Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Štatistická inferencia I a II

Zadania príkladov a domácich úloh a niektoré riešenia

Stanislav Katina

katina@math.muni.cz

3. júna 2014

Obsah

1 Model rozdelenia pravdepodobnosti a štatistický model **1**
1.1 Simulačný experiment ako nástroj štúdia teoretických vlastností modelov 14
1.2 Štatistika 16
1.3 Funkcia vierohodnosti 17

2 Charakteristiky polohy a variability **21**

3 Testovanie hypotéz **22**
3.1 Testy dobrej zhody 24
3.2 Asymptotické testy o jednom parametri 26

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 závislá na premennej X pohlavie alebo (2) model náhodná premenná Y hrúbka kožných rias u dospelých zdravých žien závislá 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]*. 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 R

```
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] pred.

Riešenie (aj v R); (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
```

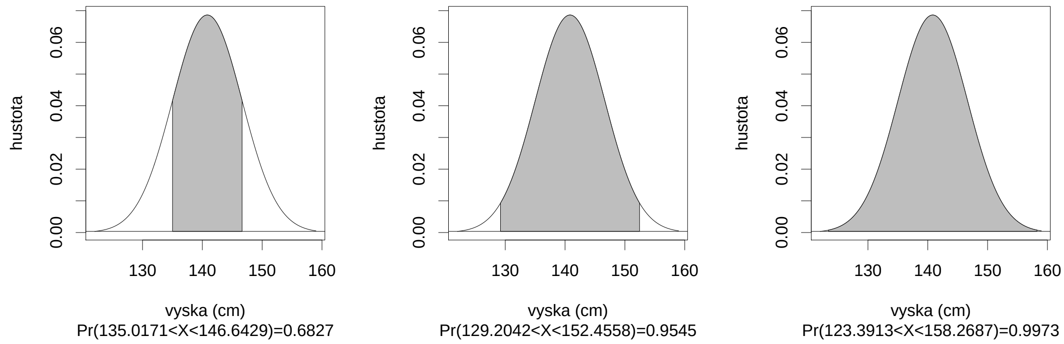
Dostaneme pravidlo $68.27 - 95.45 - 99.73$ (tzv. "**miery normálneho rozdelenia**").

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* $\Pr(a \leq X \leq b) = 1 - \alpha$, *bola rovná* $0.90, 0.95$ *a* 0.99. *Číslo* $x_{1-\alpha}$ *je kvantil normálneho normovaného rozdelenia, t.j.* $\Pr(Z = \frac{X-\mu}{\sigma} < x_{1-\alpha}) = 1 - \alpha, Z \sim N(0, 1)$. pred.

Riešenie (aj v R); (pozri obrázok 2)
$\Pr(\mu - x_{1-\alpha}\sigma < X < \mu + x_{1-\alpha}\sigma) = \Pr(X < \mu + x_{1-\alpha}\sigma) - \Pr(X < \mu - x_{1-\alpha}\sigma) = 1 - \alpha = 0.9$. Z-transformáciou[4] na normálne normované rozdelenie dostaneme $\Pr(-x_{1-\alpha} < X < x_{1-\alpha}) = 0.9$, kde $\frac{\mu - x_{1-\alpha}\sigma - \mu}{\sigma} = -x_{1-\alpha}$, $\frac{\mu + x_{1-\alpha}\sigma - \mu}{\sigma} = x_{1-\alpha}$, $x_{1-\alpha} = x_{0.9} = 1.64$, t.j. 90.00 % dát leží v intervale $\mu \pm 1.64\sigma$.
$\Pr(a < X < b) = 0.95$. Potom $x_{0.95} = 1.96$, t.j. 95.00 % dát leží v intervale $\mu \pm 1.96\sigma$.
$\Pr(a < X < b) = 0.99$. Potom $x_{0.99} = 2.58$, t.j. 99.00 % dát leží v intervale $\mu \pm 2.58\sigma$.

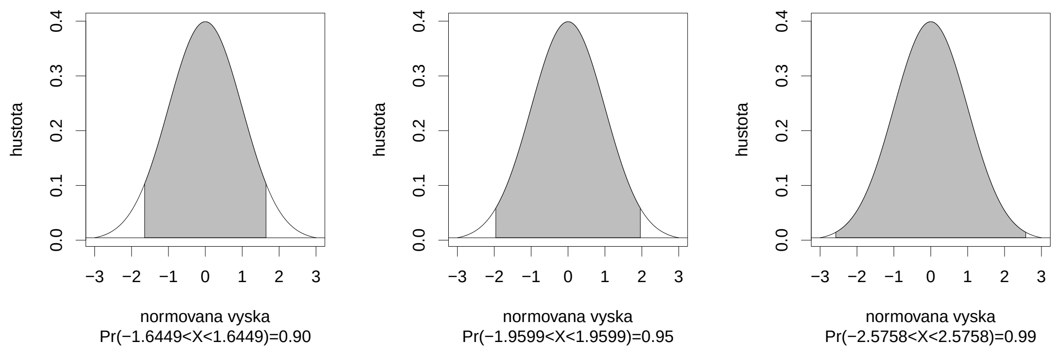
[3] Pravdepodobnosť $\Pr(a < X < b) = \Pr(a \leq X \leq b)$, pretože pravdepobnosť v bode (tu a a b) je rovná nule pre spojité premenné, t.j. $\Pr(a) = \Pr(b) = 0$. Pre diskrétne premenné to neplatí.

[4] Z-transformácia je spôsob transformácie náhodnej premennej $X \sim N(\mu, \sigma^2)$ pomocou centrovania strednou hodnotou μ a normovania smerodajnou odchýlkou σ, kde $Z = \frac{X-\mu}{\sigma}$; $Z \sim N(0, 1)$.



Obr. 1: 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 výškou v rozpätí týchto kvantilov

```
Q95 <- qnorm(0.95,0,1) # 1.644854
Q05 <- qnorm(0.05,0,1) # -1.644854
Q975 <- qnorm(0.975,0,1) # 1.959964
Q025 <- qnorm(0.025,0,1) # -1.959964
Q995 <- qnorm(0.995,0,1) # 2.575829
Q005 <- qnorm(0.005,0,1) # -2.575829
```



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

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.

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?* pred.

Riešenie (aj v R)
Pomocou Z-transformácie dostaneme
$\Pr(X > 160) = \Pr\left(\frac{X-132}{13} > \frac{160-132}{13}\right) = \Pr\left(\frac{X-132}{13} > 2.154\right) = 0.016.$

```
(1-pnorm(160,mean=132,sd=13))*100 # 1.562612 %
z.transf <- (160-132)/13
(1-pnorm(z.transf))*100 # 1.562612 %
```

Teda asi 1.6 % populácie z $N(132, 13^2)$ bude mať systolický krvný tlak väčší ako 160 mm Hg.

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?* pred.

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 (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 9 (poznámka k označeniu) *Pojem „model rozdelenia pravdepodobnosti“ sa často skracuje na „rozdelenie“. Potom hovoríme, že „X má rozdelenie* $F_X(x)$*“, „X je charakterizované rozdelením* $F_X(x)$*“ alebo „X pochádza z rozdelenia* $F_X(x)$*“, čo označujeme ako* $X \sim F_X(x)$, *kde symbol „*$\sim$*“ čítame ako „je rozdelená ako“ alebo „pochádza z rozdelenia“ (často sa uvádza aj pojem* ***„asymptoticky“****, čo znamená „pre veľké n“). Mohli by sme písať aj* $X \sim f_X(x)$, *to sa však používa len zriedkavo. Ak porovnávame rozdelenia dvoch náhodných premenných X a Y, hovoríme „X a Y majú rovnaké rozdelenie“ alebo „X a Y sú rovnako rozdelené“, ozn.* $X \sim Y$ *alebo* $F_X(x) \sim F_Y(y)$. *Pojem „štatistický model“ sa často skracuje na „model“.* pred.

Definícia 10 (aproximácia binomického rozdelenia normálnym) *Ak* $X \sim Bin(N,p)$, $Np > 5$ *a* $Nq > 5$, *kde* $q = 1-p$, *potom rozdelenie náhodnej premennej X môžeme aproximovať normálnym rozdelením , kde* $V \sim N(Np, Npq)$. pred.

Príklady minimálnych N pre fixované p potrebných na aproximáciu

p	0.1	0.2	0.3	0.4	0.5
q	0.9	0.8	0.7	0.6	0.5
N	51	26	17	13	11

Príklad 11 (aproximácia binomického rozdelenia normálnym) [5]*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 $N(Np, Npq)$.* cvič.

Riešenie (aj v R) (pozri obrázok 3 a 4)
(a) $E[X] = Np = 5 \times 0.515 = 2.575, E[Y] = 5 \times 0.485 = 2.425$,
$\Pr(X \leq 3) = \sum_{k \leq 3} \binom{5}{k} 0.515^k 0.485^{5-k} = 0.793$,
$\Pr(X \leq 3) = 0.648, N(5 \times 0.515, 5 \times 0.515 \times 0.485)$.
(b) $E[X] = 10 \times 0.515 = 5.15, E[Y] = 10 \times 0.485 = 4.85$,
$\Pr(X \leq 5) = \sum_{k \leq 5} \binom{10}{k} 0.515^k 0.485^{10-k} = 0.586$,
$\Pr(X \leq 5) = 0.462, N(10 \times 0.515, 10 \times 0.515 \times 0.485)$.
(c) $E[X] = 50 \times 0.515 = 25.75, E[Y] = 50 \times 0.485 = 24.25$,
$\Pr(X \leq 25) = \sum_{k \leq 25} \binom{50}{k} 0.515^k 0.485^{50-k} = 0.471$,
$\Pr(X \leq 25) = 0.416, N(50 \times 0.515, 50 \times 0.515 \times 0.485)$.

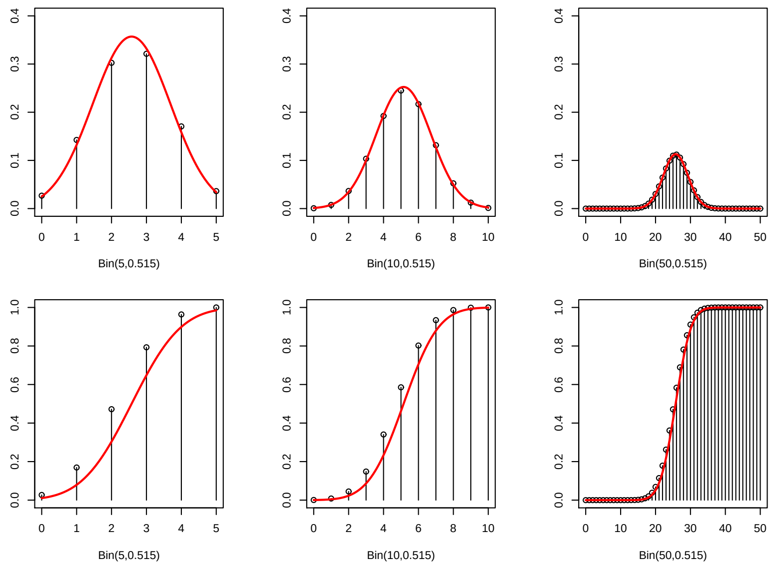
```
pbinom(3,size=5,prob=0.515) # 0.7931878
pnorm(3,mean=5*0.515,sd=sqrt(5*0.515*0.485)) # 0.6481396
pbinom(5,size=10,prob=0.515) # 0.5856244
pnorm(5,mean=10*0.515,sd=sqrt(10*0.515*0.485)) # 0.4621927
pbinom(25,size=50,prob=0.515) # 0.4712842
pnorm(25,mean=50*0.515,sd=sqrt(50*0.515*0.485)) # 0.4159648
```

Z vyššie uvedeného príkladu vyplýva, že pre pravdepodobnosť $p = 0.515$ a $N = 50$ aproximácia stále nie je postačujúca (ani na jedno desatinné miesto) a pre $N = 10$ a $N = 5$ ju nie je možné použiť. Pre pravdepodobnosti p blížiace sa jednotke alebo nule sú potrebné väčšie početnosti ako pre pravdepodobnosti p blízke hodnote 0.5.

