M}$.

[2] Kombinácie s opakovaním n-tej triedy z N prvkov množiny $\mathcal{M}$.

Riešenie v

```
sample(x=1:30,size=5,replace = FALSE)
sample(x=1:30,size=5,replace = TRUE)
sample(x=1:30,size=5, prob=c(rep(1/39,27),rep(4/39,3)), replace = FALSE)
```

Príklad 4 (normálne rozdelenie) *Majme náhodnú premennú X (môže to byť napr. výška postavy 10-ročných dievčat) a predpokladáme, že má normálne rozdelenie s parametrami μ (stredná hodnota) a σ^2 (rozptyl), čo zapisujeme ako $X \sim N(\mu, \sigma^2)$, $\mu = 140.83, \sigma^2 = 33.79$. Normálne rozdelenie predstavuje model rozdelenia pravdepodobnosti pre túto náhodnú premennú. Vypočítajte pravdepodobnosť $\Pr(a < X < b) = \Pr(X < b) - \Pr(X < a) = F_X(b) - F_X(a)$, kde $a = \mu - k\sigma$, $b = \mu + k\sigma$, $k = 1, 2, 3$.[3] Nakreslite hustotu rozdelenia pravdepodobnosti, vyfarbite oblasť medzi bodmi a a b a popíšte osi x a y tak, ako je uvedené na obrázku 1.* cvič

Riešenie (aj v **)**; (pozri obrázok 1)
$a = \mu - \sigma = 135.0171$, $b = \mu + \sigma = 146.6429$,
$\Pr(|X - \mu| > \sigma) = 0.3173, \Pr(|X - \mu| < \sigma) = 1 - 0.3173 = 0.6827$,
$a = \mu - 2\sigma = 129.2042$, $b = \mu + 2\sigma = 152.4558$,
$\Pr(|X - \mu| > 2\sigma) = 0.0455, \Pr(|X - \mu| < 2\sigma) = 1 - 0.0455 = 0.9545$,
$a = \mu - 3\sigma = 123.3913$, $b = \mu + 3\sigma = 158.2687$,
$\Pr(|X - \mu| > 3\sigma) = 0.0027, \Pr(|X - \mu| < 3\sigma) = 1 - 0.0027 = 0.9973$.

Alternatívny výpočet cez štandardizované normálne rozdelenie (syn. normálne normované rozdelenie) je nasledovný:

```
mu <- 0
sig <- 1
bin <- seq(mu-3*sig,mu+3*sig,by=sig)
pnorm(bin[7]) - pnorm(bin[1]) # 0.9973002
pnorm(bin[6]) - pnorm(bin[2]) # 0.9544997
pnorm(bin[5]) - pnorm(bin[3]) # 0.6826895
```

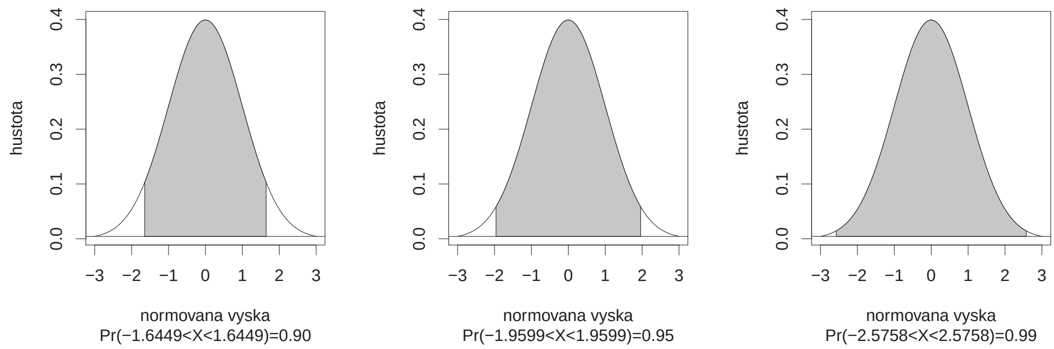
Dostaneme pravidlo $68.27 - 95.45 - 99.73$ (tzv. „**miery normálneho rozdelenia**“).

Príklad 5 (normálne rozdelenie) *Majme $X \sim N(\mu, \sigma^2)$, kde $\mu = 150, \sigma^2 = 6.25$. Vypočítajte $a = \mu - x_{1-\alpha}\sigma$ a $b = \mu + x_{1-\alpha}\sigma$ tak, aby $\Pr(a \le X \le b) = 1 - \alpha$, bola rovná $0.90, 0.95$ a 0.99. Číslo $x_{1-\alpha}$ je kvantil normálneho normovaného rozdelenia, t.j. $\Pr(Z = \frac{X-\mu}{\sigma} < x_{1-\alpha}) = 1 - \alpha, Z \sim N(0,1)$. Nakreslite hustotu rozdelenia pravdepodobnosti, vyfarbite oblasť medzi bodmi a a b a popíšte osi x a y tak, ako je uvedené na obrázku 2.* cvič

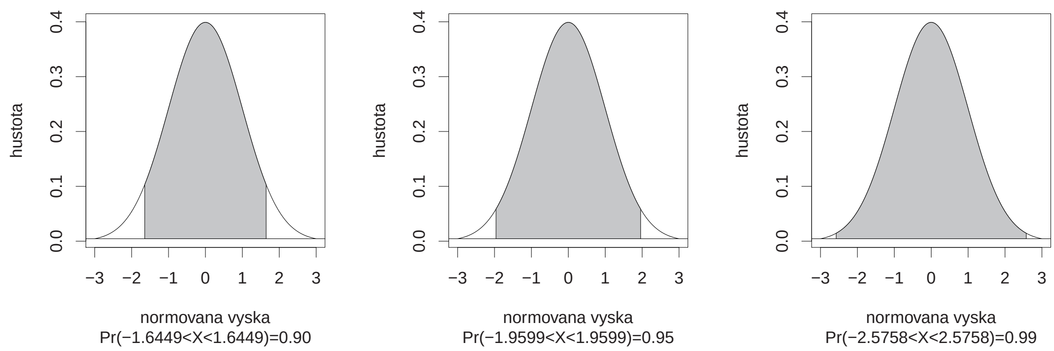
Dostaneme pravidlo $90 - 95 - 99$ (tzv. „**upravené miery normálneho rozdelenia**“). Použili sme nerovnosť $\Pr\left(u_{\alpha/2} < Z < u_{1-\alpha/2}\right) = \Phi\left(u_{1-\alpha/2}\right) - \Phi\left(u_{\alpha/2}\right) = 1 - \alpha$, kde Φ je distribučná funkcia normálneho normovaného rozdelenia a všeobecne $\alpha \in (0, 1/2)$; v príklade $\alpha = 0.1$, 0.05 a 0.01. Pozri obrázok 2.

Príklad 6 (normálne rozdelenie) *Predpokladajme model normálneho rozdelenia $N(132, 13^2)$ pre systolický krvný tlak. Aká časť populácie (v %) bude mať hodnoty väčšie ako 160 mm Hg?* cvič

[3]Pravdepodobnosť $\Pr(a < X < b) = \Pr(a \le X \le b)$, pretože pravdepobnosť v bode (tu a a b) je rovná nule pre spojité premenné, t.j. $\Pr(a) = \Pr(b) = 0$. Pre diskrétne premenné to neplatí.



Obr. 1: Upravené miery normálneho rozdelenia; krivka hustoty s vyfarbeným obsahom pod touto krivkou medzi príslušnými kvantilmi na osi x; obsah je rovný pravdepodobnosti výskytu subjektov s danou normovanou výškou v rozpätí týchto kvantilov



Obr. 2: Upravené miery normálneho rozdelenia; krivka hustoty s vyfarbeným obsahom pod touto krivkou medzi príslušnými kvantilmi na osi x; obsah je rovný pravdepodobnosti výskytu subjektov s danou normovanou výškou v rozpätí týchto kvantilov

Príklad 7 (binomické rozdelenie) *Predpokladajme, že počet ľudí uprednostňujúcich liečbu A pred liečbou B sa správa podľa modelu binomického rozdelenia s parametrami p (pravdepodobnosť výskytu udalosti) a N (rozsah náhodného výberu), ozn.* $Bin\,(N, p)$, *kde* $N = 20, p = 0.5$, *t.j. ľudia preferujú oba typy liečby rovnako. (a) Aká je pravdepodobnosť, že bude 16 a viac pacientov uprednostňovať liečbu A pred liečbou B? (b) Aká je pravdepodonosť, že bude 16 a viac a zároveň 4 alebo menej pacientov uprednostňovať liečbu A pred liečbou B?* cvič

Riešenie (aj v R)
(a) $\Pr(X \geq 16) = 1 - \sum_{i:x_i \leq 15} \Pr(X = x_i) = 1 - \sum_{i:x_i \leq 15} \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i} = 1 - \sum_{i:x_i \leq 15} \binom{20}{x_i} 0.5^{x_i} (1-0.5)^{20-x_i} = 0.006$.

```
pbinom(16,size=20,prob=0.5) # 0.9987116
1-pbinom(16,size=20,prob=0.5) # 0.001288414
```

Z vyššie uvedeného R-kódu vyplýva, že ide o pravdepodobnosť $\Pr(X \leq 16)$ a $\Pr(X > 16)$, ale my potrebujeme $\Pr(X \geq 16)$. Preto R-kód upravíme nasledovne

```
1-pbinom(15,size=20,prob=0.5) # 0.005908966
```

```
sum(choose(20,16:20)*0.5^(16:20)*0.5^(20-16:20)) # 0.005908966
```

(b) $\Pr(X \leq 4, X \geq 16) = 1 - \sum_{i:x_i \leq 15} \Pr(X = x_i) + \sum_{i:x_i \leq 4} \Pr(X = x_i) = 0.012$. Táto pravdepodobnosť je dvojnásobkom predchádzajúcej pravdepodobnosťi, lebo $Bin(N, 0.5)$ je symetrické okolo 0.5, t.j.

```
1-pbinom(15,size=20,prob=0.5) + pbinom(4,size=20,prob=0.5) # 0.01181793
```

Príklad 8 (binomické rozdelenie) *Predpokladajme, že* $\Pr(\textit{vír}) = 0.533 = p_1$ *je pravdepodobnosť výskytu dermatoglyfického vzoru vír na palci pravej ruky mužov českej populácie a* $\Pr(\textit{ostatné}) = 0.467 = p_2$ *je pravdepodobnosť výskytu ostatných vzorov na palci pravej ruky mužov českej populácie, pričom* X *je počet vírov a* Y *je počet ostatných vzorov, kde* $X \sim Bin(N, p_1)$ *a* $Y \sim Bin(N, p_2)$. *Vypočítajte (1)* $\Pr(X \leq 120)$*, keď* $N = 300$ *a (2)* $\Pr(Y \leq 120)$*, keď* $N = 300$. cvič

Príklad 9 (parametre) *Príklady parametrov* θ *– stredná hodnota* μ*, rozptyl* σ^2*, korelačný koeficient* ρ*, pravdepodobnosť* p *výskytu nejakej udalosti, rozdiel dvoch stredných hodnôt* $\mu_1 - \mu_2$*, podiel dvoch rozptylov* σ_1^2/σ_2^2*, rozdiel dvoch korelačných koeficientov* $\rho_1 - \rho_2$*, rozdiel dvoch pravdepodobností* $p_1 - p_2$ *a pod.* pred

Príklad 10 (binomické rozdelenie) *Ak* $X \sim Bin(N, \theta), \theta = p \in \langle 0, 1 \rangle$*, potom* $\mathcal{Y}_\theta$ *je rovnaký pre všetky* θ *a koinciduje s výberovým priestorom* $\mathcal{Y} = \{0, 1, \ldots, N\}$. pred

Príklad 11 (aproximácia binomického rozdelenia normálnym) [4]*Nech* $\Pr(\textit{muž}) = 0.515$ *znamená pravdepodobnosť výskytu mužov v populácii a* $\Pr(\textit{žena}) = 0.485$ *pravdepodobnosť výskytu žien. Nech* X *je počet mužov a* Y *počet žien. Za predpokladu modelu* $Bin(N, p)$ *vypočítajte (a)* $\Pr(X \leq 3)$*, ak* $N = 5$*, (b)* $\Pr(X \leq 5)$*, ak* $N = 10$ *a (c)* $\Pr(X \leq 25)$*, ak* $N = 50$*. Porovnajte vypočítané pravdepodobnosti s pravdepodobnosťami aproximovanými normálnym rozdelením* $N(Np, Npq)$*. Nakreslite hustotu rozdelenia pravdepodobnosti normálneho rozdelenia a superponujte ju pravdepodobnostou funkciou binomického rozdelenia tak, ako je uvedené na obrázku 3. Nakreslite distribučnú funkciu normálneho rozdelenia a superponujte ju distribučnou funkciou binomického rozdelenia tak, ako je uvedené na obrázku 4.* cvič

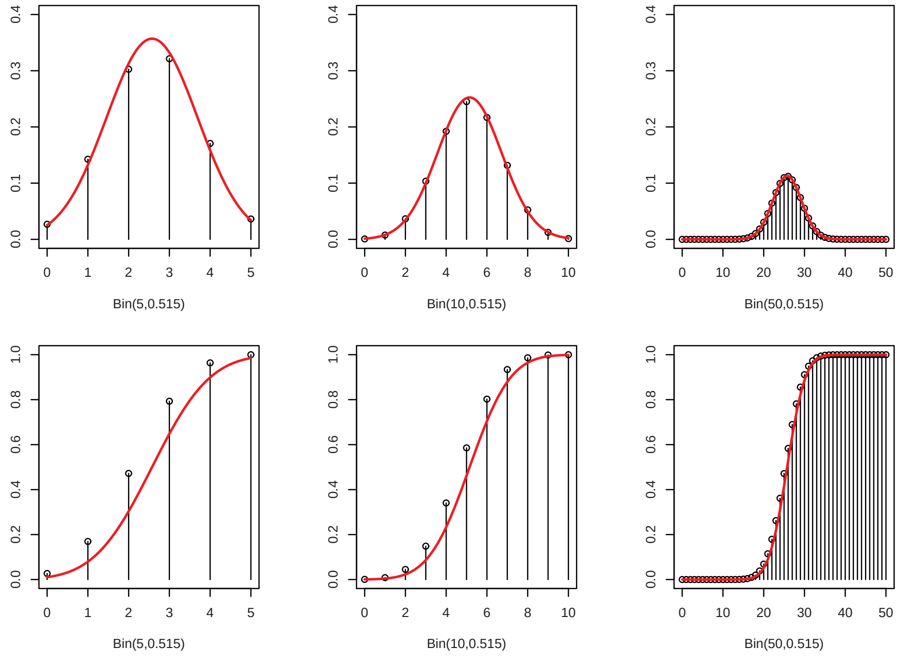
Príklad 12 (normálne rozdelenie) *Model pre náhodný výber* $X_1, X_2, \ldots, X_n$ *je* $N(\mu, \sigma^2)$ *a hovoríme, že* $X_1, X_2, \ldots, X_n$ *pochádza z normálneho rozdelenia, t.j.* $X \sim N(\mu, \sigma^2)$*. Parameter modelu* $N(\mu, \sigma^2)$ *je vektor* $\boldsymbol{\theta} = (\mu, \sigma^2)$*. Hustota tohto rozdelenia má tvar* $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$. pred

Príklad 13 (štandardizované normálne rozdelenie) *Model pre náhodný výber* $X_1, X_2, \ldots, X_n$ *je* $N(0, 1)$ *a hovoríme, že* $X_1, X_2, \ldots, X_n$ *pochádza zo štandardizovaného normálneho rozdelenia, t.j.* $X \sim N(\mu, \sigma^2)$*, kde* $\mu = 0$ *a* $\sigma^2 = 1$*. Parameter modelu* $N(\mu, \sigma^2)$ *je vektor* $\boldsymbol{\theta} = (0, 1)$*. Hustota tohto rozdelenia má tvar* $\phi(x) = f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}$. pred

Príklad 14 (dvojrozmerné normálne rozdelenie) *Náhodný vektor* $(X, Y)^T$ *má dvojrozmerné normálne rozdelenie*

$$N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \ \textit{kde } \boldsymbol{\mu} = (\mu_1, \mu_2)^T \ \textit{ a } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

[4]Aproximácia znamená „približné vyjadrenie“, t.j. buď nejaké rozdelenie aproximujeme iným (majúcim isté výhody oproti tomu, ktoré aproximujeme), alebo aproximujeme dáta nejakým rozdelením (ktoré popisuje dáta pomocou ľahko interpretovateľných parametrov).