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.

[5]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)

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

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

s hustotou

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

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*

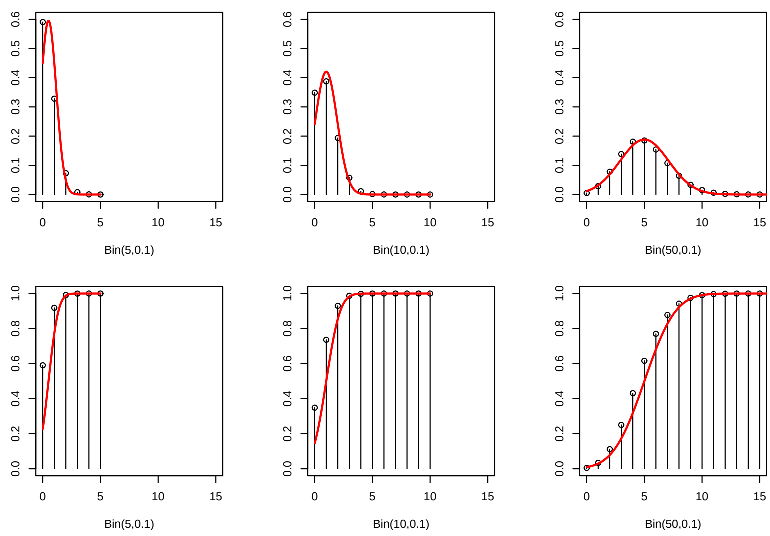
$$-\frac{1}{2}\begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix},$$

marginálne rozdelenia[6] *sú* $X \sim N(\mu_1, \sigma_1^2)$ *a* $Y \sim N(\mu_2, \sigma_2^2)$, ρ *je koeficient korelácie*[7] *(pozri obrázok 5).* cvič.

Príklad 15 (dvojrozmerné normálne rozdelenie) *Nech náhodnou premennou* X *je najväčšia výška mozgovne (`skull.pH`; v mm) a náhodnou premennou* Y *je morfologická výška tváre (`face.H`; v mm). 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*

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

[7] 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)

hodnôt a σ_1^2, σ_2^2 a ρ, čo sú parametre kovariančnej matice $\mathbf{\Sigma}$, kde sila lineárneho vzťahu týchto dvoch premenných je daná veľkosťou a znamienkom ρ. Možno predpokladať, že oba rozmery spolu pomerne silno korelujú (ρ bude číslo blížiace sa jednotke) a tvar dvojrozmernej hustoty sa bude blížiť prostrednému stĺpcu na obrázku 5. cvič.

Príklad 16 (štandardizované dvojrozmerné normálne rozdelenie) *Náhodný vektor $(X, Y)^T$ má dvojrozmerné normálne rozdelenie*

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

s hustotou

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

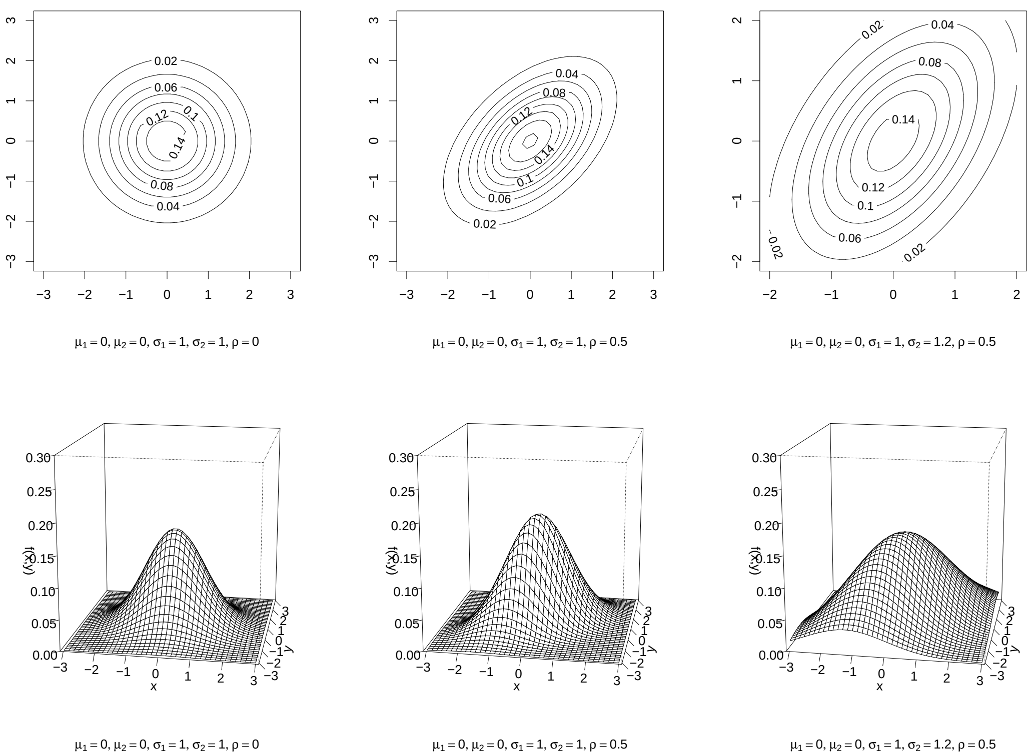
kde $(x, y)^T \in \mathbb{R}^2$, $\rho \in \langle -1, 1 \rangle$ sú parametre, potom $\boldsymbol{\theta} = (0, 0, 1, 1, \rho)$. Výraz v exponente môžeme písať ako

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

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

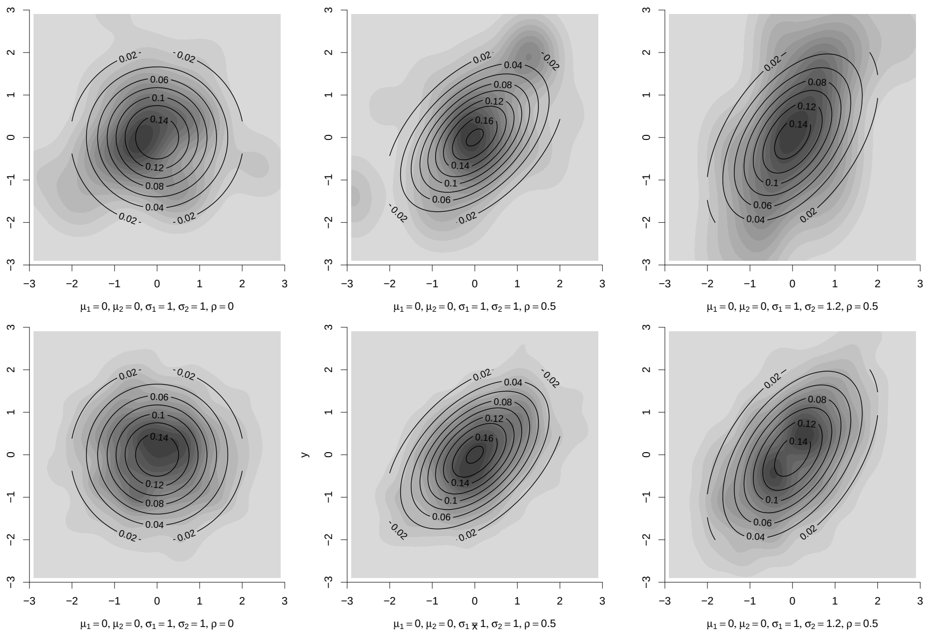
Príklad 17 (štandardizované dvojrozmerné normálne rozdelenie) *Nech náhodnou premennou $X \sim N(\mu_1, \sigma_1^2)$ je najväčšia výška mozgovne (`skull.pH`; v mm) a náhodnou premennou $Y \sim N(\mu_2, \sigma_2^2)$ je morfologická výška tváre (`face.H`; v mm). Nech X a Y majú dvojrozmerné normálne rozdelenie s parametrami $(\mu_1, \mu_2)^T$ a σ_1^2, σ_2^2 a ρ sú parametre kovariančnej matice $\mathbf{\Sigma}$. 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.



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)

Príklad 18 (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()`*;*
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.



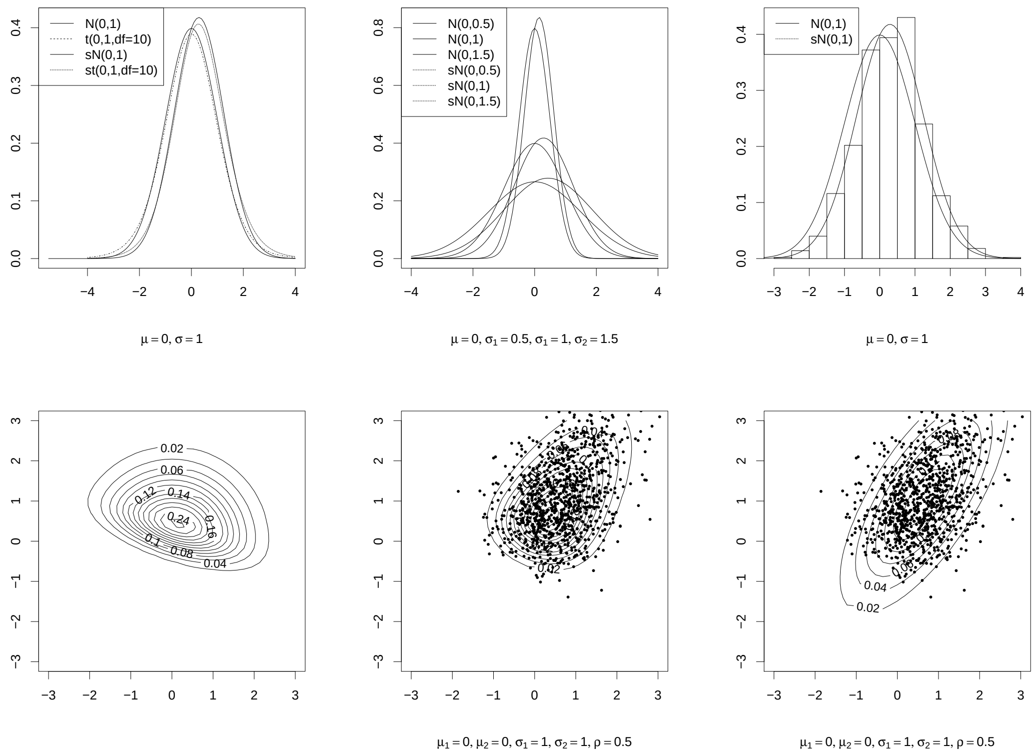
Obr. 6: Hustoty dvojrozmerného normálneho rozdelenia (prvý riadok $n = 50$; druhý riadok $n = 1000$)

Odlišnosti od teoretického rozdelenia. Odlišnosti empirického rozdelenia (rozdelenia realizácií) od teoretického (napr. normálneho) rozdelenia, môžeme charakterizovať napr. ako pravostranne alebo ľavostranne zošikmené rozdelenie (obrázok 7, prvý riadok vľavo a vpravo), ploché alebo špicaté rozdelenie (obrázok 7, 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 7, dolný riadok).

Príklad 19 (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 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).* cvič.

Riešenie v R (pozri obrázok 8)

```
par(mfrow=c(1,2),mar = c(6,5,1,1), pty="s")
plot(0:5, dbinom(0:5,5,0.5), type="h", xlab="x",ylab="Pr(X=x)",
xlim=c(-1,6))
title(sub="hustota pre X~Bin(5, 0.5)")
plot(0:5, pbinom(0:5,5,0.5), type="n", xlab="x", ylab="F(x)",
xlim=c(-1,6), ylim=c(0,1))
segments(-1,0,0,0)
segments(0:5, pbinom(0:5,5,.5), 1:6, pbinom(0:5,5,.5))
lines(0:5, pbinom(0:5,5,.5), type="p", pch=16)
segments(-1,1,9,1, lty=2)
title(sub="distribucna funkcia pre X~Bin(5, 0.5)")
```



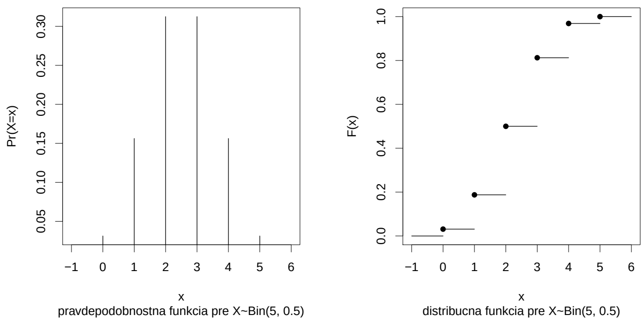
Obr. 7: Hustoty normálneho rozdelenia a zošikmeného normálneho rozdelenia pri rôznych parametroch (prvý riadok); hustoty dvojrozmerného zošikmeného normálneho rozdelenia (druhý riadok vľavo a uprostred) a dvojrozmerného normálneho rozdelenia (druhý riadok vpravo) pri rôznych parametroch

Príklad 20 (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:



Obr. 8: Pravdepodobnostná funkcia a distribučná funkcia $Bin(5, 0.5)$

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

Príklad 21 (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

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 22 (farba očí a vlasov) *Majme premenné farba vlasov (blond BlH, hnedá BrH, ryšavá RH) a farba očí (modrá BlE, hnedá BrE, zelená GE). Ich interakcie sú usporiadané v tabuľke ako* X_1 *(BlH-BlE),* X_2 *(BlH-BrE),* X_3 *(BlH-GE),* X_4 *(BrH-BlE),* X_5 *(BrH-BrE),* X_6 *(BrH-GE),* X_7 *(RH-BlE),* X_8 *(RH-BrE),* X_9 *(RH-GE). Nim zodpovedajúce pravdepodobnosti* $p_j, j = 1, 2, \ldots, 9$*, sú v tabuľke (pozri tabuľku).* $\mathbf{X} = (X_1, X_2, \ldots, X_9)^T \sim Mult_9(N, \mathbf{p})$*. Transformujte multinomický model na súčinový multinomický model nasledovne – vypočítajte (a) riadkové marginálne pravdepodobnosti* $p_{1\cdot} = \sum_{j=1}^{3} p_j$*,* $p_{2\cdot} = \sum_{j=4}^{6} p_j$*,* $p_{3\cdot} = \sum_{j=7}^{9} p_j$*, (b) stĺpcové marginálne pravdepodobnosti* $p_{\cdot 1} = p_1 + p_4 + p_7$*,* $p_{\cdot 2} = p_2 + p_5 + p_8$*,* $p_{\cdot 3} = p_3 + p_6 + p_9$*, (c) podmienené pravdepodobnosti* $p_{j|k} = p_{kj}$*; (d) podmienené pravdepodobnosti* $p_{k|j} = p_{jk}$*; (d) akému číslu sú rovné sumy* $\sum_{j=1}^{3} p_{j|k}$ *pre každé* k *a* $\sum_{k=1}^{3} p_{k|j}$ *pre každé* j*?*

farba vlasov/farba očí	modrá (BlE)	hnedá (BrE)	zelená (GE)
blond (BlH)	0.12	0.15	0.03
hnedá (BrH)	0.22	0.34	0.04
ryšavá (RH)	0.06	0.01	0.03

Riešenie (čiastkové)
Marginálne pravdepodobnosti sú
$\Pr(BlH) = 0.3$, $\Pr(BrH) = 0.6$, $\Pr(RH) = 0.1$,
$\Pr(BlE) = 0.4$, $\Pr(BrE) = 0.5$, $\Pr(GE) = 0.1$.
Podmienené pravdepodobnosti $p_{k|j}$ sú
$\Pr(BlH|BlE) = \Pr(BlH \cap BlE) / \Pr(BlE) = 0.12/0.4 = 0.3$,
$\Pr(BlH|BlE) = \Pr(BlH)$,
$\Pr(BrH|BlE) = 0.22/0.4 = 0.55$,
$\Pr(BrH) = 0.6$.
Ak vieme, že niekto má modré oči, potom bude menej pravdepodobné, že má hnedé vlasy v porovnaní s tým, keď nevieme, akej farby má oči. Teda
$\Pr(BlE|BlH) = 0.12/0.3 = 0.4$,
$\Pr(BlE|BlH) = \Pr(BlE)$,
$\Pr(BrE|BlH) = \Pr(BrE)$,
$\Pr(GE|BlH) = \Pr(GE)$.
Informácia, že má niekto blond vlasy, nám nedáva ďalšiu informáciu o farbe jeho očí.

Binomické, multinomické a súčinové multinomické rozdelenie sú vhodné v prípadoch, keď máme počet pokusov N nie príliš veľký a pravdepodobnosti výskytu udalostí p nie príliš malé. V opačnom prípade je vhodné Poissonovo rozdelenie.

Príklad 23 (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* $50mil \times 0.000002 = 100$*.* pred.

Príklad 24 (Poissonovo rozdelenie; pruské armádne jednotky) *Nech početnosti úmrtí* X *ako následok kopnutia koňom v Pruských armádnych jednotkách, 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* $n = 1, 2, 3, 4, \geq 5$*, 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 prepokladu* $X \sim Poiss(\lambda)$*, kde* $\lambda = \frac{\sum_n n m_n}{\sum_n m_n}$*.* DÚ

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

Príklad 25 (Poissonove rozdelenie; tri typy havárií) *Nech n_1 je počet ľudí, ktorí zahynú pri automobilovej nehode, n_2 je počet ľudí, ktorí zahynú pri havárii lietadla, n_3 je počet ľudí, ktorí zahynú pri havárii vlaku v Taliansku budúci týždeň. Potom Poissonov model pre (X_1, X_2, X_3) vytvára nezavislé poissonovské náhodné premenné s parametrami $(\lambda_1, \lambda_2, \lambda_3)$ a $X_1 + X_2 + X_3 \sim Poiss(\lambda_1 + \lambda_2 + \lambda_3)$.* pred.

Príklad 26 (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 27 (overdispersion v binomickom modeli) *V klasickej štúdii pomeru pohlaví u ľudí z roku 1889 na základe záznamov z nemocníc v Sasku Geissler 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 $\pi = \frac{\sum_{n=0}^{N} n m_n}{NM} = 0.5192$ 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 28 (overdispersion v Poissonovom modeli) *Majme početnosti úrazov n medzi robotníkmi v továrni, kde početnosti robotníkov m_n pri danom n pozri v tabuľke.*

n	0	1	2	3	≥ 4
m_n	447	132	42	21	5

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

Riešenie

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

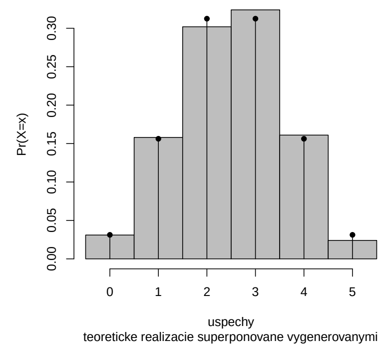
Keď porovnáme pozorované m_n a vypočítané (teoretické, očakávané) m_n zistíme, že pozorované poukazujú na *overdispersion*, t.j. máme viac robotníkov bez úrazu ako aj viac robotníkov s väčším množstvom úrazov v porovnaní s teroretickými početnosťami.

1.1 Simulačný experiment ako nástroj štúdia teoretických vlastností modelov

Príklad 29 (binomický experiment, simulačná štúdia) *Vygenerujte pseudonáhodné čísla opakované M-krát ($M = 1000$) z $Bin(5, 0.5)$. Vytvorte tabuľku vygenerovaných ako aj teoretických realizácií (pre $n = 0, 1, \ldots, 5$), superponujte histogram vygenerovaných realizácií s pravdepodobnostnou funkciou teoretických realizácií (pozri obrázok 9).* cvič.

Riešenie

r	0	1	2	3	4	5
simulované realizácie	0.031	0.158	0.302	0.324	0.161	0.024
teoretické realizácie	0.031	0.156	0.312	0.312	0.156	0.031



Obr. 9: Teoretické realizácie (relatívne početnosti) z $Bin(5, 0.5)$ superponované vygenerovanými ($M = 1000, n = 5$); histogram superponovaný spojnicovým grafom

Príklad 30 (CLV pre binomické rozdelenie) [8]*Na základe CLV môžeme tvrdiť, že pre náhodnú premennú X_N platí $Z_N = \frac{X_N - Np}{\sqrt{Np(1-p)}} \sim N(0, 1)$, t.j. X_N má pre dostatočne veľké N asymptoticky normálne rozdelenie $X_N \sim N(Np, Np(1-p))$. Ukážte, že CLV platí pre $X_N \sim Bin(N, p)$, ak $N = 100, p = 1/2$, na tri desatinné miesta.* cvič.

Riešenie (aj v R)
$E[X_N] = Np = 50$, $\sqrt{Var[X_N]} = \sqrt{Np(1-p)} = \sqrt{5}$.
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 $Z_{100} \sim N(50, 5)$.

```
pbinom(51,100,.5) - pbinom(48,100,.5) # 0.2356466
pnorm(51.5,50,5) - pnorm(48.5,50,5) # 0.2358228
```

Výsledky sa zhodujú na tri desatinné miesta. Všeobecne platí $X_M \sim N(M/2, M/4)$ a $Y_M = X_M/M \sim N(1/2, 1/(4M))$.

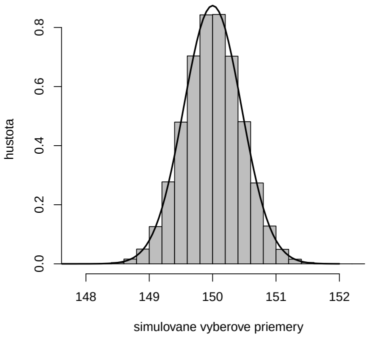
[8]Príklad hovorí o tom, ako dobre normálne rozdelenie aproximuje binomické pri rozsahu $N = 100$, čo je dôležité pri testovaní hypotéz.

Príklad 31 (CLV pre normálne rozdelenie) [9]*Na základe simulačnej štúdie (M=500000) preverte, že ak* $X_n \sim N(150, 6.25)$*, potom* $\overline{X}_n \sim N(150, 6.25/n)$ *pre* $n = 30$*. Vypočítajte* $\Pr(\overline{X}_n > 151)$ *zo simulovaných dát a porovnajte tento výsledok s teoretickou (očakávanou) pravdepodobnosťou.* cvič.

Riešenie (aj v R) (pozri obrázok 10)
$\Pr(\overline{X}_n > 151) = \Pr(\frac{\overline{X}_n - 150}{\sqrt{6.25}/\sqrt{n}} > \frac{151-150}{\sqrt{6.25}/\sqrt{n}}) \approx \Phi(2.190890) = 0.01422987.$

```
1-pnorm((151-150)/sqrt(6.25/30)) # 0.01422987
M <- 500000; n <- 30
x <- rnorm(M*n,150,sqrt(6.25))
x.mat <- matrix(x,M)
x.bar <- rowMeans(x.mat)
mean(x.bar > 151) # 0.014238
hist(x.bar, probability = TRUE, col="gray", main="",
     ylab="hustota",xlab="simulovane vyberove priemery")
curve(dnorm(x,150,sqrt(6.25/30)),from=147,to=152,lwd=2, add = TRUE)
```

Pri dostatočne veľkom počte opakovaní vidíme zhodu medzi teoretickým a simulovaným rozdelením $\overline{X}_n$ na tri desatiné miesta (pri výpočte zadanej pravdepodobnosti).