Obr. 3: Aproximácia binomického rozdelenia normálnym pre $p = 0.515$ a $N = 5, 10$ a 50; spojnicový graf superponovaný hustotou (prvý riadok) a distribučnou funkciou (druhý riadok)

s hustotou

$$f(x,y) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left\{\tfrac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\tfrac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \tfrac{(y-\mu_2)^2}{\sigma_2^2}\right\}\right\},$$

kde $(x,y)^T \in \mathbb{R}^2$, $\mu_j \in \mathbb{R}^1$, $\sigma_j^2 > 0$, $j = 1,2$, $\rho \in \langle -1, 1\rangle$ *sú parametre, potom* $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. *Výraz v exponente môžeme písať ako*

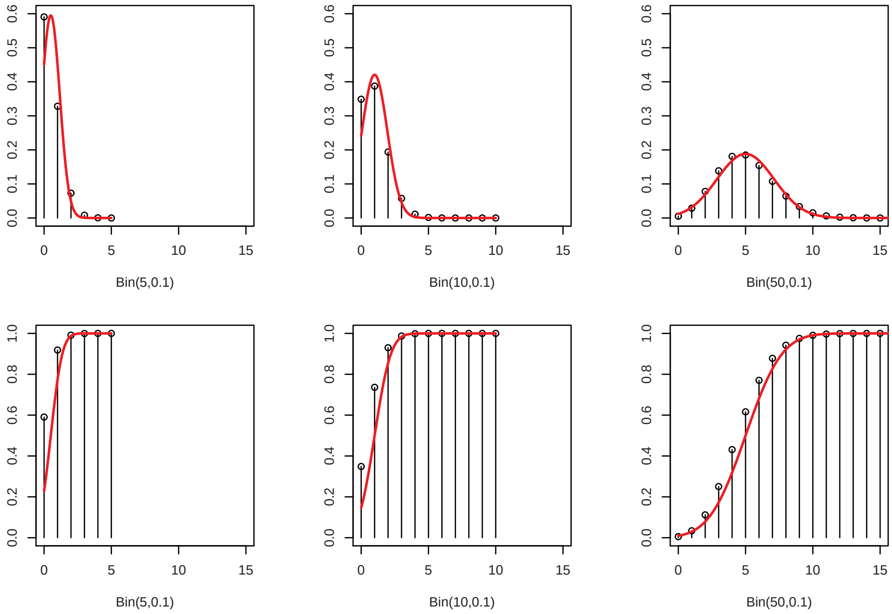
$$-\frac{1}{2}\begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix},$$

marginálne rozdelenia[5] *sú* $X \sim N(\mu_1, \sigma_1^2)$ *a* $Y \sim N(\mu_2, \sigma_2^2)$, ρ *je koeficient korelácie*[6] *(pozri obrázok 5).* cvič.

Príklad 15 (dvojrozmerné normálne rozdelenie) *(1) Nakreslite hustotu dvojrozmerného normálneho rozdelenia* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *pomocou funkcie* `image()` *a superponujte ho s kontúrovým grafom hustoty toho istého rozdelenia pomocou funkcie* `contour()`*. (2) Nakreslite hustotu dvojrozmerného normálneho rozdelenia* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *pomocou funkcie* `persp()`*. Hustotu rozsekajte na 12 intervalov, kde hodnoty v týchto intervaloch budú zodpovedať farbám* `terrain.colors(12)`*. Použite nasledovné parametre*
(a) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1,\ \rho = 0$;
(b) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1,\ \rho = 0.5$;
(c) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1.2,\ \rho = 0.5$.
Vzorové riešenie pozri na obrázku 5. cvič.

[5] Margináne rozdelenie je rozdelenie marginálnej náhodnej premennej, tu X nezávisle na Y a naopak Y nezávisle na X.

[6] Z tohto príkladu je zrejmé, že na dostatočný popis dvojrozmerného normálneho rozdelenia potrebujeme päť parametrov, t.j. strednú hodnotu a rozptyl pre marginálne rozdelenie náhodných premenných X a Y a korelačný koeficient $\rho = \rho(X, Y)$ popisujúci silu lineárneho vzťahu X a Y.



Obr. 4: Aproximácia binomického rozdelenia normálnym pre $p = 0.1$ a $N = 5, 10$ a 50; spojnicový graf superponovaný hustotou (prvý riadok) a distribučnou funkciou (druhý riadok)

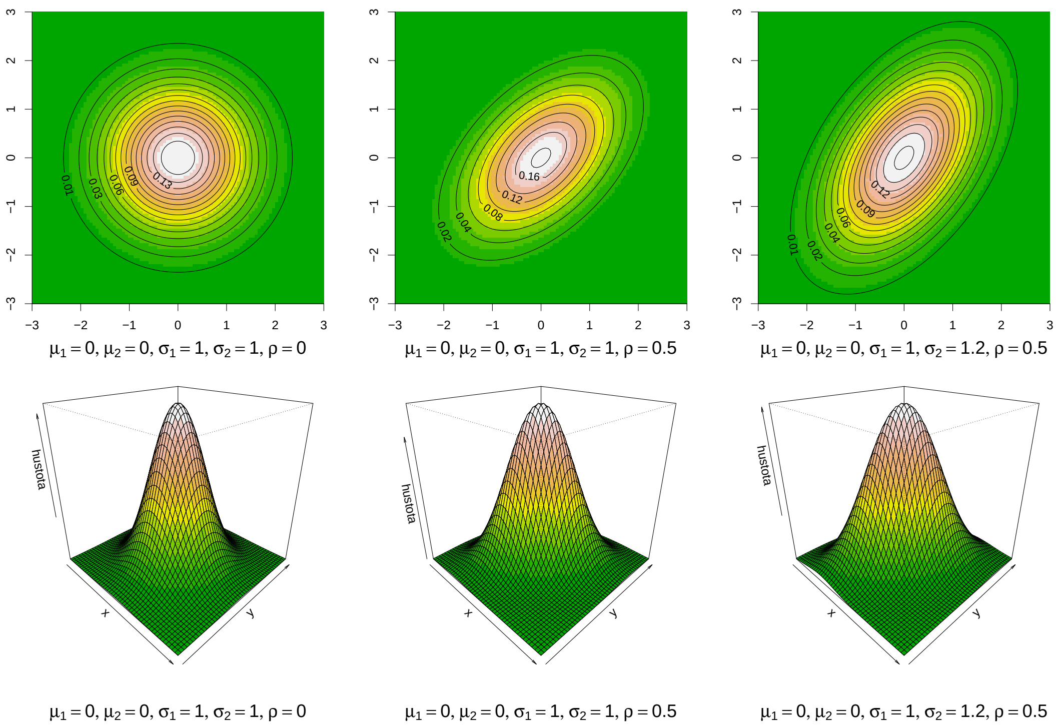
Príklad 16 (dvojrozmerné normálne rozdelenie) *Nech náhodnou premennou X je najväčšia výška mozgovne u mužov (`skull.pH`; v mm) a náhodnou premennou Y je morfologická výška tváre u mužov (`face.H`; v mm); dáta: `one-sample-correlation-skull.txt`. Nech $E[X] = \mu_1$ je stredná hodnota najväčšej výšky mozgovne a $Var[X] = \sigma_1^2$ je rozptyl najväčšej výšky mozgovne, $E[Y] = \mu_2$ je stredná hodnota morfologickej výšky tváre a $Var[Y] = \sigma_2^2$ je rozptyl morfologickej výšky tváre. Predpokladajme, že najväčšia výška mozgovne X má normálne rozdelenie $N(\mu_1, \sigma_1^2)$ a morfologická výška tváre Y má normálne rozdelenie $N(\mu_2, \sigma_2^2)$. Potom $(X, Y)^T$ má dvojrozmerné normálne rozdelenie $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ s parametrami $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$, čo je vektor stredných hodnôt a σ_1^2, σ_2^2 a ρ, čo sú parametre kovariančnej matice $\boldsymbol{\Sigma}$, kde sila lineárneho vzťahu týchto dvoch premenných je daná veľkosťou a znamienkom ρ. Potom $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$. (1) Nakreslite hustotu dvojrozmerného normálneho rozdelenia $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pomocou funkcie `image()` a superponujte ho s kontúrovým grafom hustoty toho istého rozdelenia pomocou funkcie `contour()`. (2) Nakreslite dvojrozmerný jadrový odhad hustoty pomocou funkcií `kde2d()` a `image()` a superponujte ho s kontúrovým grafom hustoty dvojrozmerného normálneho rozdelenia $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pomocou funkcie `contour()`. Hustotu rozsekajte na 12 intervalov, kde hodnoty v týchto intervaloch budú zodpovedať farbám `terrain.colors(12)`. Namiesto $\boldsymbol{\theta}$ použite vektor $\widehat{\boldsymbol{\theta}}$ odhadnutý z dát. Riešenie pozri na obrázku 6.* cvič.

Príklad 17 (štandardizované dvojrozmerné normálne rozdelenie) *Náhodný vektor $(X, Y)^T$ má dvojrozmerné normálne rozdelenie*

$$N_2(\mathbf{0}, \boldsymbol{\Sigma}), \ \textit{kde } \mathbf{0} = (0, 0)^T \ \textit{ a } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

s hustotou

$$\phi(x, y) = f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right\},$$



Obr. 5: Hustoty dvojrozmerného normálneho rozdelenia pri rôznych parametroch (prvý riadok – kontúrový graf, druhý riadok – perspektívny trojrozmerný graf v podobe plochy); čím je ρ odlišnejšie od nuly, tým viac sa kontúry líšia od kruhov (menia sa na elipsy); so zväčšujúcim sa rozdielom medzi σ_1 a σ_2 sa zväčšuje rozdiel rozptýlenia koncetrických kruhov v smere jednotlivých osí (hovoríme, že rozdiel variability premenných X_1 a X_2 sa zväčšuje)

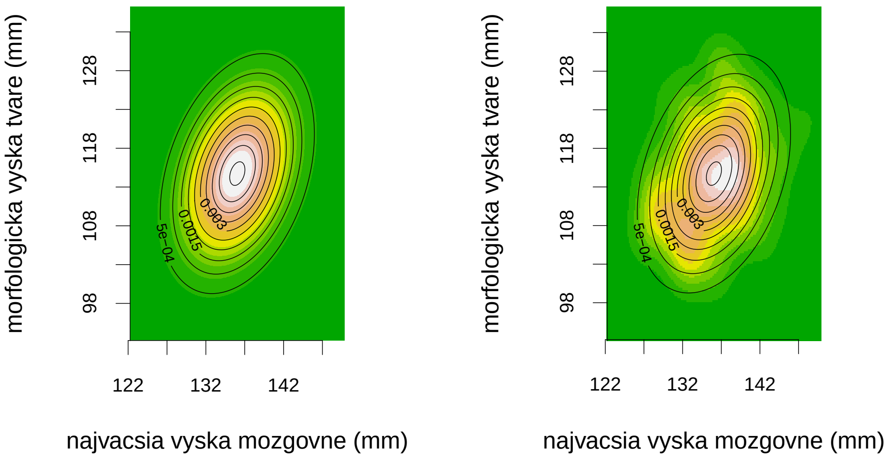
kde $(x, y)^T \in \mathbb{R}^2$, $\rho \in \langle -1, 1\rangle$ sú parametre, potom $\boldsymbol{\theta} = (0, 0, 1, 1, \rho)$. Výraz v exponente môžeme písať ako

$$-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix},$$

marginálne rozdelenia sú obe $N(0, 1)$ a ρ je koeficient korelácie. cvič.

Príklad 18 (štandardizované dvojrozmerné normálne rozdelenie) *Nech náhodnou premennou $X \sim N(\mu_1, \sigma_1^2)$ je najväčšia výška mozgovne (`skull.pH`; v mm) a náhodnou premennou $Y \sim N(\mu_2, \sigma_2^2)$ je morfologická výška tváre (`face.H`; v mm). Nech X a Y majú dvojrozmerné normálne rozdelenie s parametrami $(\mu_1, \mu_2)^T$ a σ_1^2, σ_2^2 a ρ sú parametre kovariančnej matice $\boldsymbol{\Sigma}$. Keď od X odpočítame jej strednú hodnotu μ_1 a tento rozdiel vydelíme odmocninou z rozptylu σ_1, dostaneme náhodnú premennú Z_X, ktorá má asymptoticky normálne rozdelenie so strednou hodnotou $\mu_1 = 0$ a rozptylom $\sigma_1^2 = 1$, čo zapisujeme ako $Z_X \sim N(0, 1)$. Keď od Y odpočítame jej strednú hodnotu μ_2 a tento rozdiel vydelíme odmocninou z rozptylu σ_2, dostaneme náhodnú premennú Z_Y, ktorá má asymptoticky normálne rozdelenie so strednou hodnotou $\mu_2 = 0$ a rozptylom $\sigma_2^2 = 1$, čo zapisujeme ako $Z_Y \sim N(0, 1)$. Potom $(Z_X, Z_Y)^T$ má štandardizované dvojrozmerné normálne rozdelenie $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ s parametrami $\boldsymbol{\mu} = (0, 0)^T$ a $\sigma_1^2 = 1$, $\sigma_2^2 = 1$ a ρ sú parametre kovariančnej matice $\boldsymbol{\Sigma}$.* cvič.

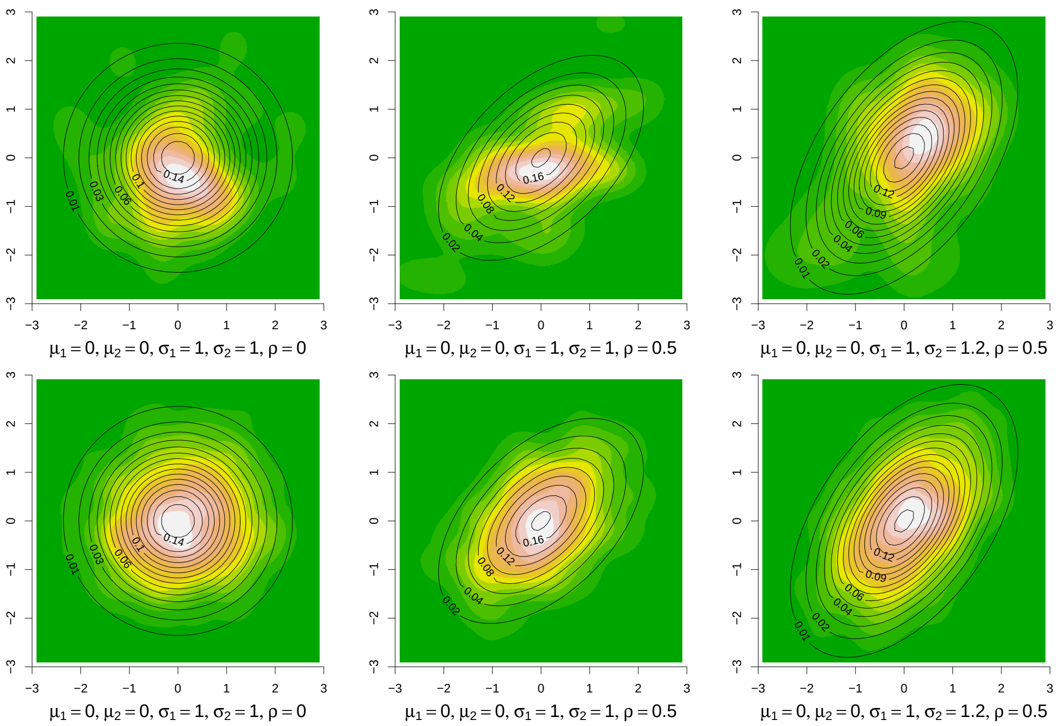
Príklad 19 (dvojrozmerné normálne rozdelenie) *Simuláciu pseudonáhodných čísel z $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ môžeme v R urobiť nasledovne použitím:*
1) knižnice `library(MASS)` a funkcie `mvrnorm()`;



Obr. 6: Hustota dvojrozmerného normálneho rozdelenia s parametrom $\widehat{\boldsymbol{\theta}}$, ktorý je odhadnutý z dát (vľavo) a superimpozícia kontúr hustoty dvojrozmerného normálneho rozdelenia s parametrom $\widehat{\boldsymbol{\theta}}$, ktorý je odhadnutý z dát a dvojrozmerného jadrového odhadu hustoty (vpravo)

2) knižnice `library(mvtnorm)` a funkcie `rmvnorm()`;
3) funkcie `rnorm()` a nasledovného algoritmu – nech $X_1 \sim N(0,1)$ a $X_2 \sim N(0,1)$; potom $(Y_1,Y_2) \sim N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$, kde $\boldsymbol{\mu} = (\mu_1,\mu_2)^T$, čo je vektor stredných hodnôt a σ_1^2, σ_2^2 a ρ, čo sú parametre kovariančnej matice $\boldsymbol{\Sigma}$, kde sila lineárneho vzťahu Y_1 a Y_2 je daná veľkosťou a znamienkom ρ; $Y_1 = \sigma_1 X_1 + \mu_1$ a $Y_2 = \sigma_2(\rho X_1 + \sqrt{1-\rho^2} X_2) + \mu_2$. Nasimulujte pseudonáhodné čísla Y_1 a Y_2 z $N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Vypočítajte dvojrozmerný jadrový odhad hustoty $(Y_1,Y_2)^T$ pomocou funkcie `kde2d()`. Nakreslite ho pomocou funkcie `image()` a superponujte ho s kontúrovým grafom hustoty dvojrozmerného normálneho rozdelenia $N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$ pomocou funkcie `contour()`. Pri simulácii použite nasledovné parametre
(a) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1,\ \rho = 0$; (1) $n = 50$ a (2) $n = 1000$;
(b) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1,\ \rho = 0.5$; (1) $n = 50$ a (2) $n = 1000$;
(c) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1.2,\ \rho = 0.5$; (1) $n = 50$ a (2) $n = 1000$.
Vzorové riešenie pozri na obrázku 6. cvič.