Obr. 10: Teoretické realizácie z $N(150, 6.25/30)$ v podobe krivky hustoty superponované vygenerovanými ($M = 500000$, $n = 30$) v podobe histogramu

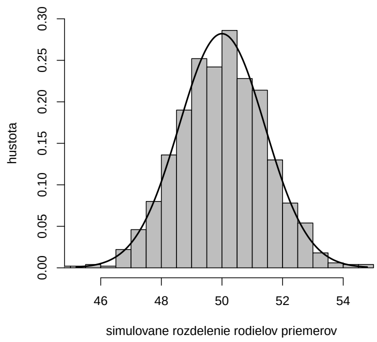
Príklad 32 (CLV pre normálne rozdelenie, jeden náhodný výber) [10]*Majme náhodné výbery s rozsahmi* $n = 2, 5, 20, 50, 100$ *a* 500 *z rozdelení (a)* $N(\mu, \sigma^2), \mu = 0, \sigma^2 = 1$*, (b)* $Exp(\lambda), \lambda = 1/3$*, (c)* $Unif(\min, \max), \min = 0, \max = 1$*, (d) zmes dvoch* $N(\mu, \sigma^2)$*:* $0.1 \times N(0, 10) + 0.9 \times N(0, 1)$*. Použite R na simuláciu náhodných výberov (počet simulácií je* $M = 1000$*), pre každú simuláciu vypočítajte aritmetické priemery* $\overline{x}_m, m = 1, 2, \ldots, M$ *a zobrazte ich do histogramu superponovaného krivkou hustoty teoretického rozdelenia* $N(\mu, \sigma^2/n)$ *prislúchajúceho danej simulácii.* cvič.

[9]Príklad hovorí o tom, že ak má náhodná premenná X_n normálne rozdelenie, bude mať normálne rozdelenie aj aritmetický priemer $\overline{X}_n$, čo je dôležité pri testovaní hypotéz.

[10]Príklad slúži na zistenie vlastností rozdelenia aritmetického priemeru pri rôznych situáciách. $Exp(\lambda)$ je exponenciálne rozdelenie s parametrom λ, $Unif(\min, \max)$ je rovnomerné rozdelenie s parametrami min a max. Zmes dvoch normálnych rozdelení predstavuje 10% prímes normálneho rozdelenia s väčším rozptylom rovným $\sigma^2 = 10$ v normálnom rozdelení s menším rozptylom rovným $\sigma^2 = 1$, čím sme docielili výskyt 10 % odľahlých pozorovaní.

Príklad 33 (CLV pre normálne rozdelenie, dva náhodné výbery) [11] *Preverte normalitu rozdelenia rozdielu $\overline{X}_{n_1} - \overline{Y}_{n_2}$, teda $\overline{X}_{n_1} - \overline{Y}_{n_2} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$, pomocou simulačnej štúdie. Generujte pseudonáhodné čísla $M = 1000$-krát z rozdelení $N(\mu_j, \sigma_j^2), j = 1, 2$, kde $\mu_1 = 100, \sigma_1 = 10, \mu_2 = 50, \sigma_2 = 9$ pre (a) $n_1 = 4, n_2 = 5$, (b) $n_1 = 100, n_2 = 81$. Pre prípad (a) aj (b) vypočítajte* $\Pr(\overline{X}_{n_1} - \overline{Y}_{n_2}) < 52$ *na základe empirického (zo simulácie) a teoretického rozdelenia* $\overline{X}_{n_1} - \overline{Y}_{n_2}$. DÚ

Pri dostatočne veľkom počte opakovaní vidíme zhodu medzi teoretickým a simulovaným rozdelením $\overline{X}_{n_1} - \overline{Y}_{n_2}$ na dve desatiné miesta (pri výpočte zadanej pravdepodobnosti; pozri obrázok 11).



Obr. 11: Teoretické realizácie rozdelenia $\overline{X}_{n_1} - \overline{Y}_{n_2}$ v podobe krivky hustoty superponované vygenerovanými v podobe histogramu ($M = 1000$, $n_1 = 100$, $n_2 = 81$)

1.2 Štatistika

Príklad 34 (štatistika) [12] *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 35 (CLV pre binomické rozdelenie, testovacia štatistika) [13] *Ak náhodná premenná $X \sim Bin(N, p)$, preverte normalitu rozdelenia Z, kde testovacia štatistika $Z = \frac{X/N - p}{\sqrt{(p(1-p)/N}} \sim N(0, 1)$ testovacej štatistiky pomocou simulačnej štúdie ($p = 0, 0.1, 0.5, 0.9, 1$; $N = 5, 10, 30, 50, 100$; $M = 1000$). Okomentujte výsledky v spojitosti s Haldovou podmienkou $Np(1-p) > 9$.* cvič.

Príklad 36 (CLV pre normálne rozdelenie, testovacia štatistika) [14] *Zistite pomocou simulačnej štúdie (počet opakovaní $M = 1000$), či testovacia štatistika $F = \frac{(n-1)S^2}{\sigma^2}$ má asymptoticky χ^2_{n-1} rozdelenie s $n - 1$ stupňami voľnosti, ak (a) $X \sim N(\mu, \sigma^2)$, kde $\mu = 0, \sigma^2 = 1$ a (b) $X \sim Exp(\lambda)$ (exponenciálne rozdelenie s parametrom $\lambda = 1$), kde $E[X] = 1$ a $\sigma^2 = 1$. Rozsahy náhodných výberov sú pre oba prípady $n = 15$ a $n = 100$.* DÚ

[11] Príklad slúži na zistenie vlastností rozdelenia rozdielu dvoch aritmetických priemerov pri rôznych situáciách.

[12] Štatistiky teda môžu byť náhodné premenné alebo náhodné vektory, ktoré sumarizujú informáciu o dátach, zjednodušujú pohľad na ne a umožňujú na ich základe dáta jednoduchšie popísať a ľahšie interpretovať.

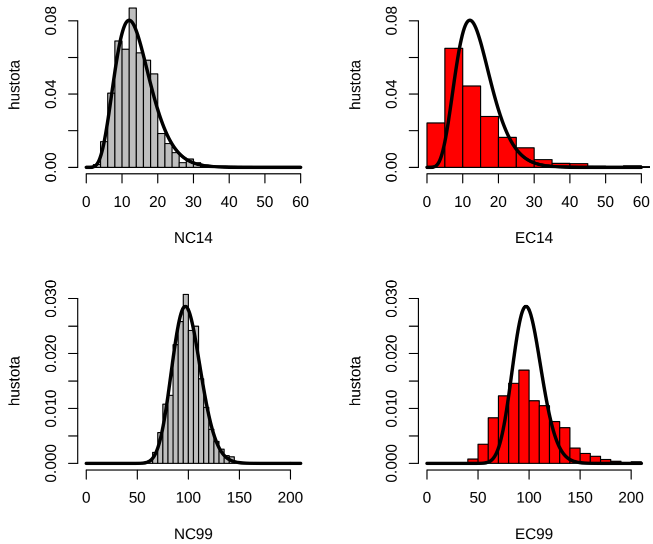
[13] Príklad hovorí o použití jednovýberovej testovacej štatistiky pre parameter binomického rozdelenia (pravdepodobnosť) pre rôzne pravdepodobnosti a rôzne početnosti. Ak Haldova podmienka nie je splnená, nie je možné testovaciu štatistiku použiť.

[14] Príklad hovorí o použití jednovýberovej testovacej štatistiky pre parameter normálneho rozdelenia (rozptyl) pre rôzne teoretické rozdelenia a rôzne rozsahy náhodných výberov. Ak sú výchylky od normality príliš veľké, nie je možné testovaciu štatistiku použiť.

Riešenie
Vieme, že stredná hodnota $E[F] = n-1$ a $Var[F] = 2(n-1)$, t.j. chceme, aby sa výsledky simulačnej štúdie priblížili týmto teoretickým výsledkom (pozri tabuľku).

odhady počítané pri simulácii	$E[S^2]$	$Var[S^2]$	$E[F]$	$Var[F]$
normálne rozdelenie, $n = 15$	0.9900	0.1451	13.8599	28.4403
exponenciálne rozdelenie, $n = 15$	1.0637	0.6629	14.8920	129.9219
normálne rozdelenie, $n = 100$	0.9952	0.0202	98.5274	198.3750
exponenciálne rozdelenie, $n = 100$	0.9958	0.0766	98.5866	750.4624

Pri dostatočne veľkom počte opakovaní vidíme zhodu medzi teoretickým a simulovaným rozdelením F, len ak ide o dáta z normálneho rozdelenia (pozri obrázok 12).



Obr. 12: Teoretické realizácie rozdelenia F superponované vygenerovanými (empirickými) pre $X \sim N(0,1)$ (ľavý stĺpec) a $X \sim Exp(1)$ (pravý stĺpec) ($M = 1000$, $n = 15$ (horný riadok), $n = 100$ (dolný riadok); histogram empirického rozdelenia realizácií v relatívnej škále superponovaný krivkou hustoty normálneho rozdelenia

1.3 Funkcia vierohodnosti

Príklad 37 (princípy vierohodnosti) *Majme binomické rozdelenie (N je fixované a náhodná premenná je počet úspechov) a negatívne binomického rozdelenia (počet úspechov je fixovaný vopred a náhodná premenná je počet zlyhaní pozorovaný pred zastavením sekvencie pokusov). Ak x_1 je počet úspechov a x_2 počet neúspechov a θ pravdepodobnosť úspechu, potom*

$$f_1(x_1, \theta) = \binom{N}{x_1} \theta^{x_1}(1-\theta)^{x_2}, x_1 = 1, 2, \dots, n$$

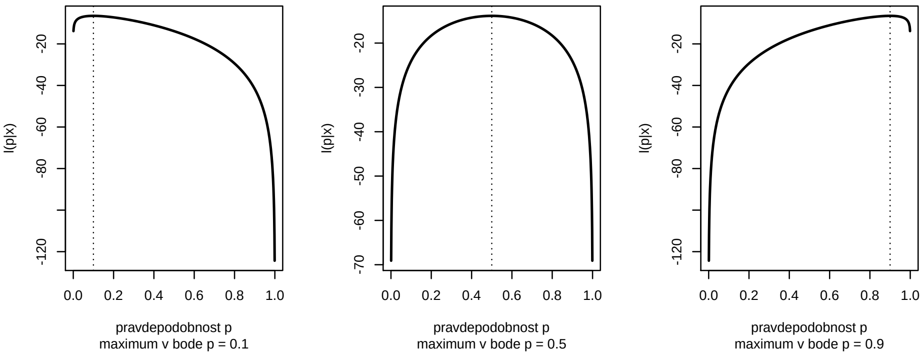
a

$$f_2(x_2, \theta) = \binom{x_1 + x_2 - 1}{x_2} \theta^{x_1}(1-\theta)^{x_2}, x_2 = 1, 2, \dots,$$

kde funkcia vierohodnosti pre oba prípady bude $L(\theta) = c\theta^{x_1}(1-\theta)^{x_2}$. pred.

Príklad 38 (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.

Riešenie (pozri obrázok 13)
(a) $\widehat{p} = x/N = 2/20 = 0.1$,
(b) $\widehat{p} = x/N = 10/20 = 0.5$,
(c) $\widehat{p} = x/N = 18/20 = 0.9$.
Logaritmus funkcie vierohodnosti pre p má tvar $l((p|\mathbf{x}) = n\log(p) + (N-n)\log(1-p)$, kde $p \in (0, 1)$.



Obr. 13: Funkcia vierohodnosti pre $X \sim Bin(N, p)$ ($p = 0.1, 0.5, 0.9$ a $N = 20$)

Z grafov (Obr. 13) je zreteľné, že funkcia vierohodnosti pre p je symetrická len pre $p = 0.5$, pre ostatné p je asymetrická. Naviac pre p a $1 - p$ dostaneme grafy, ktoré možno transformovať jeden na druhý pomocou osi zrkadlenia definovanej ako vertikálna priamka v $p = 0.5$.

Príklad 39 (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, odhadnite jeho parameter λ, ktorý predstavuje očakávanú početnosť zemetrasení za rok.* cvič.

Riešenie
Pomocou logaritmu funkcie vierohodnosti $l(\lambda|\mathbf{x}) = \sum_{i=1}^{N} x_i \ln\lambda - N\lambda, N = 5$, vieme vypočítať $\widehat{\lambda} = \frac{\sum x_i}{N} = \overline{x}$, ktorý je rovný 2.8.