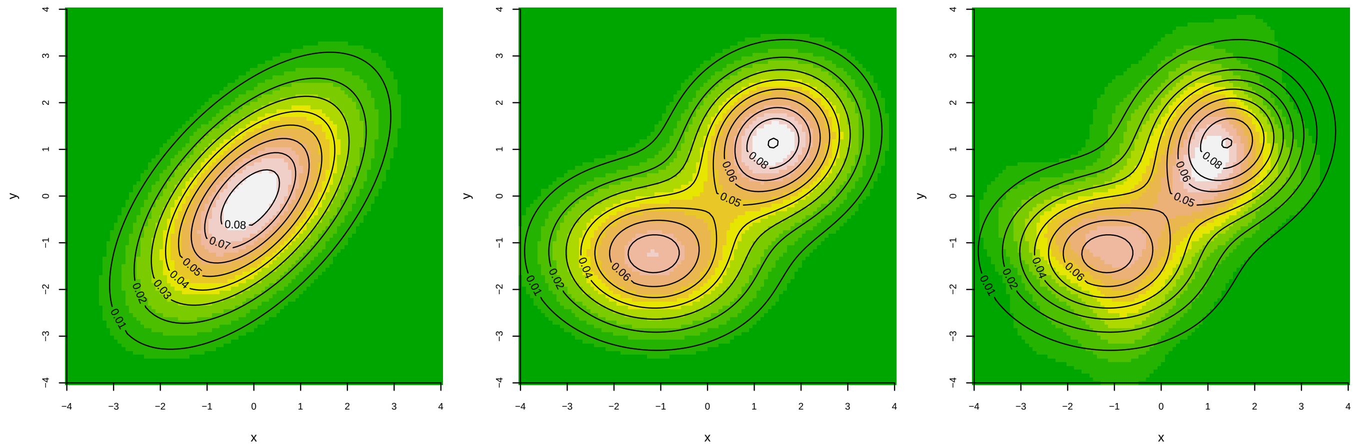
Príklad 20 (dvojrozmerné normálne rozdelenie) *Simuláciu pseudonáhodných čísel z $N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$ môžeme v R urobiť použitím nasledovných alternatívnych funkcií:*
1) knižnice `library(MASS)` a funkcie `mvrnorm()`;
2) knižnice `library(mvtnorm)` a funkcie `rmvnorm()`;
3) funkcie `rnorm()` a nasledovného algoritmu – nech $X_1 \sim N(0,1)$ a $X_2 \sim N(0,1)$; potom $(Y_1,Y_2) \sim N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$, kde $\boldsymbol{\mu} = (\mu_1,\mu_2)^T$, čo je vektor stredných hodnôt a σ_1^2, σ_2^2 a ρ, čo sú parametre kovariančnej matice $\boldsymbol{\Sigma}$, kde sila lineárneho vzťahu Y_1 a Y_2 je daná veľkosťou a znamienkom ρ; $Y_1 = \sigma_1 X_1 + \mu_1$ a $Y_2 = \sigma_2(\rho X_1 + \sqrt{1-\rho^2} X_2) + \mu_2$. Nasimulujte pseudonáhodné čísla Y_1 a Y_2 z $N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Vypočítajte dvojrozmerný jadrový odhad hustoty $(Y_1,Y_2)^T$ pomocou funkcie `kde2d()`. Nakreslite ho pomocou funkcie `image()` a superponujte ho s kontúrovým grafom hustoty dvojrozmerného normálneho rozdelenia $N_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$ pomocou funkcie `contour()`. Hustotu rozsekajte na 12 intervalov, kde hodnoty v týchto intervaloch budú zodpovedať farbám `terrain.colors(12)`. Pri simulácii použite nasledovné parametre
(a) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1,\ \rho = 0$; (1) $n = 50$ a (2) $n = 1000$;
(b) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1,\ \rho = 0.5$; (1) $n = 50$ a (2) $n = 1000$;
(c) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1.2,\ \rho = 0.5$; (1) $n = 50$ a (2) $n = 1000$.
Vzorové riešenie pozri na obrázku 7.



Obr. 7: Hustoty dvojrozmerného normálneho rozdelenia (prvý riadok $n = 50$; druhý riadok $n = 1000$)

Príklad 21 (zmes dvoch dvojrozmerných normálnych rozdelení) *Simuláciu pseudonáhodných čísel zo zmesi dvoch normálnych rozdelení* $pN_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-p)N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *môžeme v R urobiť použitím jedného z alternatínych postupov z príkladu 20. Nasimulujte pseudonáhodné čísla* X *a* Y *(1) zo zmesi* $pN_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-p)N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$*, kde* $\boldsymbol{\theta} = (\mu_{11}, \mu_{12}, \sigma^2_{11}, \sigma^2_{12}, \rho_1, \mu_{21}, \mu_{22}, \sigma^2_{21}, \sigma^2_{22}, \rho_2)^T$ *a (2) z dvojrozmerného rozdelenia* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, kde parametre predstavujú spoločný vektor stredných hodnôt a spoločnú kovariančnú maticu. t.j.* $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2_1, \sigma^2_2, \rho)^T$*. Pre (1) vypočítajte dvojrozmerný jadrový odhad hustoty* $(X, Y)^T$ *pomocou funkcie* `kde2d()`*.*
(a) Nakreslite teoretickú hustotu (2) pomocou funkcie `image()` *a superponujte ju s kontúrovým grafom teoretickej hustoty (2) pomocou funkcie* `contour()`*.*
(b) Nakreslite teoretickú hustotu (1) pomocou funkcie `image()` *a superponujte ju s kontúrovým grafom teoretickej hustoty (1) pomocou funkcie* `contour()`*.*
(c) Nakreslite dvojrozmerný jadrový odhad hustoty realizácí (1) pomocou funkcie `image()` *a superponujte ju s kontúrovým grafom teoretickej hustoty (1) pomocou funkcie* `contour()`*.*
Hustotu rozsekajte na 12 intervalov, kde hodnoty v týchto intervaloch budú zodpovedať farbám `terrain.colors(12)`*. Pri simulácii použite* $\boldsymbol{\theta} = (-1.2, -1.2, 1, 1, 0, 1, 1, 1, 1, 0)^T$,
(1) $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_{11}, \widehat{\mu}_{12}, \widehat{\sigma}^2_{11}, \widehat{\sigma}^2_{12}, \widehat{\rho}_1, \widehat{\mu}_{21}, \widehat{\mu}_{22}, \widehat{\sigma}^2_{21}, \widehat{\sigma}^2_{22}, \widehat{\rho}_2)^T$, $n_1 = n_2 = 50$ *a* $p = 0.5$ *(odhady pochádzajú z nasimulovaných dát).*
(2) $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}^2_1, \widehat{\sigma}^2_2, \widehat{\rho})^T$ *a* $n_1 = n_2 = 50$ *(odhady pochádzajú zo spoločného výberu nasimulovaných dát).*
Vzorové riešenie pozri na obrázku 8. DÚ

Príklad 22 (zmes dvoch dvojrozmerných normálnych rozdelení) *Nech* $(X_1, Y_1)^T$ *pochádza z rozdlenia* $N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$*, kde* X_1 *je priemerná dĺžka dolnej končatiny* `lowex.L` *v milimetroch a* Y_1 *dĺžka trupu* `tru.L` *v milimetroch (u mužov). Nech* $(X_2, Y_2)^T$ *pochádza z rozdlenia* $N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$*, kde* X_2 *je*

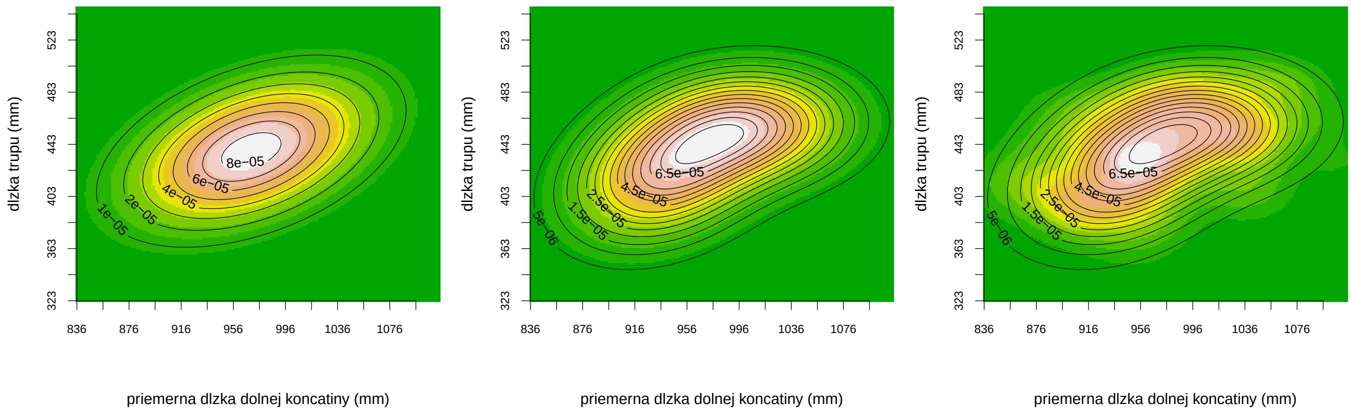


Obr. 8: Spoločná hustota dvojrozmerného normálneho rozdelenia (vľavo), hustota zmesi dvoch dvojrozmerných normálnych rozdelení (uprostred) a dvojrozmerný jadrový odhad superponovaný hustotou zmesi dvoch dvojrozmerných normálnych rozdelení (vpravo)

priemerná dĺžka dolnej končatiny `lowex.L` *v milimetroch a* Y_2 *dĺžka trupu* `tru.L` *v milimetroch (u žien). Predpokladajme, že* X *je priemerná dĺžka dolnej končatiny a* Y *dĺžka trupu pochádzajú (1) zo zmesi* $pN_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-p)N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, *kde* $\boldsymbol{\theta} = (\mu_{11}, \mu_{12}, \sigma_{11}^2, \sigma_{12}^2, \rho_1, \mu_{21}, \mu_{22}, \sigma_{21}^2, \sigma_{22}^2, \rho_2)^T$ *a (2) z dvojrozmerného rozdelenia* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *kde parametre predstavujú spoločný vektor stredných hodnôt a spoločnú kovariančnú maticu. t.j.* $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$. *Pre (1) vypočítajte dvojrozmerný jadrový odhad hustoty* $(X, Y)^T$ *pomocou funkcie* `kde2d()`.
(a) Nakreslite teoretickú hustotu (2) pomocou funkcie `image()` *a superponujte ju s kontúrovým grafom teoretickej hustoty (2) pomocou funkcie* `contour()`.
(b) Nakreslite teoretickú hustotu (1) pomocou funkcie `image()` *a superponujte ju s kontúrovým grafom teoretickej hustoty (1) pomocou funkcie* `contour()`.
(c) Nakreslite dvojrozmerný jadrový odhad hustoty realizácií (1) pomocou funkcie `image()` *a superponujte ju s kontúrovým grafom teoretickej hustoty (1) pomocou funkcie* `contour()`.
Hustotu rozsekajte na 12 intervalov, kde hodnoty v týchto intervaloch budú zodpovedať farbám `terrain.colors(12)`.
(1) $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_{11}, \widehat{\mu}_{12}, \widehat{\sigma}_{11}^2, \widehat{\sigma}_{12}^2, \widehat{\rho}_1, \widehat{\mu}_{21}, \widehat{\mu}_{22}, \widehat{\sigma}_{21}^2, \widehat{\sigma}_{22}^2, \widehat{\rho}_2)^T$ *a* $p = n_1/(n_1+n_2)$; *parametre sú odhadnuté z dát.*
(2) $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \widehat{\rho})^T$; *parametre sú odhadnuté zo spoločného výberu.*
Vzorové riešenie pozri na obrázku 9 (dáta `two-samples-correlations-trunk.txt`*).* DÚ

Odlišnosti od teoretického rozdelenia. Odlišnosti empirického rozdelenia (rozdelenia realizácií) od teoretického (napr. normálneho) rozdelenia, môžeme charakterizovať napr. ako pravostranne alebo ľavostranne zošikmené rozdelenie (obrázok 10, prvý riadok vľavo a vpravo), ploché alebo špicaté rozdelenie (obrázok 10, prvý riadok uprostred). Pri viacrozmerných rozdeleniach je situácia komplikovanejšia. Pri dvojrozmernom normálnom rozdelení može byť napr. zošikmená jedna alebo obe premenné (príklad zošikmenia oboch premenných zľava pozri na obrázku 10, dolný riadok).

Príklad 23 (binomické rozdelenie, binomický experiment) *Experiment pozostávajúci z fixného počtu Bernouliho experimentov (ozn.* N*) sa nazýva binomický experiment. Pravdepodobnosť úspechu ozn.* p, *pravdepodobnosť neúspechu* $q = 1-p$. *Náhodná premenná* X *je počet pozorovaných úspechov počas experimentu. Pravdepodobnosť* $X = x$ *za podmienky, že* X *pochádza z binomického rozdelenia* $Bin(N, p)$ *píšeme ako* $\Pr(X = x) = \binom{N}{x} p^x (1-p)^{N-x}, x = 0, 1, \ldots, N$ *(Ugarte a kol. 2008). Stredná*



Obr. 9: Spoločná hustota dvojrozmerného normálneho rozdelenia (vľavo), hustota zmesi dvoch dvojrozmerných normálnych rozdelení (uprostred) a dvojrozmerný jadrový odhad superponovaný hustotou zmesi dvoch dvojrozmerných normálnych rozdelení (vpravo) – reálne dáta

hodnota $E[X] = Np$ a rozptyl $Var[X] = Np(1-p)$. Naprogramujte a zobrazte v R pravdepodobnostnú funkciu a (kumulatívnu) distribučnú funkciu pre Bin(5,0.5). Riešenie pozri na obrázku 11. cvič.

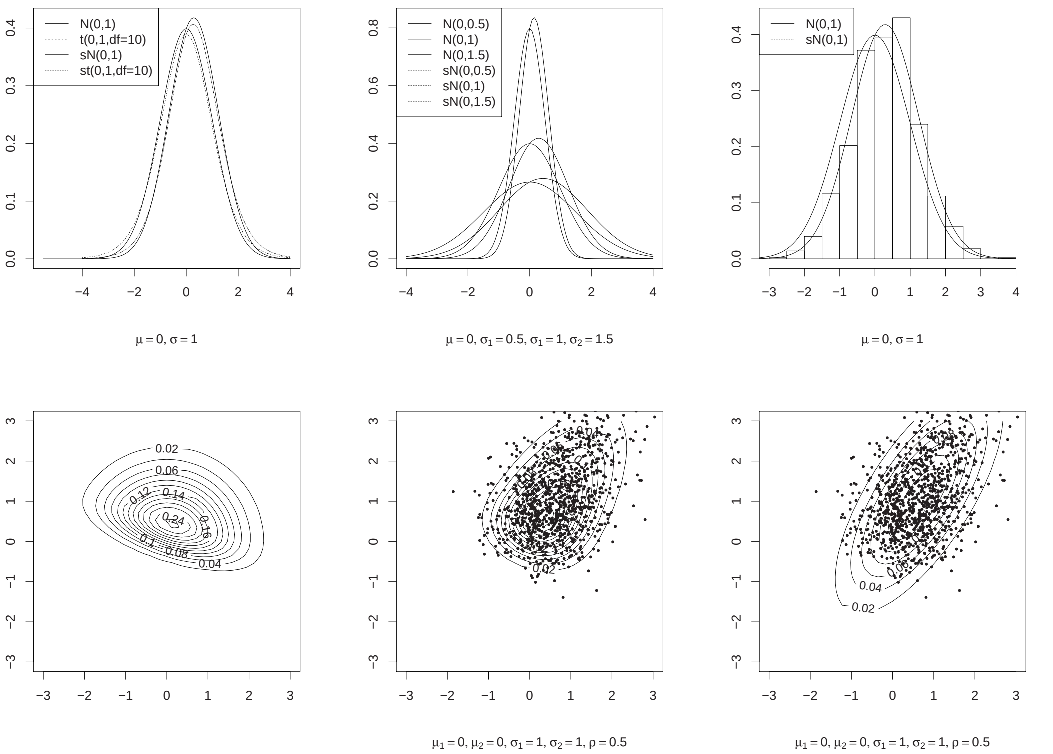
Príklad 24 (multinomické rozdelenie) *Majme náhodné premenné (1) socioekonomický status (vysoký – H, nízky – Lo), (2) politická prislušnosť (demokrat – D, republikán – R) a (3) politická filozofia (liberál – Li, konzervatívec – C). Označme ich interakcie nasledovne X_1 (H-D-Li), X_2 (H-D-C), X_3 (H-R-Li), X_4 (H-R-C), X_5 (Lo-D-Li), X_6 (Lo-D-C), X_7 (Lo-R-Li) a X_8 (Lo-R-C). Predpokladajme, že máme náhodný výber s rozsahom $N = 50$. Pravdepodobnosti p_j sú nasledovné*