Vo všeobecnosti píšeme funkciu vierohodnosti pre Poissonove rozdelenie s parametrom λ a pozorovanými početnosťami m_n ako $L(\lambda|\mathbf{x}) = \prod_n p_n^{m_n}$, kde $p_n = \Pr(X = n) = e^{-\lambda}\lambda^n/n!$ a logaritmus funkcie vierohodnosti ako $l(\lambda|\mathbf{x}) = -\lambda\sum_n m_n + \sum_n nm_n \ln\lambda$. Maximálne vierohodný odhad $\widehat{\lambda} = \frac{\sum_n nm_n}{\sum_n m_n}$.

Príklad 40 (overdispersion v binomickom modeli, pokrač.) *Majme početnosti úrazov n medzi robotníkmi v továrni, kde početnosti robotníkov m_n pri danom n pozri v tabuľke.*

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.

Riešenie Aby sme mohli fitovať negatívne binomické rozdelenie, potrebujeme funkciu vierohodnosti

$$L(\alpha,\pi|\mathbf{x}) = \Pr(X=x)^{\sum_{n=0}^4 m_n} \left(\Pr(X\geq 5)\right)^{m_{\geq 5}}.$$

a jej logaritmus

$$l(\alpha,\pi|\mathbf{x}) = \sum_{n=0}^{4} m_n \ln \Pr(X=x) + m_{\geq 5}\ln\left(\Pr(X\geq 5)\right).$$

Numerickou optimalizáciou dostaneme $\widehat{\alpha} = 0.84$ a $\widehat{\pi} = 0.64$. Pomer zlyhaní $\widehat{\mu} = \frac{1-\widehat{\pi}}{\widehat{\pi}}\widehat{\alpha} = 0.47$. Keď porovnáme pozorované m_n a vypočítané (teoretické) m_n zistíme, že početnosti sú veľmi podobné (pozri tabuľku).

n	0	1	2	3	4	≥ 5
očakávané m_n	446	134	44	15	5	3

Príklad 41 (kvadraticá aproximácia funkcie vierohodnosti) *(1) Nakreslite škálovaný logaritmus funkcie vierohodnosti binomického rozdelenia. Na x-ovej osi bude p a na y-ovej osi $l^*(p|\mathbf{x}) = l(p|\mathbf{x}) - \max(l(p|\mathbf{x}))$. Porovnajte $l^*(p|\mathbf{x})$ s kvadratickou aproximáciou vypočítanou pomocou Taylorovho rozvoja $\ln(\frac{L(p|\mathbf{x})}{L(\widehat{p}|\mathbf{x})}) \approx -\frac{1}{2}\mathcal{I}(\widehat{p})(p-\widehat{p})^2$. (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* 14. DÚ

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

Riešenie
Logaritmus funkcie vierohodnosti má tvar

$$l(\mu,\sigma^2) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

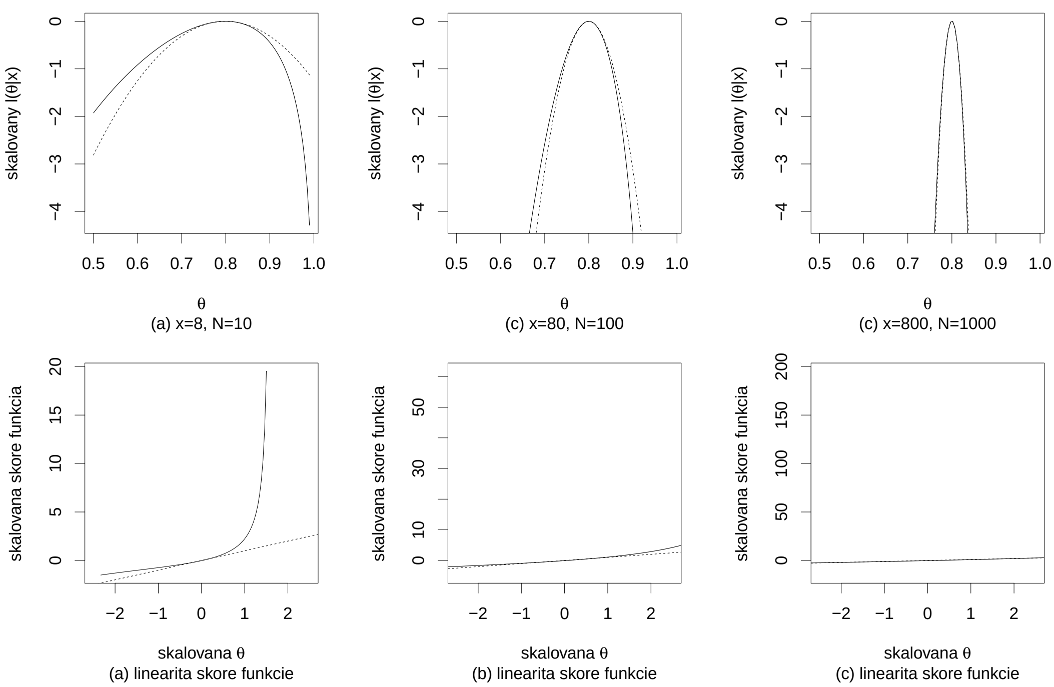
Derivácie funkcie vierohodnosti v μ a σ^2 budú nasledovné

$$S_1(\mu,\sigma^2) = \frac{\partial}{\partial\mu} l((\mu,\sigma^2)|\mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu),$$

$$S_2(\mu,\sigma^2) = \frac{\partial}{\partial\sigma^2} l((\mu,\sigma^2)|\mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Potom

$$\mathcal{I}(\widehat{\mu},\widehat{\sigma}^2) = \begin{pmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\widehat{\sigma}^4} \end{pmatrix}.$$



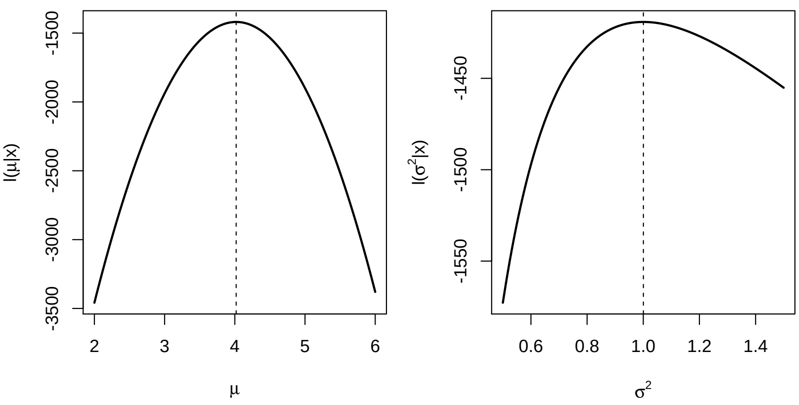
Obr. 14: Porovnanie škálovaného logaritmu funkcie vierohodnosti (plná čiara) s jeho kvadratickou aproximáciou (čiarkovaná čiara) v prvom riadku a porovnanie škálovanej skóre funkcie a priamky s nulovým interceptom a jednotkovým sklonom v druhom riadku

Príklad 43 (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 44 (maximálne vierohodný odhad μ a σ^2) [15] *Vygenerujte pseudonáhodné čísla z* $X \sim N(4,1)$*,* $n=1000$*. (a) Napíšte profilovú funkciu vierohodnosti pre* μ *a* σ^2 *a preverte, či sú simulované 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 simulovaných maxím týchto funkcií. (b) Napíšte funkciu vierohodnosti pre* $\boldsymbol{\theta}=(\mu,\sigma^2)$ *a preverte, či je simulovaný maximálne vierohodný odhad* $\boldsymbol{\theta}=(\mu,\sigma^2)$ *dostatočne blízko k jeho skutočnej hodnote.* cvič.

Riešenie (pozri obrázok 15)
Logaritmus funkcie vierohodnosti pre jednotlivé parametre má tvar
$l(\mu|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma_1^2 - \frac{1}{2\sigma_1^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right)$, kde $\mu\in(2,6), \sigma_1 = 1$;
$l(\sigma^2|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\sum_{i=1}^n(x_i-\mu_1)^2}{2\sigma^2}$, kde $\mu_1 = 4, \sigma\in(0.5,1.5)$;
$l((\mu,\sigma^2)|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\sum_{i=1}^n(x_i-\mu)^2}{2\sigma^2}$, kde $\mu\in(2,6)$ a $\sigma\in(0.5,1.5)$.
Výsledky simulácie: $\widehat{\mu} = 4.019708$ a $\widehat{\sigma}^2 = 1.000038$.

[15] Ak náhodná premenná X nebude mať normálne rozdelenie, funkcia vierohodnosti pre strednú hodnotu nemusí mať symetrický parabolický tvar okolo strednej hodnoty. Odhad strednej hodnoty môže byť potom vychýlený.



Obr. 15: Funkcia vierohodnosti pre μ a σ^2 ($X \sim N(4,1)$); odhad strednej hodnoty (aritmetický priemer) a odhad rozptylu sú označené zvislou čiarkovanou čiarou a v nich má funkcia vierohodnosti maximum

2 Charakteristiky polohy a variability

Príklad 45 (argument minima; DÚ) *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) sumu 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* $x_{(j)}$.

Príklad 46 (vygenerované pásy normality; cvič., DÚ) *Na základe vygenerovaných pseudonáhodných čísel* $X \sim N(\mu, \sigma^2)$, $n = 50, \mu = 0, \sigma^2 = 1$, *kde* $M = 1000$ *odhadnite (a) hustotu i-tej realizácie pomocou funkcie* `density()` *(ponechajte argument* `n=512` *a nastavte* `from=-3` *a* `to=3`*), (b) distribučnú funkciu i-tej realizácie pomocou (a) a funkcie* `cumsum()` *a (c) empirické kvantily i-tej realizácie pomocou funkcie* `qqnorm()`. *Vygenerované čísla* $x_{ij}, i = 1, 2, \ldots, 1000$ *a* $j = 1, 2, \ldots, 50$ *uložte do matice* $\mathbf{X}$, *ktorá bude mať rozmery* 1000×50. *Odhadnuté hustoty a distribučné funkcie uložte do matíc* $\mathbf{H}$ *a* $\mathbf{D}$, *ktoré bude mať rozmery* 1000×512 *a empirické kvantily do matice* $\mathbf{K}$, *ktorá bude mať rozmery* 1000×50. *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)* $X \sim N(0,1)$, $n = 50$, *(2)* $X \sim pN(0,1) + (1-p)N(0,4)$, $n = 50$ *a* $p = 0.95$ *a (3)* $X \sim pN(0,1) + (1-p)N(0,4)$, $n = 50$ *a* $p = 0.9$. *Zobrazte separátne (1), (2) a (3) do grafov (a), (b) a (c). Okomentujte.*

3 Testovanie hypotéz

Príklad 47 (MC experiment pre IS; cvič., DÚ) *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 $100(1-\alpha)\%$ IS pre μ. Zistite, koľko IS obsahuje strednú hodnotu $\mu = 0$. Toto číslo podelené M predstavuje simulovanú hladinu významnosti α. Okomentujte.*

Príklad 48 (tri typy testovacích štatistík; DÚ) *Predpokladajme, že $X \sim N(\mu, \sigma^2)$, kde σ^2 je známa. Testujme $H_0 : \theta = \theta_0$ oproti $H_1 : \theta \neq \theta_0$, kde $\theta = \mu$. Ukážte, že všetky tri testovacie štatistiky sú rovnaké, t.j. $U_{LR} = U_W = U_S$.*

1. $U_{LR} = n\frac{(\overline{X}-\mu_0)^2}{\sigma^2}$;
2. $U_W = n\frac{(\overline{X}-\mu_0)^2}{\sigma^2}$;
3. $U_S = n\frac{(\overline{X}-\mu_0)^2}{\sigma^2}$.