	D-Li	*D-C*	*R-Li*	*R-C*	*spolu*
H	0.12	0.12	0.04	0.12	0.4
Lo	0.18	0.18	0.06	0.18	0.6
spolu	0.30	0.30	0.10	0.30	1.0

Vypočítajte $Var[X_1]$, $Var[X_3]$, $Cov\,[X_1, X_3]$, $Cor\,[X_1, X_3]$ a očakávané početnosti $Np_j, j = 1, 2, \ldots, 8$. pred

Riešenie
$X = (X_1, X_2, \ldots, X_8) \sim Mult(N, \mathbf{p})$, kde $N = 50$, $\mathbf{p} = (p_1, p_2, \ldots, p_8)^T$, vieme, že $X_j \sim Bin(N, p_j)$, p_j sú v tabuľke v zadaní príkladu a $j = 1, 2, \ldots, 8$. Potom
$Var[X_1] = 50 \times 0.12 \times (1 - 0.12) = 5.28$,
$Var[X_3] = 50 \times 0.04 \times (1 - 0.04) = 1.92$.
Vybraná kovariancia a korelácia (medzi počtami prislušných skupín) je rovná
$Cov\,[X_1, X_3] = -50 \times 0.12 \times 0.04 = -0.24$, $Cor\,[X_1, X_3] = -0.24/\sqrt{5.28 \times 1.92} = -0.075$.
Očakávané počtetnosti pre každú bunku tabuľky sú (všeobecne nemusia byť) celé čísla:

	D-Li	D-C	R-Li	R-C
H	6	6	2	6
Lo	9	9	3	9



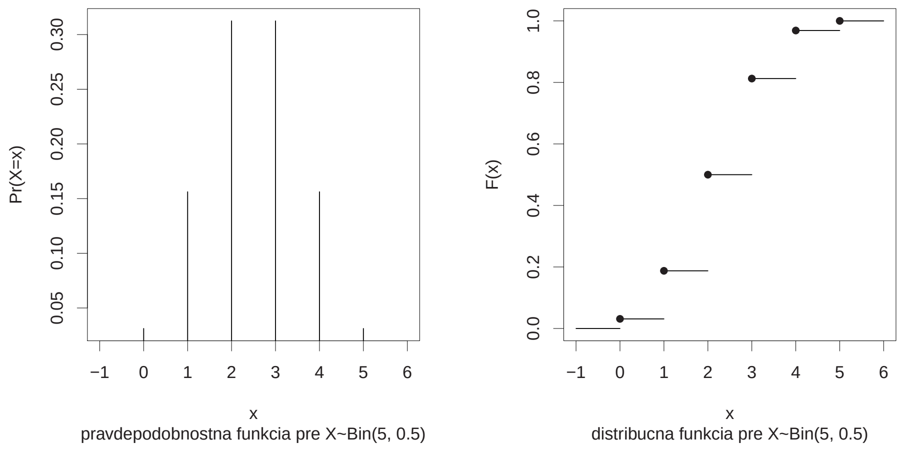
Obr. 10: Hustoty normálneho rozdelenia a zošikmeného normálneho rozdelenia pri rôznych parametroch (prvý riadok); hustoty dvojrozmerného zošikmeného normálneho rozdelenia (druhý riadok vľavo a uprostred) a dvojrozmerného normálneho rozdelenia (druhý riadok vpravo) pri rôznych parametroch

Príklad 25 (súčinové multinomické rozdelenie) *Majme dáta z predchádzajúceho príkladu a náhodný výber s rozsahom* $N_1 = 30$ *zo skupiny H, ďalší náhodný výber s rozsahom* $N_2 = 20$ *zo skupiny Lo. Označme interakcie premenných nasledovne* $X_{11} = X_{1|1}$ *(H-D-Li),* $X_{12} = X_{2|1}$ *(H-D-C),* $X_{13} = X_{3|1}$ *(H-R-Li),* $X_{14} = X_{4|1}$ *(H-R-C),* $X_{21} = X_{1|2}$ *(Lo-D-Li),* $X_{22} = X_{2|2}$ *(Lo-D-C),* $X_{23} = X_{3|2}$ *(Lo-R-Li) a* $X_{24} = X_{4|2}$ *(Lo-R-C), kde* $\mathbf{X}_1 = (X_{11}, X_{12}, X_{13}, X_{14})^T$ *a* $\mathbf{X}_2 = (X_{21}, X_{22}, X_{23}, X_{24})^T$*. Potom* $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ *má súčinové multinomické rozdelenie s* $K = 2, N_1 = 30, J_1 = 4, N_2 = 20, J_2 = 4$*. Zápis s* $X_{j|k}$*, kde* $j = 1, 2, 3, 4$ *a* $k = 1, 2$ *zvýrazňuje fakt, že rozdelenie je podmienené socioekonomickým statusom (vysoký – H, nízky – Lo), t.j. rozdelenie v stĺpcoch tabuľky je podmienené jej riadkom. Realizácie* $X_{j|k}$ *označujeme ako* $n_{j|k} = n_{kj}$*, pravdepodobnosti ekvivalentné* $X_{j|k} = X_{kj}$ *ako* $p_{j|k} = p_{kj}$*. Vypočítajte podmienené pravdepodobnosti* $p_{j|k}$*, očakávané početnosti* $N_k p_{kj}$*,* $Var[X_{13}]$*,* $Cov\,[X_{21}, X_{23}]$ *a* $Cor\,[X_{11}, X_{23}]$. pred

Riešenie

Pravdepodobnosti štyroch kategórií asociovaných s H statusom sú podmienené pravdepodobnosti dané H statusom. Napr. $\Pr(X_{3|1}) = 0.04/0.4 = 0.1$. $\Pr(X_{1|1}) = 0.12/0.4 = 0.3$, $\Pr(X_{3|2}) = 0.06/0.6 = 0.1$. Musíme ale tabuľku prepísať na súčinovo-multinomický model, teda podmienené pravdepodobnosti $p_{j|i}$ dané socioekonomickým statusom i budú

	D-Li	D-C	R-Li	R-C	spolu
H	0.3	0.3	0.1	0.3	1.0
Lo	0.3	0.3	0.1	0.3	1.0



Obr. 11: Pravdepodobnostná funkcia a distribučná funkcia $Bin(5, 0.5)$

Pre $N_1 = 30$ a $N_2 = 20$ máme očakávané počty nasledovné

	D-Li	D-C	R-Li	R-C	spolu
H	9	9	3	9	30
Lo	6	6	2	6	20

$Var(X_{3|1}) = 30 \times 0.1 \times (1 - 0.1) = 2.7$.
Vybrané kovariancie (medzi počtami prislušných skupín) sú rovné
$Cov\left[X_{1|2}, X_{3|2}\right] = -20 \times 0.3 \times 0.1 = -0.6$,
$Cov\left[X_{1|1}, X_{3|2}\right] = 0$, lebo $\mathbf{X}_1$ a $\mathbf{X}_2$ sú nezávislé.

Príklad 26 (Poissonovo rozdelenie; počet havárií za týždeň) *Ak každý z 50 miliónov ľudí šoféruje auto v Taliansku budúci týždeň nezávisle, potom pravdepodobnosť smrti pri autonehode bude* 0.000002*, kde počet úmrtí má binomické rozdelenie* $Bin(50mil, 0.000002)$ *alebo limitne Poissonovo rozdelenie s parametrom* $50mil \times 0.000002 = 100$.

Príklad 27 (Poissonovo rozdelenie; pruské armádne jednotky) *Nech početnosti úmrtí* X *ako následok kopnutia koňom v Pruských armádnych jednotkách (Bortkiewicz 1898), má Poissonovo rozdelenie s parametrom* λ*, t.j.* $X \sim Poiss(\lambda)$*. Pravdepodobnosť, že niekto bude smrteľne zranený v danom dni je extrémne malá. Majme* 10 *vojenských jednotiek za* 20*-ročnú periódu s rozsahom* $M = 200$ *(*$200 = 10 \times 20$*), kde popri početnostiach úmrtí* $n = 1, 2, 3, 4, 5+$*, v danej jednotke a v danom roku, zaznamenávame aj početnosti vojenských jednotiek* m_n *pri danom* n*, kde* $M = \sum m_n$ *(pozri tabuľku). Vypočítajte očakávané početnosti, za predpokladu* $X \sim Poiss(\lambda)$*, kde* $\lambda = \frac{\sum_n n m_n}{\sum_n m_n}$. cvič.

n	*0*	*1*	*2*	*3*	*4*	*5+*
m_n	*109*	*65*	*22*	*3*	*1*	*0*

Príklad 28 (podiel chlapcov a dievčat v rodinách) *Nech* X *predstavuje početnosť chlapcov medzi deťmi v rodinách. Tu môžeme predpokladať, že* $X \nsim Bin(N, p)$*, t.j. rodina môže mať vychýlený pomer pohlaví detí v smere ku chlapcom alebo dievčatám. V realite teda môžeme mať príliš veľa rodín len s chlapcami alebo len s dievčatami a nemáme dostatok rodín s pomerom pohlaví blízkym* $51 : 49$ *(pomer chlapcov ku dievčatám). Z toho nám vyplýva, že rozptyl početnosti chlapcov bude v skutočnosti väčší ako rozptyl predpokladaný binomickým modelom* $Bin(N, p)$. pred

Príklad 29 (overdispersion v binomickom modeli) *V klasickej štúdii pomeru pohlaví u ľudí z roku 1889 na základe záznamov z nemocníc v Sasku (bližšie pozri Lindsey a Altham (1998)) Geissler (1889) zaznamenal rozdelenie počtu chlapcov v rodinách. Medzi $M = 6115$ rodinami s $N = 12$ deťmi pozoroval nasledovné početnosti chlapcov (n sú početnosti chlapcov a m_n početnosti rodín s n chlapcami)*

n	0	1	2	3	4	5	6	7	8	9	10	11	12
m_n	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

Vypočítajte m_n za predpokladu, že početnosti chlapcov X v rodinách majú binomické rozdelenie s parametrami $\pi = \frac{\sum_{n=0}^{N} nm_n}{NM} = 0.5192$ a $N = 12$, ozn. $X \sim Bin(N, \pi)$. cvič.

Riešenie

n	0	1	2	3	4	5	6	7	8	9	10	11	12
očakávané m_n	1	12	72	258	628	1085	1367	1266	854	410	133	26	2

Keď porovnáme pozorované m_n a vypočítané (teoretické) m_n zistíme, že pozorované poukazujú na *overdispersion*, t.j. máme väčšie početnosti rodín s malým a veľkým množstvom chlapcov v porovnaní s teroretickými početnosťami.

Príklad 30 (overdispersion v Poissonovom modeli) *Majme početnosti úrazov n medzi robotníkmi v továrni, kde početnosti robotníkov m_n pri danom n pozri v tabuľke (Greenwood a Yule 1920).* cvič.

n	0	1	2	3	4	≥ 5
m_n	447	132	42	21	3	2

Vypočítajte očakávané početnosti robotníkov za predpokladu, že početnosti úrazov na robotníka X majú Poissonove rozdelenie s parametrom $\lambda = \frac{\sum_n nm_n}{\sum_n m_n} = 0.47$, ozn. $X \sim Poiss(\lambda)$.

Riešenie

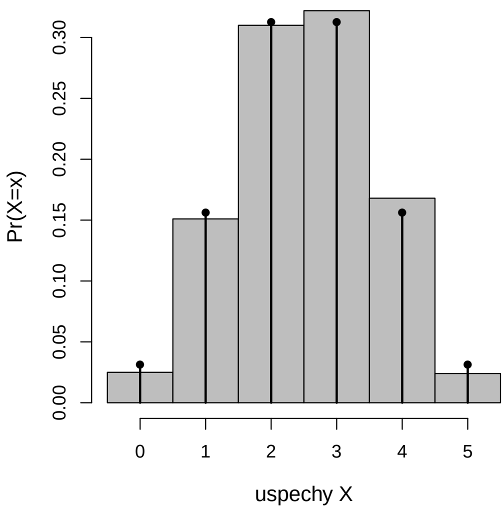
n	0	1	2	3	≥ 4
očakávané m_n	406	189	44	7	1

Keď porovnáme pozorované m_n a vypočítané (teoretické, očakávané) m_n zistíme, že pozorované poukazujú na *overdispersion*, t.j. máme viac robotníkov bez úrazu ako aj viac robotníkov s väčším množstvom úrazov v porovnaní s teroretickými početnosťami.

Príklad 31 (binomický rozdelenie, simulačná štúdia) *Vygenerujte pseudonáhodné čísla X (početnosti úspechov) opakovane M-krát ($M = 1000$) z $Bin(N, p)$, kde $N = 5$ a $p = 0.5$. Vytvorte tabuľku vygenerovaných (simulovaných) ako aj teoretických relatívnych početností (pre $n = 0, 1, \ldots, 5$). Superponujte histogram vygenerovaných pseudonáhodných čísel s teoretickou pravdepodobnostnou funkciou.* cvič.

Riešenie (pozri obrázok 12)

r	0	1	2	3	4	5
simulované relatívne početnosti	0.021	0.152	0.318	0.324	0.155	0.030
teoretické relatívne početnosti	0.031	0.156	0.312	0.312	0.156	0.031



Obr. 12: Histogram vygenerovaných pseudonáhodných čísel superponovaný spojnicovým grafom teoretickej pravdepodobnostnej funkcie X

Príklad 32 (binomické vs normálne rozdelenie) *Nech* $X_N \sim Bin(N, p)$, *potom môžeme aproximovať binomické rozdelenie normálnym nasledovne –* $X_N \sim N(Np, Np(1-p))$, *kde tiež platí* $Z_N = \frac{X_N - Np}{\sqrt{Np(1-p)}} \sim N(0, 1)$. *Ukážte, že CLV platí pre* $N = 100$ *a* $p = 1/2$ *na tri desatinné miesta.*

Riešenie
Príklad hovorí o tom, ako dobre normálne rozdelenie aproximuje binomické pri rozsahu $N = 100$, čo je dôležité pri testovaní hypotéz.
$E[X_N] = Np = 50$, $\sqrt{Var[X_N]} = \sqrt{Np(1-p)} = \sqrt{5}$.
Ak $Y_N = X_N/N$, potom $\Pr(|Y_N - 1/2| < \epsilon) = 0.236$, kde $\epsilon = 0.02$. $\Pr(0.48 < Y_{100} < 0.52) = \Pr(48 < X_{100} < 52) = \Pr(48.5 < X_{100} < 51.5) = \Pr(\frac{48.5-50}{\sqrt{5}} < Z_{100} < \frac{51.5-50}{\sqrt{5}})$, kde $X_{100} \sim N(50, 5)$ s použitím úpravy na spojitosť.

Príklad 33 (normálne rozdelenie, simulačná štúdia) *Na základe simulačnej štúdie preverte, že ak* $X \sim N(150, 6.25)$, *potom* $\overline{X}_n \sim N(150, 6.25/n)$. *Použite* $n = 30$. *Pre každú simuláciu* X *vypočítajte aritmetické priemery* $\overline{x}_m$, $m = 1, 2, \ldots, M$, *kde* $M = 500000$. *Superponujte ich histogram v relatívnej škále s teoretickou krivkou hustoty pre* $\overline{X}_n$. *Vypočítajte* $\Pr(\overline{X}_n > 151)$ *zo simulovaných dát a porovnajte tento výsledok s teoretickou (očakávanou) pravdepodobnosťou. Riešenie pozri na obrázku 13.* cvič.

Príklad 34 (normálne rozdelenie, simulačná štúdia) *Nech* $X \sim N(\mu_1, \sigma_1^2)$ *a* $Y \sim N(\mu_2, \sigma_2^2)$. *Potom* $\overline{X}_{n_1} - \overline{Y}_{n_2} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$. *Generujte pseudonáhodné čísla* X *a* Y *rozdelení* $N(\mu_j, \sigma_j^2), j = 1, 2$, *kde* $\mu_1 = 100, \sigma_1 = 10, \mu_2 = 50, \sigma_2 = 9$ *pri (a)* $n_1 = 4, n_2 = 5$, *(b)* $n_1 = 100, n_2 = 81$. *Pre každú simuláciu* X *a* Y *vypočítajte rozdiel* $\overline{x}_m - \overline{y}_m, m = 1, 2, \ldots, M$, *kde* $M = 1000$. *Superponujte histogram týchto rozdielov v relatívnej škále s teoretickou krivkou hustoty rozdielu* $\overline{X}_{n_1} - \overline{Y}_{n_2}$. *Pre prípad (a) aj (b) vypočítajte* $\Pr(\overline{X}_{n_1} - \overline{Y}_{n_2}) < 52$ *na základe empirického (vygenerovaného) a teoretického rozdelenia* $\overline{X}_{n_1} - \overline{Y}_{n_2}$. cvič.