Príklad 49 (pred., DÚ) *Predpokladajme, že $X \sim N(\mu, \sigma^2)$, kde σ^2 je známa. Nech $\theta = \mu$. Testujme tri typy hypotéz*

1. $H_{01} : \mu = \mu_0$ *oproti* $H_{11} : \mu \neq \mu_0$;
2. $H_{02} : \mu \leq \mu_0$ *oproti* $H_{12} : \mu > \mu_0$;
3. $H_{03} : \mu \geq \mu_0$ *oproti* $H_{13} : \mu < \mu_0$;

(a) Vypočítajte pravdepodobnosti chýb druhého druhu $\Pr_\mu(CHDD)$ pri danej alternatíve pre všetky tri typy hypotéz, t.j. β_{11}, β_{12} a β_{13}.
(b) Vypočítajte a zobrazte silofunkcie pre všetky tri typy hypotéz, t.j. $1-\beta_{11}(\mu)$, $1-\beta_{12}(\mu)$ a $1-\beta_{13}(\mu)$. Pre zobracovanie si zvoľte $\mu_0 = 0$, $\mu \in (-10, 10)$, $\sigma = 6.4, \alpha = 0.05$ a $n = 10, 20, 30, 40$ a 50 (jeden obrázok pre každú z hypotéz (1), (2) a (3)).

Načrtnutému zodpovedá nasledovná situácia

$$\begin{array}{llll}
H_0 & H_1 & \mathcal{W} & \beta(\mu) \\
\mu = \mu_0 & \mu \neq \mu_0 & \mathcal{W}_1 = \left\{Z_W; |Z_W| \geq u_{\alpha/2}\right\} & \Phi\left(u_{\alpha/2} - \frac{|\mu_0-\mu|}{\sigma}\sqrt{n}\right) \\
\mu \leq \mu_0 & \mu > \mu_0 & \mathcal{W}_2 = \{Z_W; Z_W \geq u_\alpha\} & \Phi\left(u_\alpha + \frac{\mu_0-\mu}{\sigma}\sqrt{n}\right) \\
\mu \geq \mu_0 & \mu < \mu_0 & \mathcal{W}_3 = \{Z_W; Z_W \leq -u_\alpha\} & \Phi\left(u_\alpha - \frac{\mu_0-\mu}{\sigma}\sqrt{n}\right)
\end{array}$$

$\beta(\mu)$ v tabuľke pre H_{01} oproti H_{11} je približná (často sa používa v praxi namiesto presnej $\beta(\mu)$), jej zodpovedajúca presná silofunkcia je definovaná ako:

$$\begin{aligned}
1 - \beta(\mu) &\leq \Phi\left(u_{1-\alpha/2} + \frac{\mu - \mu_0}{\sigma}\sqrt{n}\right) + \Phi\left(u_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma}\sqrt{n}\right) \\
&= \Phi\left(u_{1-\alpha/2} - \frac{\mu_0 - \mu}{\sigma}\sqrt{n}\right) + \Phi\left(u_{1-\alpha/2} + \frac{\mu_0 - \mu}{\sigma}\sqrt{n}\right)
\end{aligned}$$

Potom

$$n \geq \left(\frac{u_{\alpha/2} + u_\beta}{c}\right)^2 = \left(\frac{u_{\alpha/2} + u_\beta}{\frac{\mu-\mu_0}{\sigma}}\right)^2 = \left(\frac{u_{\alpha/2} + u_\beta}{\mu - \mu_0}\right)^2 \sigma^2.$$

Nech Z je **nejaká testovacia štatistika** a z_W je jej realizácia (**pozorovaná, vypočítaná testovacia štatistika**), potom p-hodnotu počítame nasledovne:

$$\text{p-hodnota} = \begin{cases} 2\Pr(Z \geq |z_W||H_0), \text{ ak} & H_1 : \mu \neq \mu_0 \\ \Pr(Z \geq z_W|H_0), \text{ ak} & H_1 : \mu > \mu_0 \\ \Pr(Z \leq z_W|H_0), \text{ ak} & H_1 : \mu < \mu_0 \end{cases}.$$

$100(1-\alpha)\%$ empirické IS pre všetky tri typy hypotéz majú nasledovný tvar:

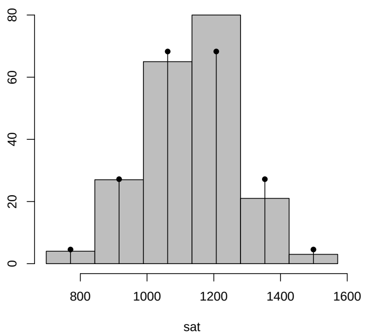
$$\begin{array}{lll} H_0 & H_1 & \text{hranice } (d,h) \text{ pre } 100(1-\alpha)\% \text{ empirický IS} \\ \mu = \mu_0 & \mu \neq \mu_0 & \mathcal{CS}_{1-\alpha} = \left\{\mu_0 : \mu_0 \in \left(\overline{x} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right\} \\ \mu \leq \mu_0 & \mu > \mu_0 & \mathcal{CS}_{1-\alpha} = \left\{\mu_0 : \mu_0 \in \left(\overline{x} - u_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right)\right\} \\ \mu \geq \mu_0 & \mu < \mu_0 & \mathcal{CS}_{1-\alpha} = \left\{\mu_0 : \mu_0 \in \left(-\infty, \overline{x} + u_{\alpha}\frac{\sigma}{\sqrt{n}}\right)\right\} \end{array}$$

Príklad 50 (jednovýberový Z-test strednej hodnoty μ; DÚ) *Porovnajte presnú a približnú silofunkciu pre test $H_0 : \mu = \mu_0$ (oproti $H_1 : \mu \neq \mu_0$, ak σ^2 je známe) nakreslením oboch do jedného obrázka pre $n = 20, \mu_0 = 0$ a $\sigma = 1$. Okomentujte.*

Príklad 51 (vylepšená vierohodnosť pomocou $g(\theta)$; DÚ) *Nakreslite (a) logaritmus funkciu vierohodnosti parametra p binomického rozdelenia $Bin(N,p)$, kde $N = 10$ a $n = 8$, superponovaný jeho kvadratickou aproximáciou. Nakreslite (b) logaritmus funkciu vierohodnosti $g(p) = logit(p) = \ln\frac{p}{1-p}$ (pri rovnakom zadaní N a n ako v (a)) superponovaný jeho kvadratickou aproximáciou. Je funkcia vierohodnosti $g(p)$ regulárnejšia ako funkcia vierohodnosti pre p? (c) Vypočítajte Waldov a vierohodnostný $100 \times (1-\alpha)$ % empirický IS pre p. (d) Vypočítajte Waldov $100 \times (1-\alpha)$ % empirický IS pre $g(p)$ z (a) a (b) a transformujte ho späť do originálnej škály. (e) Ukážte, že vierohodnostný IS pre p v škále p je identický s vierohodnostným IS v škále $g(p)$ z (a) a (b) po jeho spätnej transformácii do originálnej škály. Okomentujte.*

3.1 Testy dobrej zhody

Príklad 52 (χ^2-test dobrej zhody) *Majme dáta* `Grades`*, ktoré reprezentujú SAT skóre ($n = 200$) náhodne vybranej vzorky študentov z jednej univerzity v USA. Otestujte na hladine významnosti $\alpha = 0.05$, či majú dáta normálne rozdelenie. Použite intervaly $\langle \mu - 3\sigma, \mu - 2\sigma \rangle$, $(\mu - 2\sigma, \mu - \sigma \rangle$, $(\mu - \sigma, \mu \rangle$, $(\mu, \mu + \sigma \rangle$, $(\mu + \sigma, \mu + 2\sigma \rangle$ a $(\mu + 2\sigma, \mu + 3\sigma \rangle$. Nakreslite histogram použitím vyššie spomenutých intervalov a superponujte ho s očakávanými hodnotami SAT skóre v každej kategórii, keď $F_0(x) \sim N(\mu, \sigma^2)$.* cvič.



Obr. 16: Histogram superponovaný s očakávanými hodnotami SAT skóre

Príklad 53 (χ^2-test dobrej zhody; pokrač.) *Zopakujte výpočet z predchádzajúceho príkladu na intervaloch definovaných pomocou hraníc:*
(a) kvartilové hranice – $x_{min}, \widetilde{x}_{0.25}, \widetilde{x}_{0.50}, \widetilde{x}_{0.75}, x_{max}$;
(b) decilové hranice – $x_{min}, \widetilde{x}_{0.1}, \widetilde{x}_{0.2}, \ldots, \widetilde{x}_{0.8}, \widetilde{x}_{0.9}, x_{max}$.
Nakreslite histogram použitím vyššie spomenutých intervalov a superponujte ho s očakávanými hodnotami SAT skóre v každej kategórii, keď $F_0(x) \sim N(\mu, \sigma^2)$. Porovnajte výsledky s výsledkami predchádzajúceho príkladu. DÚ

Príklad 54 (χ^2-test dobrej zhody) *Johann Gregor Mendel vo svojich pokusoch s krížením rastlín hrachu (Pisum sativum) študoval dedičnosť siedmych rôznych znakov. V každom z pokusov, pri sledovaní jedného znaku, získal po krížení dvoch čistých línií (t.j. dominantného homozygota AA s recesívnym homozygotom aa) generáciu, v ktorej mali všetky rastliny rovnaký fenotyp (t.j. heterozygoti Aa). Po ich samooplodnení (čo je prirodzený spôsob rozmnožovania hrachu) získal ďalšiu generáciu, v ktorej sa vyskytovali sledované znaky v dvoch formách, a to zakaždým v pomere veľmi blízkom 3:1. Jedným zo znakov, ktoré študoval, bola farba semien. Po krížení 258 hybridov získal celkove 8023 semien, z ktorých 6022 bolo žltých a 2001 zelených (Matalová 2008). Otestujte platnosť fenotypového štiepneho pomeru $3 : 1$ na hladine významnosti $\alpha = 0.05$.* pred

Príklad 55 (χ^2-test dobrej zhody) *χ^2-test dobrej zhody:* DÚ
(a) Otestujte zhodu početností úmrtí X ako následok kopnutia koňom v Pruských armádnych jednotkách (pozri príklad 100) s Poissonovým rozdelením s parametrom λ, t.j. $X \sim Poiss(\lambda)$ na hladine významnosti $\alpha = 0.05$. Pozrite príklad 24 (DÚ v Štatistickej inferencii I).

n	0	1	2	3	4	5+
m_n	109	65	22	3	1	0
očakávané m_n	108.7	66.3	20.2	4.1	0.6	0.1

(b) Otestujte zhodu početností chlapcov X v rodinách s binomickým rozdelením s parametrami N a π, t.j. $X \sim Bin(N, \pi)$ na hladine významnosti $\alpha = 0.05$. Pozrite príklad 27.

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
očakávané m_n	1	12	72	258	628	1085	1367	1266	854	410	133	26	2

(c) Otestujte zhodu početností úrazov medzi robotníkmi X (pozri príklad 106 a 132) (1) s Poissonovým rozdelením s parametrom λ, t.j. $X \sim Poiss(\lambda)$ a (2) s negatívne binomickým rozdelním s parametrami α a π, t.j. $Negbinom(\alpha, \pi)$ na hladine významnosti $\alpha = 0.05$. Pozrite príklad 28 a 40.

n	0	1	2	3	≥ 4
m_n	447	132	42	21	5
(1) očakávané m_n	406	189	44	7	1
(2) očakávané m_n	446	134	44	15	8

Príklad 56 (Kolmogorov-Smirnovov test dobrej zhody) *Majme výšky $n = 12$ náhodne vybraných 10-ročných dievčat $\mathbf{x} = (131, 132, 135, 141, 141, 141, 141, 142, 143, 146, 146, 151)^T$. Otestujte na hladine významnosti $\alpha = 0.05$, či majú dáta normálne rozdelenie, kde $F_0(x) \sim N(\mu, \sigma^2)$.* cvič.

Pozn.: Funkciu `ecdf()` (použitie v podobe `Fn <- ecfd(vyska); FnX <- Fn(vyska)`) nie je možné použiť, pretože pri zhodách je posunutá $\widehat{F}_n(x_{i-1})$ vypočítaná z $\widehat{F}_n(x_i)$ nesprávna.