Príklad 35 (štatistika) *Majme náhodný výber* $(X_1, X_2, \ldots, X_n)^T$, *kde* $X_i \in \mathbb{R}, i = 1, 2, \ldots, n$, *potom príkladmi štatistík sú:* $T_1 = \sum_{i=1}^n X_i \in \mathbb{R}$, $T_2 = \sum_{i=1}^n X_i^2 \in \mathbb{R}^+ \cup \{0\}$, $T_3 = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2) \in \mathbb{R}^2$. pred

Príklad 36 (testovacia štatistika, simulačná štúdia) *Na základe simulačnej štúdie preverte, že ak náhodná premenná X má asymptoticky binomické rozdelenie $Bin(N,p)$, potom testovacia štatistika $Z_W = \frac{X/N-p}{\sqrt{(p(1-p)/N}}$, má asymptoticky normálne rozdelenie $N(0,1)$. Použite $p = 0, 0.1, 0.5, 0.9$ a 1, a $N = 5, 10, 30, 50$ a 100. Okomentujte výsledky v spojitosti s Haldovou podmienkou $Np(1-p) > 9$. Pre každú simuláciu X vypočítajte $z_{W,m}$, $m = 1, 2, \ldots, M$, kde $M = 1000$. Superponujte histogram vygenerovaných testovacích štatistík v relatívnej škále s teoretickou krivkou hustoty Z_W.* cvič.

Príklad 36 hovorí o použití jednovýberovej testovacej štatistiky pre parameter binomického rozdelenia (pravdepodobnosť) pre rôzne pravdepodobnosti a rôzne početnosti. Ak Haldova podmienka nie je splnená, nie je možné testovaciu štatistiku použiť.

Príklad 37 (testovacia štatistika, simulačná štúdia) *Na základe simulačnej štúdie preverte, že ak (a) $X \sim N(\mu, \sigma^2)$, kde $\mu = 0, \sigma^2 = 1$ a (b) $X \sim Exp(\lambda)$, kde $\lambda = 1$, $E[X] = 1$ a $Var[X] = 1$, potom testovacia štatistika $F = \frac{(n-1)S^2}{\sigma^2}$ má asymptoticky χ^2_{n-1} rozdelenie s $n-1$ stupňami voľnosti. Použite rozsahy náhodných výberov $n = 15$ a $n = 100$. Pre každú simuláciu X vypočítajte $F_{poz,m}$, $m = 1, 2, \ldots, M$, kde $M = 1000$. Superponujte histogram vygenerovaných testovacích štatistík v relatívnej škále s teoretickou krivkou hustoty F.* cvič.

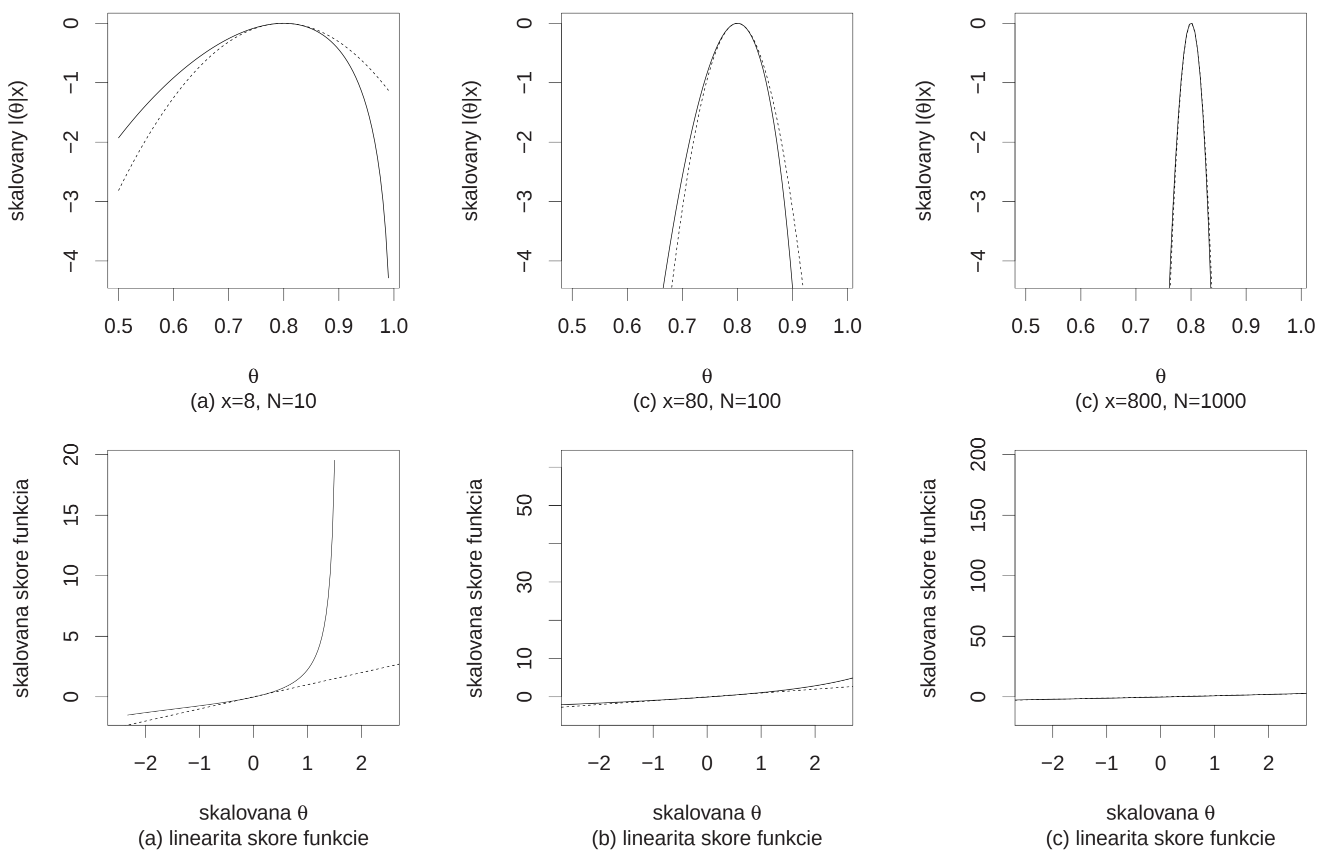
Príklad 38 (maximálne vierohodné odhady; Poissonovo rozdelenie) *Každý rok za posledných päť rokov boli v nejakom meste registrované 3, 2, 5, 0 a 4 zemetrasenia za rok. Za predpokladu, že počet zemetrasení za rok X má Poissonovo rozdelenie s parametrom λ, t.j. $X \sim Poiss(\lambda)$, odhadnite λ (λ predstavuje očakávanú početnosť zemetrasení za rok).* pred

Príklad 39 ($\mathcal{I}(\widehat{p})$ a rozptyl pre p; $X \sim Bin(N,p)$) *Z funkcie vierohodnosti odvoďte pozorovanú Fisherovu mieru informácie $\mathcal{I}(\widehat{p})$ a rozptyl $\widehat{Var[\widehat{p}]}$.* pred

Príklad 40 ($\mathcal{I}(\widehat{\lambda})$ a rozptyl pre λ; $X \sim Poiss(\lambda)$) *Každý rok za posledných päť rokov boli v nejakom meste registrované 3, 2, 5, 0 a 4 zemetrasenia za rok. Za predpokladu, že počet zemetrasení za rok $X \sim Poiss(\lambda)$, odhadnite rozptyl parametra λ a vypočítajte hodnotu tohoto odhadu rozptylu pre počet zemetrasení.* pred

Príklad 41 (kvadratická aproximácia funkcie vierohodnosti) *(1) Nakreslite škálovaný logaritmus funkcie vierohodnosti binomického rozdelenia. Na x-ovej osi bude p a na y-ovej osi $\ln \mathcal{L}(p) = l(p|\mathbf{x}) - \max(l(p|\mathbf{x}))$. Porovnajte $\ln \mathcal{L}(p)$ s kvadratickou aproximáciou vypočítanou pomocou Taylorovho rozvoja $\ln \mathcal{L}(p) = \ln(\frac{L(p|\mathbf{x})}{L(\widehat{p}|\mathbf{x})}) \approx -\frac{1}{2}\mathcal{I}(\widehat{p})(p-\widehat{p})^2$. (2) Nech skóre funkcia $S(p) = \frac{\partial}{\partial p} \ln L(p|\mathbf{x})$. Keď zoberieme deriváciu kvadratickej aproximácie uvedenej vyššie, dostaneme $S(p) \approx -\mathcal{I}(\widehat{p})(p-\widehat{p})$ alebo $-\mathcal{I}^{-1/2}(\widehat{p})S(p) \approx \mathcal{I}^{1/2}(\widehat{p})(p-\widehat{p})$. Potom zobrazením pravej strany na x-ovej osi a ľavej strany na y-ovej osi dostaneme asymptoticky lineárnu funkciu s jednotkovým sklonom. Asymptoticky tiež platí $\mathcal{I}^{1/2}(\widehat{p})(p-\widehat{p}) \sim N(0,1)$. Je postačujúce mať rozsah x-vej osi $\langle -2, 2\rangle$, pretože funkcia je asymptoticky (lokálne) lineárna na tomto intervale. Rozumne škálujte y-ovú os. Zobrazte pre (a) $n = 8, N = 10$, (b) $n = 80, N = 100$ a (c) $n = 800, N = 1000$ ($p \in (0.5, 0.99)$). Okomentujte rozdiely medzi (a), (b) a (c). Grafické riešenie je na obrázku 13.* cvič.

Príklad 42 ($\mathcal{I}(\widehat{\boldsymbol{\theta}})$ pre vektor $\boldsymbol{\theta} = (\mu, \sigma^2)^T$; $X \sim N(\mu, \sigma^2)$) *Nech $X \sim N(\mu, \sigma^2)$. Čomu je rovná pozorovaná Fisherova informačná matica $\mathcal{I}(\widehat{\boldsymbol{\theta}})$, kde $\boldsymbol{\theta} = (\widehat{\mu}, \widehat{\sigma}^2)^T$?* pred



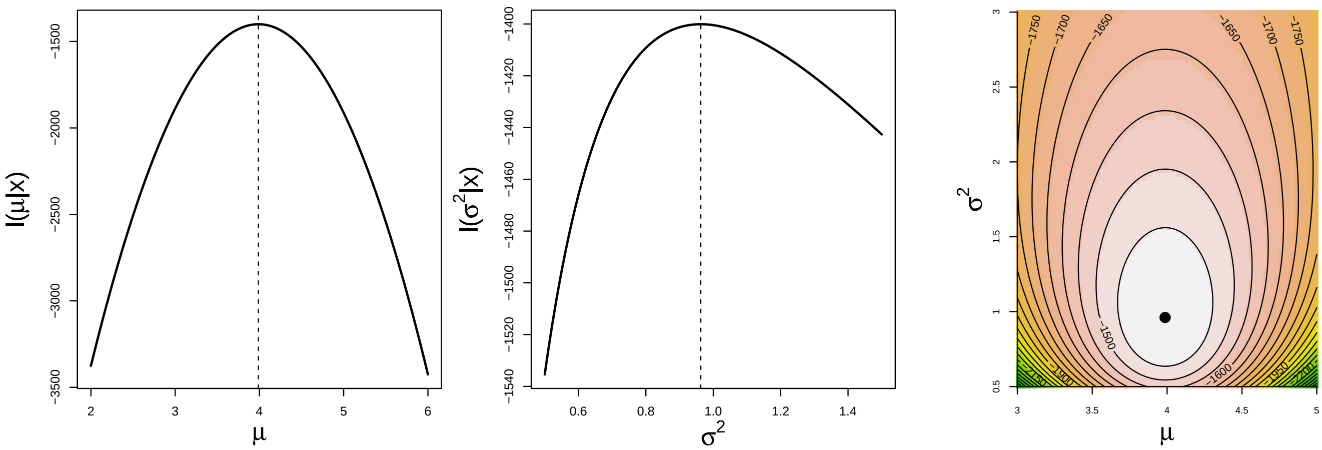
Obr. 13: Porovnanie škálovaného logaritmu funkcie vierohodnosti (plná čiara) s jeho kvadratickou aproximáciou (čiarkovaná čiara) v prvom riadku a porovnanie škálovanej skóre funkcie a priamky s nulovým interceptom a jednotkovým sklonom v druhom riadku

Príklad 43 ($\mathcal{I}(\widehat{\mathbf{p}})$ a rozptyl pre p; $\mathbf{X} \sim Mult_J(N, \mathbf{p})$) *Z funkcie vierohodnosti odvoďte pozorovanú Fisherovu informačnú maticu $\mathcal{I}(\widehat{\mathbf{p}})$ a kovariančnú maticu $\widehat{Var}[\widehat{\mathbf{p}}]$.* pred

Príklad 44 (profilová vierohodnosť; normálne rozdelenie) *Profilová funkcia vierohodnosti pre μ počítaná pre každe fixované μ, kde maximálne vierohodný odhad σ^2 bude $\widehat{\sigma}^2_\mu = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2$, má tvar $L(\mu|\mathbf{x}) = c\left(\widehat{\sigma}^2_\mu\right)^{-n/2}$, kde c je nejaká konštanta. $L(\mu|\mathbf{x})$ nie je identická s odhadnutou funkciou vierohodnosti $L(\mu, \sigma^2 = \widehat{\sigma}^2|\mathbf{x}) = c\exp\left(-\frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^n (x_i - \mu)^2\right)$, t.j. s rezom $L(\mu, \sigma^2|\mathbf{x})$ v bode $\sigma^2 = \widehat{\sigma}^2$. Obe funkcie vierohodnosti budú veľmi podobné, ak je rozptyl σ^2 dobre odhadnutý. V opačnom prípade sa preferuje profilová funkcia vierohodnosti. Profilová funkcia vierohodnosti pre σ^2 je rovná $L(\sigma^2|\mathbf{x}) = c\,(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \overline{x})^2\right) = c\,(\sigma^2)^{-n/2}\exp\left(-n\widehat{\sigma}^2/(2\sigma^2)\right)$.* pred

Príklad 45 (kvadratická aproximácia profilovej funkcie vierohodnosti) *(1) Nakreslite škálovaný logaritmus profilovej funkcie vierohodnosti normálneho rozdelenia pre μ. Na x-ovej osi bude μ a na y-ovej osi $\ln\mathcal{L}(\mu|\mathbf{x}) = l(\mu|\mathbf{x}) - \max(l(\mu|\mathbf{x}))$. Porovnajte $\ln\mathcal{L}(\mu|\mathbf{x})$ s kvadratickou aproximáciou vypočítanou pomocou Taylorovho rozvoja $\ln\mathcal{L}(\mu|\mathbf{x}) = \ln(\frac{L(\mu|\mathbf{x})}{L(\widehat{\mu}|\mathbf{x})}) \approx -\frac{1}{2}\mathcal{I}(\widehat{\mu})(\mu - \widehat{\mu})^2$. (2) Nech skóre funkcia $S(\mu) = \frac{\partial}{\partial\mu}\ln L(\mu|\mathbf{x})$. Keď zoberieme deriváciu kvadratickej aproximácie uvedenej vyššie, dostaneme $S(\mu) \approx -\mathcal{I}(\widehat{\mu})(\mu - \widehat{\mu})$ alebo $-\mathcal{I}^{-1/2}(\widehat{\mu})S(\mu) \approx \mathcal{I}^{1/2}(\widehat{\mu})(\mu - \widehat{\mu})$. Potom zobrazením pravej strany na x-ovej osi a ľavej strany na y-ovej osi dostaneme asymptoticky lineárnu funkciu s jednotkovým sklonom. Asymptoticky tiež platí $\mathcal{I}^{1/2}(\widehat{\mu})(\mu - \widehat{\mu}) \sim N(0, 1)$. Je postačujúce mať rozsah x-vej* *osi $\langle -2, 2\rangle$, pretože funkcia je asymptoticky (lokálne) lineárna na tomto intervale. Rozumne škálujte y-ovú os. Zobrazte pre (a) $n = 10$, (b) $n = 100$ a (c) $n = 1000$. Použite (1) $X \sim N(0,1)$ a (2) $X \sim (1-p)N(0,1) + pN(0,2)$, kde $p = 0.05$. Okomentujte rozdiely medzi (a), (b) a (c), ako aj rozdiely medzi (1) a (2).* DÚ

Príklad 46 (maximálne vierohodný odhad μ a σ^2) *Vygenerujte pseudonáhodné čísla z $X \sim N(4,1)$, $n = 1000$. (a) Napíšte logaritmus profilovej funkcie vierohodnosti pre μ a σ^2 a preverte, či sú maximálne vierohodné odhady μ a σ^2 dostatočne blízko k ich skutočným hodnotám. Nakreslite grafy $l(\mu|\mathbf{x})$ a $l(\sigma^2|\mathbf{x})$, kde zvýrazníte polohu maxím týchto funkcií. (b) Napíšte logaritmus funkcie vierohodnosti pre $\boldsymbol{\theta} = (\mu, \sigma^2)$ a preverte, či je maximálne vierohodný odhad $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ dostatočne blízko k jeho skutočnej hodnote. (c) Nakreslite graf $l((\mu, \sigma^2)|\mathbf{x})$ použitím funkcie* `image()` *a superponujte ho s kontúrovým grafom použitím funkcie* `contour()`. *Zvýraznite polohu maxima (pozri obrázok 14).* DÚ



Obr. 14: Profilová funkcia vierohodnosti pre μ (vľavo), σ^2 (uprostred) a funkcia vierohodnosti pre oba parametre (vpravo); $X \sim N(4,1)$); maximálne vierohodné odhady strednej hodnoty a rozptylu sú označené zvislou čiarkovanou čiarou (vľavo a uprotred) a maximálne vierohodný odhad vektora parametrov je označený • (vpravo)

Príklad 47 (binomické rozdelenie, maximálne vierohodný odhad p) *Nech $X \sim Bin(N,p)$ a realizácie X sú $x = n$. Predpokladajme, že sme pozorovali (a) $x = 2$, (b) $x = 10$ a (c) $x = 18$ úspechov v $N = 20$ pokusoch. Pomocou R vypočítajte maximálne vierohodný odhad p. Výsledok zobrazte do grafu spolu s funkciou vierohodnosti.* cvič.

Príklad 48 (maximálne vierohodné odhady; multinomické rozdelenie) *Majme dáta* `more-samples-probabilities-pubis.txt`. *Nakreslite logaritmus štandardizovanej funkcie vierohodnosti v parametroch p_1 a p_2 Európskej populácie ($n_1 = 30$, $n_2 = 20$ a $n_3 = 10$) pomocou funkcie* `contour()`. *Dokreslite do obrázku jej maximum v bode $\widehat{\boldsymbol{\theta}} = (\widehat{p}_1, \widehat{p}_2)^T$.* DÚ

Príklad 49 (overdispersion v Poissonovom modeli, pokrač.) *Majme početnosti úrazov n medzi robotníkmi v továrni, kde početnosti robotníkov m_n pri danom n pozri v tabuľke (Greenwood a Yule 1920).*

n	0	1	2	3	4	≥ 5
m_n	447	132	42	21	3	2

Vypočítajte m_n za predpokladu, že početnosti úrazov na robotníka X majú negatívne binomické rozdelenie s parametrami α a π. cvič.

2 Charakteristiky polohy a variability a štatistická grafika

Príklad 50 (argument minima) *Vygenerujte pseudonáhodné čísla* $X \sim N(\mu, \sigma^2)$, $n = 1000$, $\mu = 0, \sigma^2 = 1$. *Vygenerované čísla ozn.* $x_i, i = 1, 2, \ldots, 1000$. *Nájdite numericky také* c, *ktoré minimalizuje (a) sumu štvorcov odchýlok* $\sum_{i=1}^{1000}(x_i - c)^2$, *t.j.* $c_1 = \arg\min_{\forall c} \sum_{i=1}^{1000}(x_i - c)^2$ *a (b) sumu absolútnych odchýlok* $\sum_{i=1}^{100}|x_i - c|$, *t.j.* $c_2 = \arg\min_{\forall c} \sum_{i=1}^{1000}|x_i - c|$. *Za* c *dosadzujte postupne (1) všetky* $x_{(j)}$ ($x_{(j)}$ *sú usporiadané* x_i *podľa veľkosti od najmenšieho po najväčšie) a vybrané charakteristiky polohy ako (2) aritmetický priemer, (3) nejaké kvantily* $\widetilde{x}_p$, *kde* $p \in \langle 0, 1\rangle$ *a pod. Nakreslite obrázok závislosti (a) sumy štvorcov odchýlok na* $x_{(j)}$, *t.j. body* $[x_j, y_j]$, *kde* $y_j = \sum_{i=1}^{1000}(x_i - x_{(j)})^2$ *a (b) sumy absolútnych odchýlok na* $x_{(j)}$, *t.j. body* $[x_{(j)}, y_j]$, *kde* $y_j = \sum_{i=1}^{1000}|x_i - x_{(j)}|$. *Podobné obrázky nakreslite aj pre* $\widetilde{x}_p$ *namiesto* $x_{(j)}$. DÚ

Definícia 1 (asymptotické rozdelenie poriadkovej štatistiky) *Nech* $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ *sú poriadkové štatistiky náhodného výberu* $X_1, X_2, \ldots, X_n$. *Majme pravdepodobnosť* α, *kde* $F(t_\alpha) = \alpha$. *Asymptoticky platí, že* $\sqrt{n}(\frac{j}{n} - \alpha)$ *konverguje k* 0. *Potom je poriadková štatistika* $X_{(j)}$ *normálne rozdelená so strednou hodnotou* $E[X_{(j)}] = t_\alpha$ *a rozptylom* $\sigma^2_{X_{(j)}} = \frac{\alpha(1-\alpha)}{f^2(t_\alpha)n}$. *Ak* $X \sim N(\mu, \sigma^2)$, *potom* $\sigma^2_{X_{(j)}} = \sigma^2 \frac{\pi^2}{24 \ln n}$.

Príklad 51 (rozptyl poriadkovej štatistiky) *Pomocou delta metódy odvoďte rozptyl poriadkovej štatistiky v definícii 1.* DÚ

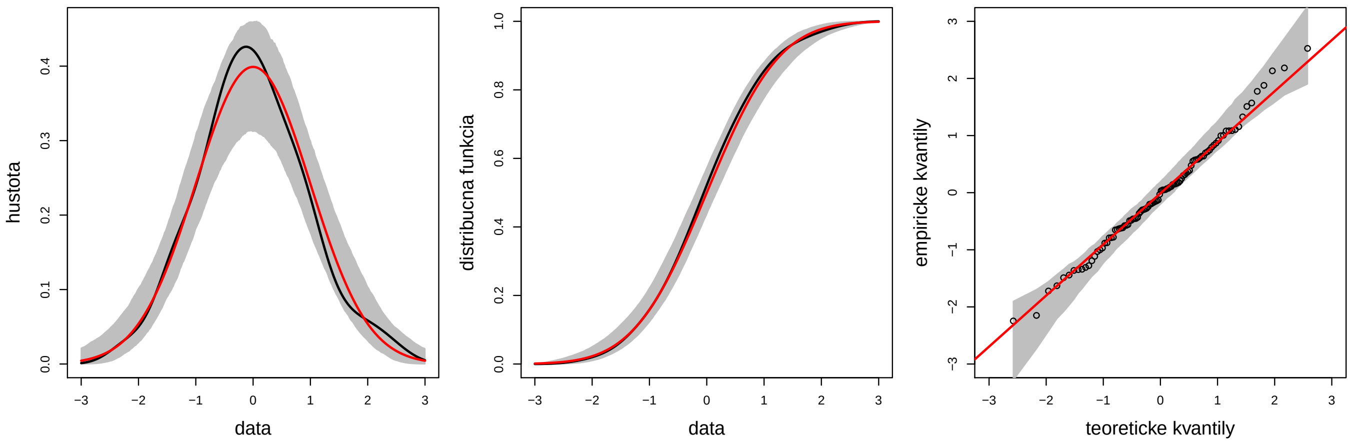
Príklad 52 (rozptyl poriadkovej štatistiky, $X \sim N(\mu, \sigma^2)$**)** *Pomocou definície 1 odvoďte rozptyl poriadkovej štatistiky, ak* $X \sim N(\mu, \sigma^2)$. DÚ

Definícia 2 (stredná hodnota a rozptyl mediánu) *Stredná hodnota mediánu* $X_{(\frac{n+1}{2})}$ *je rovná* $E[X_{(\frac{n+1}{2})}] = \widetilde{\mu}$ *a rozptyl mediánu* $\sigma^2_{X_{(\frac{n+1}{2})}} = \frac{1}{4f^2(\widetilde{\mu})n}$, *kde* n *je nepárne. Ak* $X \sim N(\mu, \sigma^2)$, *potom* $\sigma^2_{X_{(\frac{n+1}{2})}} = \sigma^2 \frac{\pi}{2n}$. pred

Príklad 53 (rozptyl mediánu) *Pomocou delta metódy odvoďte rozptyl poriadkovej štatistiky v definícii 2.* DÚ

Príklad 54 (rozptyl mediánu, $X \sim N(\mu, \sigma^2)$**)** *Pomocou definície 2 odvoďte rozptyl poriadkovej štatistiky, ak* $X \sim N(\mu, \sigma^2)$. DÚ

Príklad 55 (pásy normality) *Na základe vygenerovaných pseudonáhodných čísel* $X \sim N(\mu, \sigma^2)$, $n = 100, \mu = 0, \sigma^2 = 1$, *kde* $M = 1000$, *odhadnite (a) hustotu* m*-tej realizácie pomocou funkcie* `density()`; *ponechajte argument* `n=512` *a nastavte* `from=-3` *a* `to=3`; *(b) distribučnú funkciu* m*-tej realizácie pomocou (a) a funkcie* `cumsum()` *a (c) empirické kvantily* m*-tej realizácie pomocou funkcie* `qqnorm()`. *Vygenerované čísla* $x_{mi}, m = 1, 2, \ldots, 1000$ *a* $i = 1, 2, \ldots, 100$, *uložte po riadkoch do matice* $\mathbf{X}$, *ktorá bude mať rozmery* 1000×100. *Odhadnuté hustoty a distribučné funkcie uložte po riakodch do matíc* $\mathbf{H}$ *a* $\mathbf{D}$, *ktoré bude mať rozmery* 1000×512 *a empirické kvantily do matice* $\mathbf{K}$, *ktorá bude mať rozmery* 1000×100. *Pre každú z matíc* $\mathbf{H}$, $\mathbf{D}$ *a* $\mathbf{K}$ *vypočítajte* $\widetilde{x}_{0.05}$ *a* $\widetilde{x}_{0.95}$ *po stĺpcoch a zobrazte ich ako pásy pomocou funkcie* `polygon()`. *Do obrázkov vkreslite (a) teoretickú hustotu, (b) teoretickú distribučnú funkciu a (c) kvantilovú priamku (pomocou funkcie* `qqline()`*) červenou farbou. Obrázky usporiadajte ako trojicu vedľa seba. Dáta, ktorých normalitu chceme graficky testovať budú (1)* $X \sim N(0, 1)$, $n = 100$, *(2)* $X \sim [pN(0, 1) + (1 - p)N(0, 4)]$, $n = 100$ *a* $p = 0.95$ *a (3)* $X \sim [pN(0, 1) + (1 - p)N(0, 4)]$, $n = 100$ *a* $p = 0.9$. *Zobrazte (1), (2) a (3) oddelene do grafov (a), (b) a (c). Okomentujte. Riešenie (1) pozri na obrázku 15.*



Obr. 15: Pásy spoľahlivosti normálneho rozdelenia – pre hustotu (vľavo), distribučnú funkciu (uprostred) a kvantilovú priamku (vpravo)

3 Testovanie hypotéz

Aby sme mohli z dát vyvodiť nejaké interpretovateľné závery, musíme najprv ciele (účely, biologicky formulované hypotézy) výskumu preformulovať do matematicko-štatistickej podoby. Takáto formulácia je dôležitá pri výbere správneho modelu na dáta. V prípade parametrického modelu máme jeho parametre θ, kde hypotézy sú o týchto parametroch. Jednou hypotézou je tzv. nulová hypotéza, čo je tvrdenie o θ, ktoré testujeme prostredníctvom nejakého **štatistického testu**. Ďalšou je alternatívna hypotéza, ktorá je doplnkom nulovej hypotézy v priestore, z ktorého pochádza θ. Na základe výsledku testu vyvodíme záver o θ. Proces testovania hypotéz nazývame aj **štatistická inferencia**, ktorej základom je **štatistická teória testovania hypotéz**.

Definícia 3 (štatistický test) *Ak parametrický priestor Θ, do ktorého patrí parameter θ modelu $\mathcal{F}$ rozdelíme na dve podmnožiny Θ_0 a Θ_1, kde $\Theta = \Theta_0 \cup \Theta_1$, potom* ***štatistický test*** *$\mathcal{T} = \mathcal{T}(X)$ predstavuje funkciu X z výberového priestoru $\mathcal{Y}$ do priestoru $\Theta_0 \cup \Theta_1$. Štatistický test je teda pravidlo výberu jednej z dvoch možností, nulovej a alternatívnej, na základe dát.* ***Nulovú hypotézu*** *definujeme ako $H_0 : \theta \in \Theta_0$ a* ***alternatívnu hypotézu*** *ako $H_1 : \theta \in \Theta_1$. Prvky množiny $\mathcal{Y}$, pre ktoré zamietame H_0, predstavujú tzv.* ***oblasť (obor) nezamietania nulovej hypotézy****. Tieto prvky označujeme $\mathcal{Y}_0$. V tomto prípade H_0 nezamietame. Prvky množiny $\mathcal{Y}$, pre ktoré nezamietame H_0, predstavujú tzv.* ***oblasť (obor) zamietania nulovej hypotézy*** *alebo* ***kritickú oblasť****. Tieto prvky označujeme $\mathcal{Y}_1 = \mathcal{W}_X$. V tomto prípade H_0 zamietame. Rozdelenie priestoru Θ na obor zamietania $\mathcal{W}_T = \mathcal{W}$ a nezamietania $\mathcal{A}_T$ prebieha na základe* ***testovacej štatistiky*** *$T(X)$.*

Stretávame sa s nasledovnými štyrmi možnosťami:

A) ak platí H_0, tak naše rozhodutie je nezamietnuť H_0 (správne);

B) ak platí H_0, tak naše rozhodutie je zamietnuť H_0 (nesprávne);

C) ak neplatí H_0, tak naše rozhodutie je nezamietnuť H_0 (nesprávne);

D) ak neplatí H_0, tak naše rozhodutie je zamietnuť H_0 (správne).

V prípade B, C je naše rozhodnutie nesprávne, v prípade A, D správne. V prípade B sa dopúšťame **chyby prvého druhu** (CHPD), kde $\Pr(\text{CHPD}) \leq \alpha$ a α sa nazýva **hladina významnosti**. Doplnok ku α je pravdepodobnosť $1-\alpha$ a nazýva sa **koeficient spoľahlivosti** (prípad A). V prípade C sa dopúšťame **chyby druhého druhu** (CHDD), kde $\Pr(\text{CHDD}) = \beta$. Doplnok ku Pr(CHDD) je pravdepodobnosť $1-\beta$ a nazýva sa **sila testu** (pri nejakej alternatíve; prípad D). Sila testu závisí na zvolenej testovacej metóde a hlavne na tom, aké je skutočné rozdelenie náhodnej premennej X, aký je typ použitej testovacej štatistiky alebo aké sú skutočné hodnoty parametrov θ.

Teda

rozhodnutie/skutočnosť	H_0 platí	H_0 neplatí
H_0 nezamietnuť	správne rozhodnutie	chyba II. druhu
H_0 zamietnuť	chyba I. druhu	správne rozhodnutie

Vyššie uvedené pravdepodobnosti môžeme zosumarizovať nasledovne:

A) $1-\alpha \le \Pr(\text{nezamietni } H_0 | H_0 \text{ platí}) = \Pr(\text{nezamietni } H_0 | H_1 \text{ neplatí})$;

B) $\alpha \ge \Pr(\text{CHPD}) = \Pr(\text{zamietni } H_0 | H_0 \text{ platí}) = \Pr(\text{zamietni } H_0 | H_1 \text{ neplatí})$;

C) $\beta = \Pr(\text{CHDD}) = \Pr(\text{nezamietni } H_0 | H_0 \text{ neplatí}) = \Pr(\text{nezamietni } H_0 | H_1 \text{ platí})$;

D) $1-\beta = \Pr(\text{zamietni } H_0 | H_0 \text{ neplatí}) = \Pr(\text{zamietni } H_0 | H_1 \text{ platí})$.

V prípade (A) a (B) hovoríme o teste na hladine významnosti α (α-level test). Ak v (A) a (B) nahradíme znamienko nerovnosti rovnosťou, hovoríme o teste úrovne α (size α test). Test $\mathcal{T}(\mathbf{x})$ charakterizujeme jeho **silofunkciou** $\beta^*(\theta) = 1-\beta(\theta) = \Pr(\theta \in \Theta : \mathcal{T}(\mathbf{x}) = \Theta_1)$. Táto je závislá na θ a niekoľkých iných argumentoch (vždy sú špecifikované pri konkrétnych testoch) ako aj na type alternatívnej hypotézy. **Hladinu významnosti** α je možné definovať tiež pomocou $\beta^*(\theta)$, kde $\alpha = \sup_{\theta \in \Theta_0} \beta^*(\theta)$, t.j. ide o najväčšiu možnú pravdepodobnosť chyby I. druhu.

Hladina významnosti α je daná (určená štatistikom, experimentátorom) vopred, testovacia štatistika $T(\mathbf{x})$ sa vypočíta na základe realizácií (odpovedí) x náhodnej premennej X, kritická hodnota (alebo kvantil) sa nájde v štatistických tabuľkách alebo ju vypočítame pomocou R.

V praxi teda voľba $\mathcal{W}$ závisí na požiadavke, aby pravdepodobnosť chyby I. druhu bola menšia alebo rovná zvolenému kladnému číslu α, kde $\alpha \in (0, 1/2)$, najčastejšie $\alpha = 0.05$, 0.01 alebo 0.001. Súčasne volíme $\mathcal{W}$ tak, aby pravdepodobnosť chyby I. druhu bola čo najmenšia. Optimálne by sme chceli mať $\beta^*(\theta)$ (pri nejakej hodnote θ) tak veľké číslo (blížiace sa jednotke) ako je možné, keď $\theta \in \Theta_1$ a také malé číslo, keď $\theta \in \Theta_0$. Tieto dve požiadavky sú v konflikte, t.j.