Ak je hypotéza zložená, Kolmogorov-Smirnovov test je veľmi konzervatívny. Avšak D_n môžeme použiť na výpočet, ak odhadneme parametre príslušného rozdelenia, kde $\widehat{F}_0(x)$ substituujeme za $F_0(x)$. Potom však nastávajú problémy s rozdelením D_n. Problém rieši modifikácia Kolmogorovho-Smirnovovho testu, kedy sa tento test nazýva **Lillieforsov test normality**, použitím MC simulácií, kde kritické hodnoty označíme $D_n^{(l)}(\alpha)$. Pri testovaní sa často používa **Dallal-Wilkinsonova aproximácia p-hodnoty** v podobe

$$\widehat{\text{p-hodnota}} = \exp(-7.01256 D_n^2 (n+2.78019) + 2.99587 D_n \sqrt{n + 2.78019} - 0.122119 + \frac{0.974598}{\sqrt{n}} + \frac{1.67997}{n})$$

pre $n \in (5, 100)$ a p-hodnotu $\leq$ 0.1. Ak je $n \geq 100$, potom D_n vo vyššie uvedenom vzorci nahradíme $D_m = D_n (\frac{m}{100})^{0.49}$, kde m je skutočný rozsah a n substituujeme číslom 100. Ak p-hodnota > 0.1, potom D_n nahradíme $D_{\text{mod}} = D_n(\sqrt{n} - 0.01 + 0.85\sqrt{n})$. Podľa veľkosti D_{mod} vypočítame p-hodnotu nasledovne:

- ak $D_{\text{mod}} \leq 0.302$, $\widehat{\text{p-hodnota}} = 1$,
- ak $D_{\text{mod}} \leq 0.5$,
$$\widehat{\text{p-hodnota}} = 2.76773 - 19.828315 D_{\text{mod}} + 80.709644 D_{\text{mod}}^2 - 138.55152 D_{\text{mod}}^3 + 81.218052 D_{\text{mod}}^4,$$
- ak $D_{\text{mod}} \leq 0.9$,
$$\widehat{\text{p-hodnota}} = -4.901232 + 40.662806 D_{\text{mod}} - 97.490286 D_{\text{mod}}^2 + 94.029866 D_{\text{mod}}^3 - 32.355711 D_{\text{mod}}^4,$$
- ak $D_{\text{mod}} \leq 1.31$,
$$\widehat{\text{p-hodnota}} = 6.198765 - 19.558097 D_{\text{mod}} + 23.186922 D_{\text{mod}}^2 - 12.234627 D_{\text{mod}}^3 + 2.423045 D_{\text{mod}}^4.$$

3.2 Asymptotické testy o jednom parametri

Príklad 57 (minimálny rozsah N) *Vypočítajte minimálny rozsah n pre $p = 0.1, 0.2, \ldots, 0.9$, $p_0 = 0$ pri $\alpha = 0.05$, $\beta = 0.8$ a obojstrannej alternatíve H_{11}. Skontrolujte, či je splnená Haldova podmienka. Ak nie je, doplňte minimálne N, ktoré túto podmienku spĺňa.* cvič.

Riešenie:

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
N	71	32	19	12	8	6	4	2	1
$Np(1-p)$	6.39	5.12	3.99	2.88	2.00	1.44	0.84	0.32	0.09
$9/(p(1-p))$	**100**	**57**	**43**	**38**	**36**	**38**	**43**	**57**	**100**

Príklad 58 (minimálny rozsah N) *Vypočítajte minimálny rozsah N pre $p = 0.1, 0.2, \ldots, 0.9$, p_0 vždy o 0.1 menšie ako p, pri $\alpha = 0.05$, $\beta = 0.8$ a obojstrannej alternatíve H_{11}. Skontrolujte, či je splnená Haldova podmienka. Ak nie je, doplňte minimálne N, ktoré túto podmienku spĺňa.* cvič.

Riešenie:

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
p_0	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$p(1-p)$	0.09	0.16	0.21	0.24	0.25	0.24	0.21	0.16	0.09
N	71	**126**	**165**	**189**	**197**	**189**	**165**	**126**	71
$Np(1-p)$	6.39	20.16	34.65	45.36	49.25	45.36	34.65	20.16	20.16
$9/(p(1-p))$	**100**	57	43	38	36	38	43	57	**100**

Príklad 59 (pravdepodobnost pokrytia) *Nech $X \sim Bin(N, p)$, kde $N = 30$ a $p = 0.8$ a pravdepodobnosť úspechu $\widehat{p} = \frac{24}{30} = 0.8$, kde $x = 24$ a $N = 30$. Waldov 95% empirický DIS pre p je rovný $(d, h) = (0.657, 0.943)$. Vypočítajte pravdepodobnosť pokrytia tohoto intervalu. Pozn.: pravdepodobnosť pokrytia Waldovho 95% DIS pre p vypočítame nasledovne*

$$\Pr(pokrytie) = \textstyle\sum_j \Pr(X = Np_j : p \in \textit{Waldov 95\% DIS pre } p_j),$$

kde $p_j \in \mathcal{M}_J = \left\{\frac{1}{30}, \frac{2}{30}, \ldots, 1 - \frac{1}{30}\right\}$, t.j. ide o súčet takých funkčných hodnôt pravdepodobnostnej funkcie v bodoch Np_j, kde $p \in$ Waldovmu 95% DIS pre p_j. Výsledky usporiadajte do tabuľky, ktorej stĺpce budú x_j, p_j, d_j (dolná hranica Waldovho 95% DIS pre p_j), h_j (horná hranica Waldovho 95% DIS pre p_j), $\Pr(pokrytie)$ a pokrytie (indikácia toho, či p patrí alebo nepatrí Waldovmu 95% DIS pre p_j). cvič.

Príklad 60 (pravdepodobnost pokrytia) *Nech $X_i \sim Bin(N, p_i)$. Vypočítajte pravdepodobnosti pokrytia Waldovho 95% DIS pre každé p_i, kde p_i patria množine $\mathcal{M}_I = \left\langle\frac{1}{N}, 1 - \frac{1}{N}\right\rangle$, sú ekvidištantne vzdialené medzi $\frac{1}{N}$ a $1 - \frac{1}{N}$ a ich počet $M = 5000$. Nakreslite obrázok, kde na x-ovej osi budú p_i a na y-ovej osi pravdepodobnosť pokrytia $\Pr_i(pokrytie)$. Zvoľte (a) $N = 30$, (b) $N = 100$ a (c) $N = 1000$. Pozn.: pravdepodobnosti pokrytia Waldovho 95% DIS pre p_i vypočítame nasledovne*

$$\Pr_i(pokrytie) = \textstyle\sum_j \Pr(X = Np_j : p_i \in \textit{Waldov 95\% DIS pre } p_j),$$

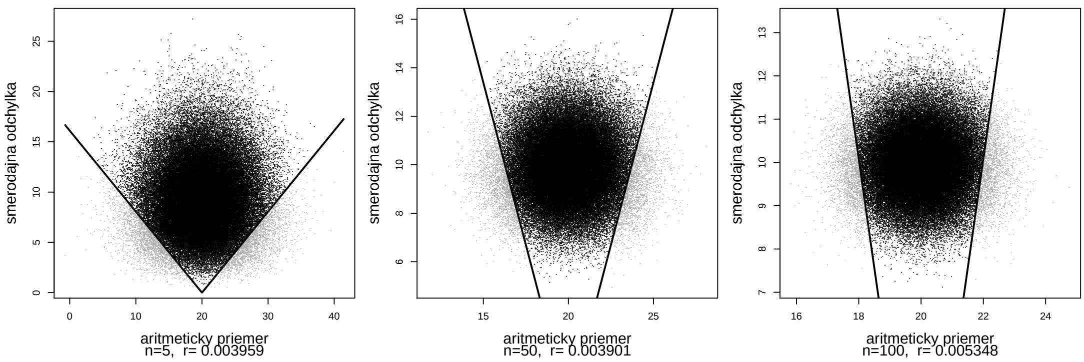
kde $p_j \in \mathcal{M}_J = \left\{\frac{1}{N}, \frac{2}{N}, \ldots, 1 - \frac{1}{N}\right\}$, t.j. ide o súčet takých funkčných hodnôt pravdepodobnostnej funkcie v bodoch Np_j, kde $p_i \in$ Waldovmu 95% DIS pre p_j. cvič.

Príklad 61 (pravdepodobnost pokrytia) *Nech* $X_i \sim Bin(N, p_i)$. *Vypočítajte pravdepodobnosti pokrytia:*
(a) vierohodnostného 95% *DIS,*
(b) skóre 95% *DIS,*
(c) spätne tranformovaného Waldovho 95% *DIS pre* $g(p_i)$ *s hranicami* $(d_g^{(i)}, h_g^{(i)})$ *na Waldov* 95% *DIS pre* p_i *s hranicami* $\left((g(d_g^{(i)}))^{-1}, (g(h_g^{(i)}))^{-1}\right)$, *kde (1)* $g(p_i) = \frac{p_i}{1-p_i}$, *(2)* $g(p_i) = \ln \frac{p_i}{1-p_i}$ *a (3)* $g(p_i) = \arcsin(\sqrt{p_i})$
pre každé p_i, *kde* p_i *patria množine* $\mathcal{M}_I = \left\langle \frac{1}{N}, 1 - \frac{1}{N} \right\rangle$, *sú ekvidištantne vzdialené medzi* $\frac{1}{N}$ *a* $1 - \frac{1}{N}$ *a ich počet* $M = 5000$. *Nakreslite obrázok, kde na x-ovej osi budú* p_i *a na y-ovej osi pravdepodobnosť pokrytia* $\Pr_i(pokrytie)$. *Zvoľte (a)* $N = 30$, *(b)* $N = 100$ *a (c)* $N = 1000$. Pozn.: *pravdepodobnosti pokrytia* 95% *DIS pre* p_i *vypočítame nasledovne*

$$\Pr_i(pokrytie) = \sum_j \Pr(X = Np_j : p_i \in 95\% \text{ } DIS \text{ } pre \text{ } p_j),$$

kde $p_j \in \mathcal{M}_J = \left\{ \frac{1}{N}, \frac{2}{N}, \ldots, 1 - \frac{1}{N} \right\}$, *t.j. ide o súčet takých funkčných hodnôt pravdepodobnostnej funkcie v bodoch* Np_j, *kde* $p_i \in 95\%$ *DIS pre* p_j. *Pre tie DIS, ktoré majú pre* $p = 0$ *a* $p = 1$ *nenulovú šírku, môžeme použiť* $\mathcal{M}_I = \left\langle \frac{0}{N}, \frac{N}{N} \right\rangle$. DÚ

Príklad 62 (nezávislosť μ a σ^2; pravdepodobnosť pokrytia) *Nech* $X \sim N(\mu, \sigma^2)$, *kde* $\mu = 20$ *a* $\sigma^2 = 100$. *Vypočítajte Pearsonov korelačný koeficient* $r_{\overline{X},S}$ *pomocou simulačnej štúdie (*$M = 100000$*). Nakreslite rozptylový graf* $(\overline{x}_i, s_i)$, *kde* $i = 1, 2, \ldots, M$ *(sivou farbou). Dokreslite do grafu také body, pre ktoré platí* $t_{W,i} = \left|\frac{\overline{x}_i - \mu}{s_i}\sqrt{n}\right| < t_{n-1}(\alpha/2)$ *(čiernou farbou) ako aj hranice, ktoré definujú také body* $(\overline{x}_i, s_i)$, *pre ktoré* $t_{W,i} = t_{n-1}(\alpha/2)$. *Vypočítajte pravdepodobnosť pokrytia* 95% *DIS pre* μ *ako podiel* $\sum_i I\left(t_{W,i} < t_{n-1}(\alpha/2)\right)/M$. *Zvoľte (a)* $n = 5$, *(b)* $n = 50$ *a (c)* $n = 100$. cvič.