- zväčšovanie oboru nezamietania H_0 zmenšuje pravdepodobnosť chyby I. druhu, ale zväčšuje pravdepodobnosť chyby II. druhu;

- zmenšovanie oboru nezamietania H_0 zväčšuje pravdepodobnosť chyby I. druhu, ale zmenšuje pravdepodobnosť chyby II. druhu.

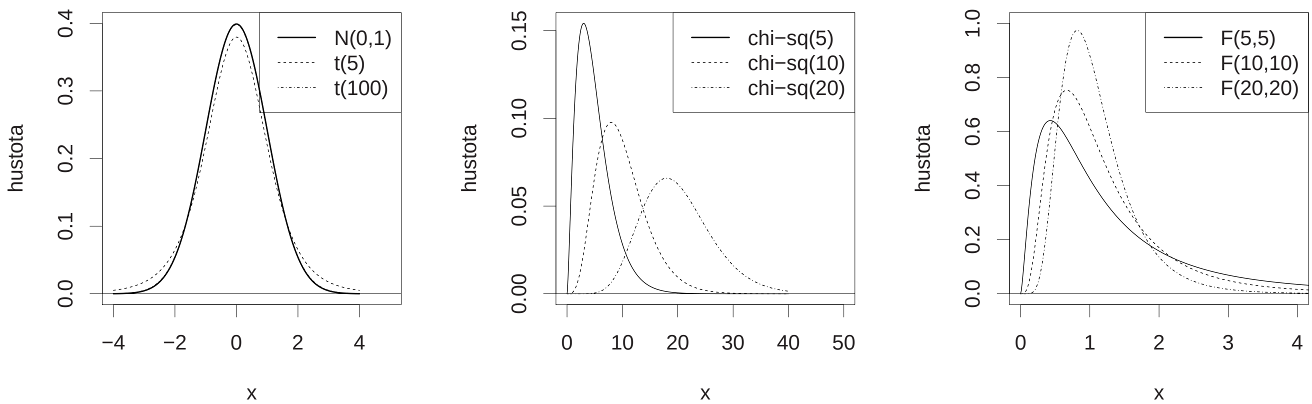
Extrémnymi prípadmi sú dve situácie, kde $\mathcal{T}_0(\mathbf{x}) = \Theta_0$ a $\mathcal{T}_1(\mathbf{x}) = \Theta_1$, t.j. každý z týchto dvoch testov je najlepší možný, keď θ patrí určitej podmnožine Θ, ale najhorší možný v opačnom prípade. Klasickým prístupom k tejto dileme je **Neyman-Pearsonov prístup**, kde dopredu fixujeme α a vyberieme taký test, ktorý má pri danej alternatíve najväčšiu silu $1-\beta$ pre $\theta \in \Theta_1$ medzi všetkými možnými na teoreticky stanovenej (danej, nominálnej) hladine významnosti α. Testy spomínané v kapitolách 5 až 8 sú najsilnejšími testami pri daných alternatívach za určitých predpokladov (bližšie pozri kap. 5 až 8). V reálnych situáciách (alebo často aj za porušenie predpokladov testov) sa nominálna hladina významnosti líši od aktuálnej (skutočnej) hladiny významnosti.

Definícia 4 (kvantil) *Nech F_X je distribučná funkcia náhodnej premennej X a $\alpha \in (0,1)$, kde $F_X(x_\alpha) = \alpha$. Potom číslo $x_\alpha = F_X^{-1}(\alpha)$ sa nazýva α-**kvantil** príslušného rozdelenia. Pre vybrané rozdelenia platí*

- $\Pr(X < x(1-\alpha)) = 1-\alpha$ *pre* $X \sim N(0,1)$ *(**normálne rozdelenie** s parametrami $\mu = 0$ a $\sigma^2 = 1$; hovorí sa mu aj **normálne normované rozdelenie** alebo **štandardizované normálne rozdelenie**); $x(1-\alpha)$ sa často zapisuje ako $x_{1-\alpha}$;*
- $\Pr\left(X < \chi^2_{df}(1-\alpha)\right) = 1-\alpha$ *pre* $X \sim \chi^2_{df}$ *(**chi-kvadrát rozdelenie** s df stupňami voľnosti);*
- $\Pr(X < t_{df}(1-\alpha)) = 1-\alpha$ *pre* $X \sim t_{df}$ *(**Studentovo t-rozdelenie** s df stupňami voľnosti);*
- $\Pr(X < F_{df_1,df_2}(1-\alpha)) = 1-\alpha$ *pre* $X \sim F_{df_1,df_2}$ *(**Fisherovo F-rozdelenie** s df_1 a df_2 stupňami voľnosti).*

Hustoty vyššie spomenutých rozdelení pozri na obrázku 16. Tiež platí

- $\Pr\left(x_{\alpha/2} < X < x_{1-\alpha/2}\right) = F_X\left(x_{1-\alpha/2}\right) - F_X\left(x_{\alpha/2}\right) = 1-\alpha$, kde $\alpha \in (0,1/2)$;
- $\Pr(X < Q_1) = 1/4$, kde $\mathrm{Q}_1 = F_X^{-1}(1/4)$ sa nazýva **výberový prvý kvartil** (**dolný kvartil**);
- $\Pr(X < Q_2) = 1/2$, kde $\mathrm{Q}_2 = F_X^{-1}(1/2)$ sa nazýva **výberový druhý kvartil** (**medián**);
- $\Pr(X < Q_3) = 3/4$, kde $\mathrm{Q}_3 = F_X^{-1}(3/4)$ sa nazýva **výberový tretí kvartil** (**horný kvartil**);
- $Q_d = F_X^{-1}(k/10)$ sa nazýva **výberový k-ty decil**, $Q_p = F_X^{-1}(k/100)$ sa nazýva **výberový k-ty percentil**.



Obr. 16: Grafické znázornenie hustôt normálneho rozdelenia, t-rozdelenia, χ^2-rozdelenia a F-rozdelenia pri rôznych stupňoch voľnosti

Definícia 5 (kritická hodnota) ***Kritická hodnota*** *príslušného rozdelenia je hodnota, ktorú náhodná premenná X prekročí s pravdepodobnosťou α. Pre vybrané rozdelenia platí*

- $\Pr(X > u(\alpha)) = \alpha$ *pre* $X \sim N(0,1)$; *$u(\alpha)$ sa často zapisuje ako u_α;*

- $\Pr\left(X > \chi^2_{df}(\alpha)\right) = \alpha$ *pre* $X \sim \chi^2_{df}$;
- $\Pr\left(X > t_{df}(\alpha)\right) = \alpha$ *pre* $X \sim t_{df}$;
- $\Pr\left(X > F_{df_1,df_2}(\alpha)\right) = \alpha$ *pre* $X \sim F_{df_1,df_2}$.

Z vyššie uvedeného je zreteľné, že $\alpha \times 100\%$ kritická hodnota je identická s $(1-\alpha)\times 100\%$ kvantilom. Pri kritickej hodnote počítame obsah pod krivkou hustoty príslušného rozdelenia nad týmto bodom a pri kvantile pod týmto bodom.

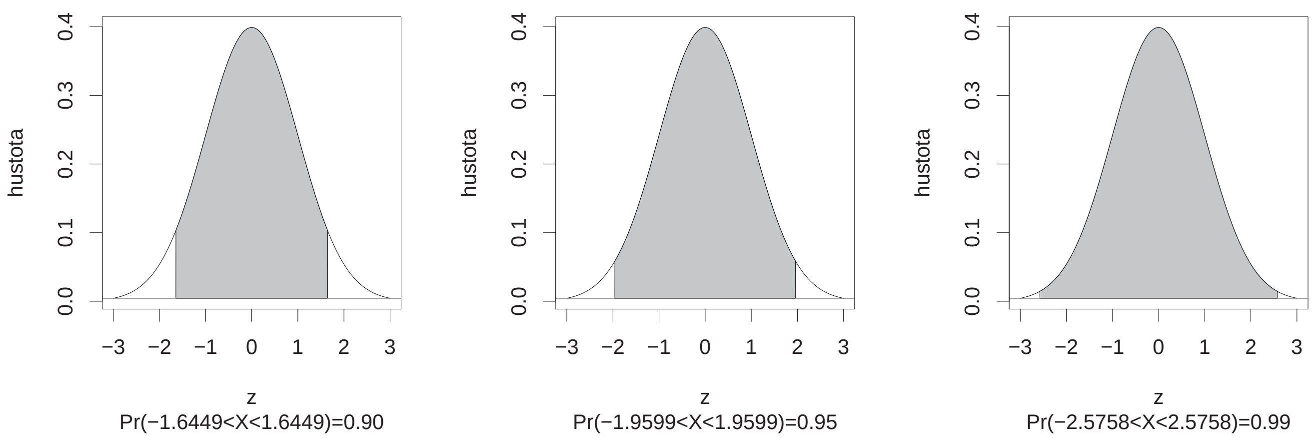
Stupne voľnosti *df* predstavujú množstvo nezávislej informácie, ktoré potrebujeme na charakterizáciu rozdelenia pravdepodobnosti nejakej náhodnej premennej. Používajú sa aj namiesto n ako prevencia pred nadhodnotením odhadu nejakého parametra (napr. rozptylu). Ide najčastejšie o rozsah náhodného výberu n zmenšený o jedna (napr. Studentovo t-rozdelenie, F-rozdelenie), kde jednotka odpočítaná od n znamená jeden voľne viazaný parameter (napr. jedna stredná hodnota, jeden rozptyl a pod.). Pri zložitejších modeloch s viacerými parametrami odpočívame od n počet voľne viazaných parametrov (napr. počet stredných hodnôt) alebo *df* môže tiež predstavovať počet hladín kategoriálnej premennej zmenšený o počet voľne viazaných parametrov.

Príklad 56 (štandardizované normálne rozdelenie) *Vypočítajte kritické hodnoty* $u(\alpha)$ *rozdelenia* $N(0,1)$, *kde* $\alpha = 0.1$, 0.05, 0.01, 0.025 *a* 0.005.

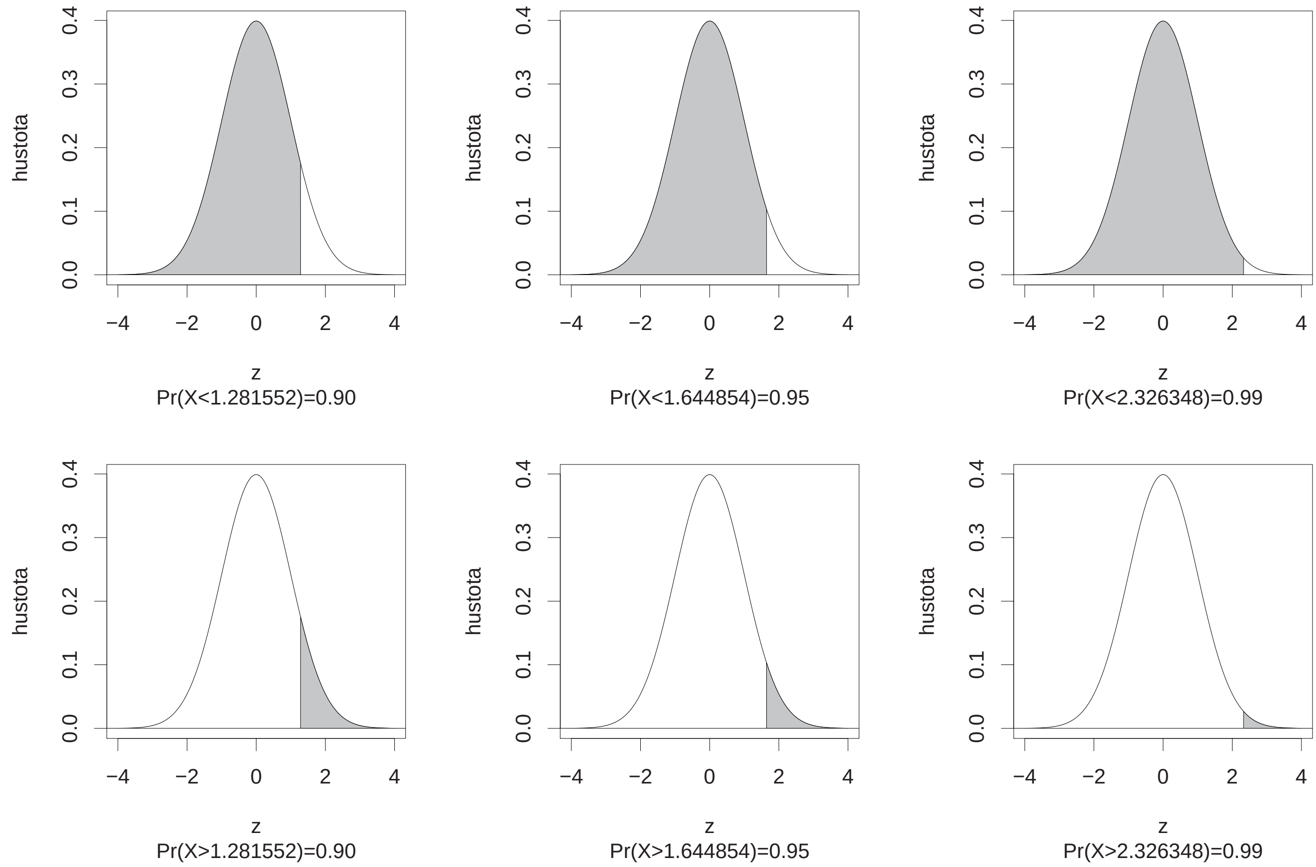
Riešenie v R (pozri obrázok 17 a 18)

```
qnorm(1-0.1) # 1.281552
qnorm(1-0.05) # 1.644854
qnorm(1-0.01) # 2.326348
qnorm(1-0.025) # 1.959964
qnorm(1-0.005) # 2.575829
```

Na výpočet pravdepodobnosti pod kvantilom sa používa funkcia `pnorm(Q)`. Na výpočet pravdepodobnosti nad kritickou hodnotou sa používa funkcia `1-pnorm(Q)`. Keďže štandardizované normálne rozdelenie je symetrické okolo nuly, $u(\alpha) = u(1-\alpha)$.



Obr. 17: Grafické znázornenie významu pravdepodobnosti pod krivkou rozdelenia medzi dvoma kvantilmi (normálne rozdelenie)



Obr. 18: Grafické znázornenie významu pravdepodobnosti (obsahu) pod krivkou normálneho rozdelenia; obsah pod (prvý riadok) a nad (druhý riadok) príslušným kvantilom

Príklad 57 (Studentove t-rozdelenie) *Vypočítajte kritické hodnoty Studentovho t-rozdelenia so stupňami voľnosti $df = 10$, t.j. $t_{df}(\alpha)$, kde $\alpha = 0.1$, 0.05, 0.01, 0.025 a 0.005.*

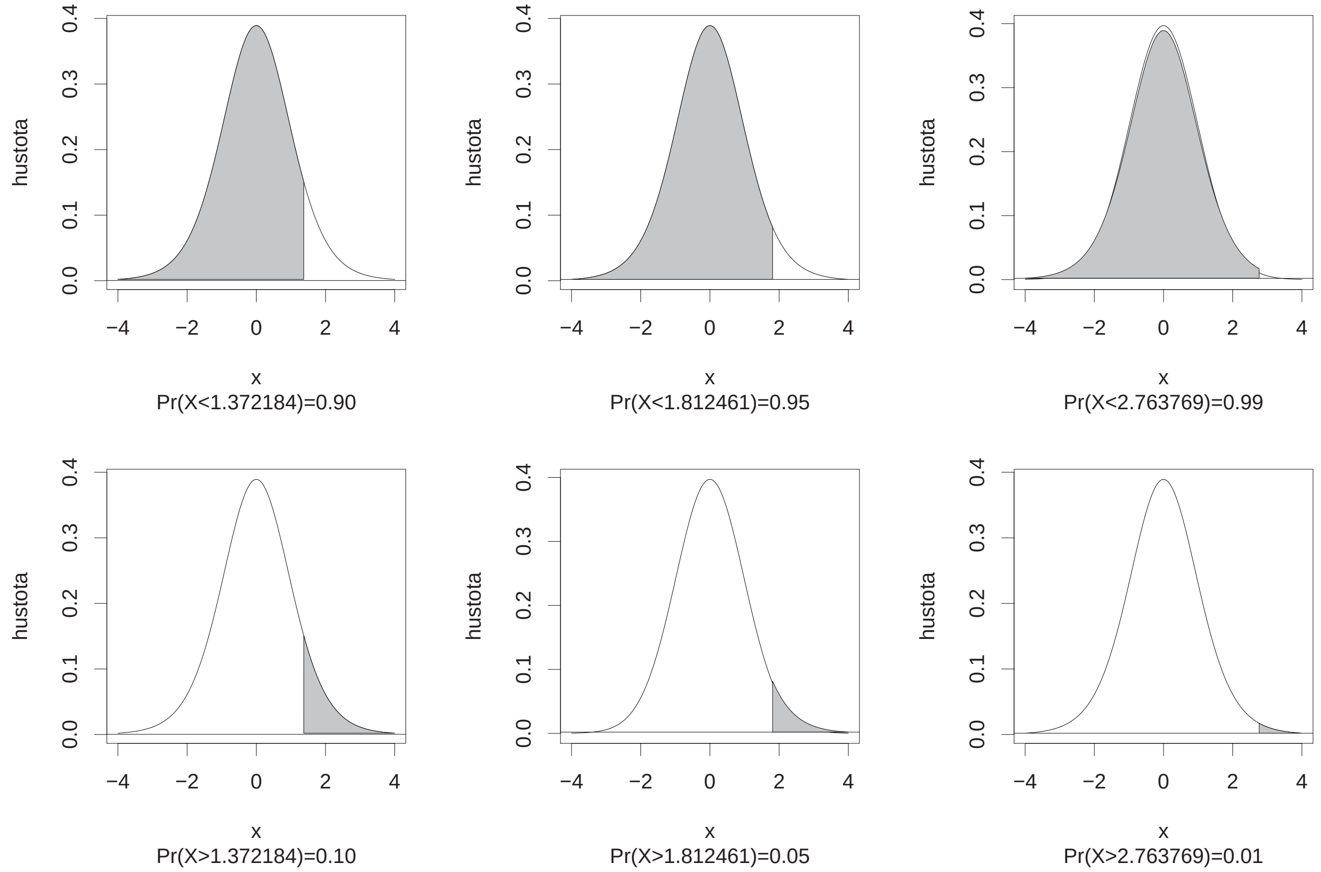
Riešenie v (pozri obrázok 19)

```
qt(1-0.1,10) # 1.372184
qt(1-0.05,10) # 1.812461
qt(1-0.01,10) # 2.763769
qt(1-0.025,10) # 2.228139
qt(1-0.005,10) # 3.169273
```

Na výpočet pravdepodobnosti pod kvantilom sa používa funkcia `pt(Q,df)`. Na výpočet pravdepodobnosti nad kritickou hodnotou sa používa funkcia `1-pt(Q,df)`. Keďže Studentovo t-rozdelenie je symetrické okolo nuly, $t_{df}(\alpha) = t_{df}(1-\alpha)$. Nejaký kvantil štandardizovaného normálneho rozdelenia je približne rovný kvantilu t-rozdelenia až pre veľmi veľké stupne voľnosti (resp. pravdepodobnosti nad kritickými hodnotami sú približne rovnaké). Napr. `1-pnorm(1.644854)`$\approx$`1-pt(1.644869,100000)`$=$ 0.05. Avšak napr. `1-pt(1.644869,100)`$\doteq 0.052$.

Príklad 58 (χ^2-rozdelenie) *Vypočítajte kritické hodnoty χ^2-rozdelenia so stupňami voľnosti $df = 10$, t.j. $\chi^2_{df}(\alpha)$, kde $\alpha = 0.1$, 0.05, 0.01, 0.025 a 0.005.*

Riešenie v (pozri obrázok 20)



Obr. 19: Grafické znázornenie významu pravdepodobnosti (obsahu) pod krivkou t-rozdelenia s $df = 10$; obsah pod (prvý riadok) a nad (druhý riadok) príslušným kvantilom

```
qchisq(1-0.1,10) # 15.98718
qchisq(1-0.05,10) # 18.30704
qchisq(1-0.01,10) # 23.20925
qchisq(1-0.025,10) # 20.48318
qchisq(1-0.005,10) # 25.18818
```

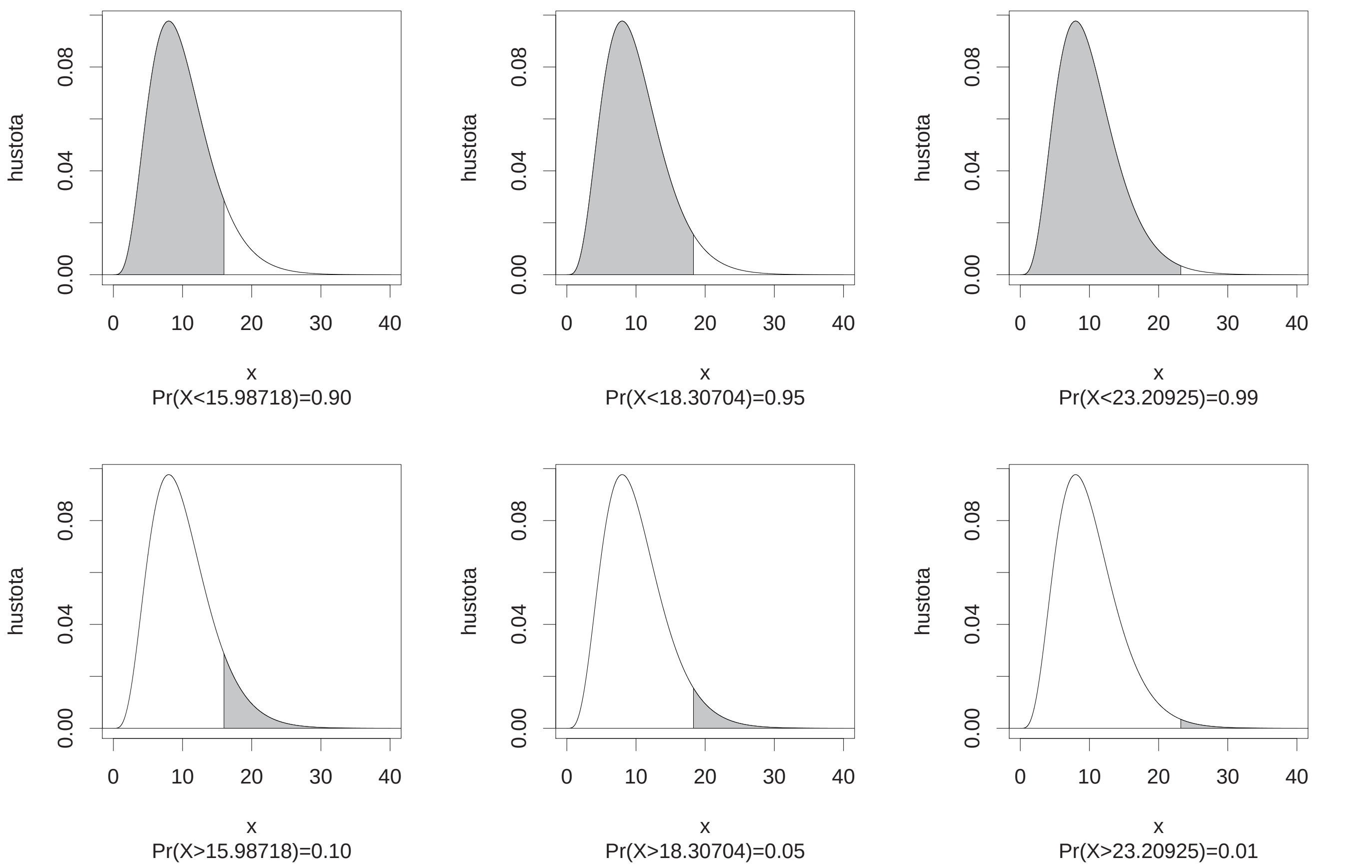
Na výpočet pravdepodobnosti pod kvantilom alebo nad kritickou hodnotou sa používa funkcia `pchisq(Q,df)`. Keďže χ^2-rozdelenie nie je symetrické, $\chi^2_{df}(\alpha) \neq \chi^2_{df}(1-\alpha)$.

Príklad 59 (F-rozdelenie) *Vypočítajte kritické hodnoty F-rozdelenia so stupňami voľnosti $df_1 = 20$ a $df_2 = 20$, t.j. $F_{df_1,df_2}(\alpha)$, kde $\alpha = 0.1$, 0.05, 0.01, 0.025 a 0.005.*

Riešenie v R (pozri obrázok 21)

```
qf(1-0.1,20,20) # 1.793843
qf(1-0.05,20,20) # 2.124155
qf(1-0.01,20,20) # 2.937735
qf(1-0.025,20,20) # 2.464484
qf(1-0.005,20,20) # 3.317786
```

Na výpočet pravdepodobnosti pod kvantilom sa používa funkcia `pf(Q,df1,df2)`. Na výpočet pravdepodobnosti nad kritickou hodnotou sa používa funkcia `1-pf(Q,df1,df2)`. Keďže F-rozdelenie nie je symetrické, $F_{df_1,df_2}(\alpha) \neq F_{df_1,df_2}(1-\alpha)$.



Obr. 20: Grafické znázornenie významu pravdepodobnosti (obsahu) pod krivkou χ^2-rozdelenia s $df = 10$; obsah pod (prvý riadok) a nad (druhý riadok) príslušným kvantilom

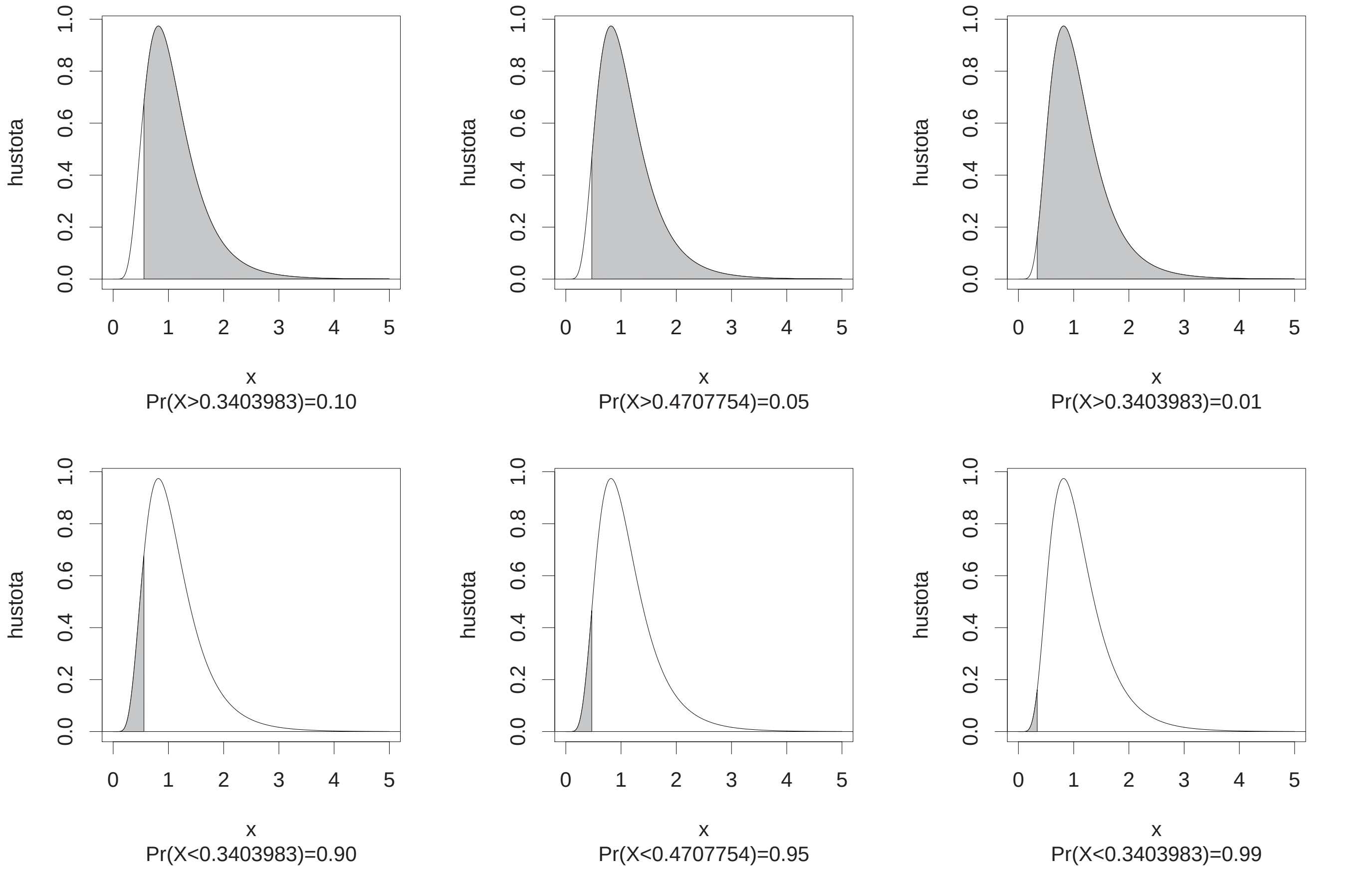
Veta 1 (koeficient variácie) *Nech náhodná premenná X pochádza z normálneho rozdelenia s parametrami μ a σ^2, $X \sim N(\mu, \sigma^2)$, kde $E[X] = \mu$ je stredná hodnota a $Var[X] = \sigma^2$ je rozptyl náhodnej premennej X. Nech $g(\boldsymbol{\theta}) = \sigma/\mu$, kde $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, $g(\widehat{\boldsymbol{\theta}}_n) = \frac{S_n}{\overline{X}_n}$ a $\Delta = \frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{\theta}) = \left(-\frac{\sigma}{\mu^2}, \frac{1}{2\sigma\mu}\right)^T$. Potom*

$$\sqrt{n}\left(\frac{S_n}{\overline{X}_n} - \frac{\sigma}{\mu}\right) \stackrel{\mathcal{D}}{\sim} N_k\left(0, \Delta^T (i(\boldsymbol{\theta}))^{-1}\Delta\right),$$

kde $\Delta^T (i(\boldsymbol{\theta}))^{-1}\Delta = \left(-\frac{\sigma}{\mu^2}, \frac{1}{2\sigma\mu}\right)\begin{pmatrix}\sigma^2 & 0 \\ 0 & 2\sigma^4\end{pmatrix}\left(-\frac{\sigma}{\mu^2}, \frac{1}{2\sigma\mu}\right)^T = \frac{\sigma^2}{\mu^2}\left(\frac{\sigma^2}{\mu^2} + \frac{1}{2}\right)$. pred

Príklad 60 (koeficient variácie) *Pomocou delta metódy odvoďte rozptyl koeficientu variácie z vety 1.* DÚ

Príklad 61 (MC experiment pre IS) *Nech (a) $X \sim N(0, 1)$ a (b) $X \sim [pN(0, 1)+(1-p)N(0, 4)]$, kde $p = 0.9$, t.j. ide o zmes dvoch normálnych rozdelení $X \sim N(0, 1)$ a $X \sim N(0, 4)$ v pomere 9:1. Vygenerujte $M = 100$ náhodných výberov s rozsahom $n = 500$ a vypočítajte $100(1-\alpha)\%$ empirický IS pre μ. Zistite, koľko IS obsahuje strednú hodnotu $\mu = 0$. Toto číslo podelené M predstavuje simulovanú hladinu významnosti α.* cvič.



Obr. 21: Grafické znázornenie významu pravdepodobnosti (obsahu) pod krivkou F-rozdelenia s $df_1 = 20$ a $df_2 = 20$; obsah pod (prvý riadok) a nad (druhý riadok) príslušným kvantilom

Definícia 6 (hodnoty distribučnej funkcie v kvantiloch) *Empirická distribučná funkcia $F_n(x)$ je definovaná nasledovne*

$$F_n(x) = \begin{cases} 0, & ak\ x < X_{(1)}, \\ \frac{i}{n}, & ak\ X_{(i)} \leq x < X_{(i+1)}, \\ 1, & ak\ x \geq X_{(n)}. \end{cases}$$

Majme transformáciu $T_{(1)} = F_n(X_{(1)})$, $T_{(2)} = F_n(X_{(2)}), \ldots, T_{(n)} = F_n(X_{(n)})$. Potom $T_{(1)}$, $T_{(2)}$, $\ldots, T_{(n)}$ sú ***poriadkové štatistiky****. Potom platí*

$$\lim_{n\to\infty} \Pr(\sup_{\forall x \in \mathcal{Y}} [F_n(x) - F(x)] n^{1/2} \leq \lambda) = \Phi(\lambda),$$

kde $F(X)$ je teoretická distribučná funkcia a $\Phi(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}$. Potom $100 \times (1-\alpha)\%$ pás spoľahlivosti pre $F_n(x)$ definujeme ako $F_n(x) \pm \lambda_\alpha 1/n^{1/2}$, kde $\Phi(\lambda_\alpha) = 1 - \alpha$ a $Var[F_n(x)] = 1/n$. Potom môžeme tvrdiť, že $F(X)$ patrí do $100 \times (1-\alpha)\%$ pásu spoľahlivosti a zároveň je medzi nulou a jednotkou s pravdepodobnosťou $1 - \alpha$. pred

Príklad 62 (graf distribučnej funkcie a jej IS) *Nakreslite graf distribučnej funkcie $N(\mu, \sigma^2)$, kde $\mu = 0$ a $\sigma^2 = 1$. Do grafu dokreslite 95% pás spoľahlivosti pre $F(x)$. Jeho hranice vypočítajte pomocou simulácie pseudonáhodných čísel z $N(0,1)$ pri $n = 50$, kde $F_n(x)$ je odhadnutá z dát.* cvič.