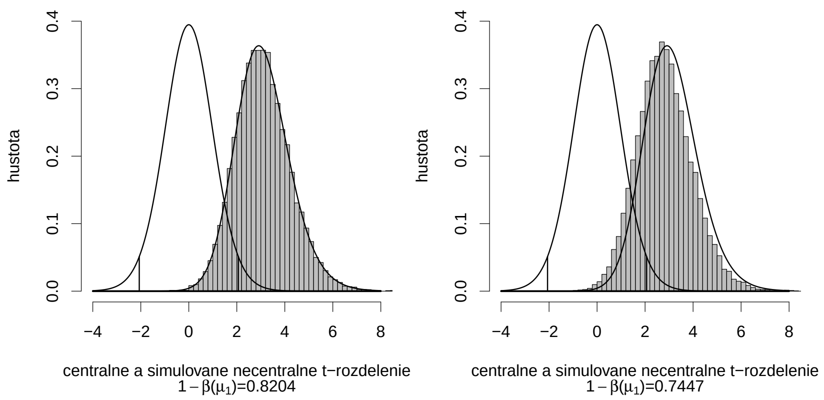
Obr. 17: Rozptylový graf $\overline{x}_i, s_i$, $i = 1, 2, \ldots, M, M = 100000$ pre $n = 5$ (vľavo), $n = 50$ (v strede) a $n = 100$ (vpravo)

Príklad 63 (nezávislosť μ a σ^2; pravdepodobnosť pokrytia) *Nech* $X \sim [pN(\mu, \sigma_1^2) + (1-p)N(\mu, \sigma_2^2)]$, *kde* $p = 0.9$, $\mu = 20$, $\sigma_1^2 = 100$ *a* $\sigma_2^2 = 400$. *Vypočítajte Pearsonov korelačný koeficient* $r_{\overline{X},S}$ *pomocou simulačnej štúdie (*$M = 100000$*). Nakreslite rozptylový graf* $(\overline{x}_i, s_i)$, *kde* $i = 1, 2, \ldots, M$ *(sivou farbou). Dokreslite do grafu také body, pre ktoré platí* $t_{W,i} = \left|\frac{\overline{x}_i - \mu}{s_i}\sqrt{n}\right| < t_{n-1}(\alpha/2)$ *(čiernou farbou) ako aj hranice, ktoré definujú také body* $(\overline{x}_i, s_i)$, *pre ktoré* $t_{W,i} = t_{n-1}(\alpha/2)$. *Vypočítajte pravdepodobnosť pokrytia* 95% *DIS pre* μ *ako podiel* $\sum_i I\left(t_{W,i} < t_{n-1}(\alpha/2)\right)/M$. *Zvoľte (a)* $n = 5$, *(b)* $n = 50$ *a (c)* $n = 100$. DÚ

Príklad 64 (necentrálne t-rozdelenie) *Nakreslite distribučnú funkciu necentrálneho t-rozdelenia $t_{n-1,\lambda}$, kde $\delta = \mu - \mu_0$ a $\lambda = \delta/(\sigma/\sqrt{n})$. Použite $\mu_0 = 0$, $\delta = 1$, $\sigma = 1.4$ a $n = 26$. Vypočítajte pravdepodobnosť nad kvantilom $x_{0.975}$ pod krivkou hustoty tohoto rozdelenia.* cvič.

Príklad 65 (necentrálne t-rozdelenie) *Nakreslite hustoty jedného centrálneho a štyroch necentrálnych t-rozdelení $t_{n-1,\lambda}$ ($\delta = \mu - \mu_0$ a $\lambda = \delta/(\sigma/\sqrt{n})$) do jedneho obrázka tak, aby boli odlíšiteľné farbou alebo typom čiary. Použite $\mu_0 = 0$, $\delta = 0, 0.5, 0.8, 1$ a 1.2, $\sigma = 1.4$ a $n = 26$.* cvič.

Príklad 66 (sila a silofunkcia) *Použite R na simuláciu hustoty rozdelenia $t_{n-1,\lambda}$ testovacích štatistík $T^{(m)}_{W,\lambda} = \frac{\bar{x}_m - \mu_0}{s_m}\sqrt{n}$ (necentrálne t-rozdelenie s $n-1$ stupňami voľnosti a parametrom necentrality λ), kde $n = 25$, $\lambda = 3$, $m = 1, 2, \ldots, M$, pri $M = 20000$ opakovaniach. Na základe tohoto rozdelenia výpočítajte silu testu pre $\mu_0 = 2.5$ a $\mu_1 = 4$ (pozri obrázok 18). (1) $X \sim N(4, 2.5^2)$ a (2) $X \sim [pN(4, 2.5^2) + (1-p)N(4, 4.5^2)$, kde $p = 0.9$.* DÚ



Obr. 18: Hustota centrálneho a necentrálneho t-rozdelenia; vľavo – hustota necentrálneho t-rozdelenia je superponovaná histogramom simulácií $X \sim N(4, 2.5^2)$ a vpravo – $X \sim [pN(4, 2.5^2) + (1 - p)N(4, 4.5^2)$, kde $p = 0.9$

Pravdepodobnosť empirickej CHPD pre MC experiment je pravdepodobnosť p signifikantných testovacích štatistík medzi ich M opakovaniami, ak H_0 platí. Potom $SE(p) = \sqrt{p(1-p)/M}$ je menšia alebo rovná $0.5/\sqrt{M}$.

Príklad 67 (pravdepodobnosť empirickej CHPD t-testu) *Nech $X \sim N(\mu, \sigma^2)$, kde $\mu = 500$ a $\sigma^2 = 100$. Testujte $H_0 : \mu = 500$ oproti $H_1 : \mu > 500$, ak $\alpha = 0.05$, σ je neznáme. Použite R na simuláciu empirickej $\Pr(CHPD)$, kde počet simulácií je $M = 10000$ a rozsah náhodného výberu je $n = 20$ pre jednovýberový Studentov t-test o strednej hodnote μ. Použite funkciu `t.test(x, alternative = "greater", mu = mu0)` a pre každú testovaciu štatistiku $t_m, m = 1, 2, \ldots, M$ vypočítajte p-hodnotu a jej štandardnú chybu za platnosti H_0. Ide o zistenie relatívnej početnosti p zamietnutých H_0 na hladine významnosti $\alpha = 0.05$ medzi M testami, kde $p = \Pr(CHPD) = \frac{\sum_{i=1}^{M} I(H_0 \text{ zamietame})}{M}$.* cvič.

Príklad 68 (vierohodnostný DIS pre μ) *Majme dáta `one-sample-mean-skull.txt` a premennú dĺžka lebky `skull.L` v mm starovekej egyptskej mužskej populácie, o ktorej predpokladáme, že má normálne rozdelenie $N(\mu, \sigma^2)$. Vypočítajte vierohodnostný 95% empirický DIS pre strednú hodnotu dĺžky lebky μ pomocou 15% cut-off relatívnej (štandardizovanej) funkcie vierohodnosti $\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = L(\boldsymbol{\theta}|\mathbf{x})/L(\widehat{\boldsymbol{\theta}}|\mathbf{x})$ a porovnajte ho s vierohodnostným 95% empirický DIS pre μ.* DÚ cvič.

Príklad 69 (MC odhad koeficientu spoľahlivosti $1-\alpha$) *Vypočítajte v R MC odhad koeficientu spoľahlivosti (pravdepodobnosti pokrytia) pre pravostranný (horný) 95% JIS pre σ^2 pri $M = 1000$ a $n = 20$. Tento JIS je ekvivalentný s testom H_{02} oproti H_{12}. (a) Nech $X \sim N(0,4)$, (b) $X \sim \chi^2(2)$ a (c) $X \sim [pN(0,4) + (1-p)N(0,9)$, kde $p = 0.9$.* cvič.

Príklad 70 (minimálny rozsah súboru) *Vypočítajte v R minimálny rozsah náhodného výberu pre test $H_{03}: \sigma^2 \geq \sigma_0^2$ oproti $H_{13}: \sigma^2 < \sigma_0^2$ pri $\alpha = 0.05$ a $1-\beta = 0.8$, ak podiel $\frac{\sigma_0^2}{\sigma^2}$ je rovný (a) 1.1, (b) 1.5 a (c) 5.* DÚ

Príklad 71 (konvergencia ρ a ξ k normálnemu rozdeleniu) *Urobte v R simuláciu pseudonáhodných čísel z $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde $\mu_1 = 0, \mu_2 = 0$, $\sigma_1 = 1, \sigma_2 = 1$ (pozri príklad 18), kde $n = 5, 10, 20, 50$ a 100, $M = 10000$. Použite (a) $\rho = 0$, (b) $\rho = 0.50$ a (c) $\rho = 0.9$. Pre každé $m = 1, 2, \dots, M$ vypočítajte Pearsonov korelačný koeficient r_m a Fisherovu Z-premennú $z_{R,m}$. Zobrazte histogramy simulovaných r_m a $z_{R,m}$ a superponujte ich teoretickými hustotami prislúchajúcich normálnych rozdelení.* cvič.

Príklad 72 (porovnanie troch DIS v extrémnej situácii) *Nech $N = 25$ študentov, ktorým sme položili otázku, či sú vegetariáni. Z nich $x = 0$ odpovedalo „áno“. Vypočítajte empirické $100\times(1-\alpha)$% DIS (a) Waldov DIS, (b) skóre DIS a (c) vierohodnostný DIS pre p ($1-\alpha = 0.95$).* cvič.

Príklad 73 (funkcia vierohodnosti v extrémnej situácii) *Nakreslite funkciu vierohodnosti pre situáciu v predchádzajúcom príklade.* DÚ

Príklad 74 (pruské armádne jednotky) *Majme $X \sim Poiss(\lambda)$. Vypočítajte (a) Waldov 95% DIS pre λ, (b) skóre 95% DIS pre λ a (c) vierohodnostný 95% DIS pre λ (dáta pozri v príklade 55).* cvič.

Príklad 75 (test o rozdiele stredných hodnotôt μ_1 a μ_2) *Majme dáta `two-samples-means-birth.txt`, premennú pôrodná hmotnosť `birth.W` v gramoch novorodencov (chlapcov) narodených v krajskej nemocnici v priebehu jedného roka a premennú počet starších súrodencov `o.sib.N`, ktorá nadobúda hodnoty 0 (žiadny) a 1 (jeden). Predpokladáme, že premenná `birth.W` chlapcov so žiadnym starším súrodencom má normálne rozdelenie $N(\mu_1, \sigma_1^2)$ a `birth.W` chlapcov s jedným starším súroden com má normálne rozdelenie $N(\mu_2, \sigma_2^2)$. (a) Otestujte hypotézu o zhode stredných hodnôt μ_1 a μ_2 na hladine významnosti $\alpha = 0.05$. (b) Vypočítajte $100 \times (1-\alpha)$% empirický DIS pre rozdiel stredných hodnôt $\mu_1 - \mu_2$, kde koeficient spoľahlivosti $1-\alpha = 0.95$. Použite (1) Waldovu testovaciu štatistiku T_W pri predpoklade (1.1) rovnosti a (1.2) nerovnosti rozptylov, (2) testovaciu štatistiku pomerom vierohodnosti U_{LR} pri predpoklade (2.1) rovnosti a (2.2) nerovnosti rozptylov a DIS prislúchajúce (1.1), (1.2), (2.1) a (2.2).* cvič. DÚ

Zhrnutie kapitoly o testoch o dvoch pravdepodobnostiach

$$\text{p-hodnota} = \begin{cases} 2\Pr(Z_W \geq |z_W||H_0), \text{ ak} & H_1: \ln \mathrm{RR} \neq \ln \mathrm{RR}_0 \\ \Pr(Z_W \geq z_W|H_0), \text{ ak} & H_1: \ln \mathrm{RR} > \ln \mathrm{RR}_0 \\ \Pr(Z_W \leq z_W|H_0), \text{ ak} & H_1: \ln \mathrm{RR} < \ln \mathrm{RR}_0 \end{cases}$$

H_0	H_1	hranice (d,h) pre $100(1-\alpha)$% empirický IS
$\ln \mathrm{RR} = \ln \mathrm{RR}_0$	$\ln \mathrm{RR} \neq \ln \mathrm{RR}_0$	$\mathcal{CS}_{1-\alpha} = \left\{\ln \mathrm{RR}_0 : \ln \mathrm{RR}_0 \in \left(\ln \widehat{\mathrm{RR}} - u_{\alpha/2}s_g, \widehat{\mathrm{RR}} + u_{\alpha/2}s_g\right)\right\}$
$\ln \mathrm{RR} \leq \ln \mathrm{RR}_0$	$\ln \mathrm{RR} > \ln \mathrm{RR}_0$	$\mathcal{CS}_{1-\alpha} = \left\{\ln \mathrm{RR}_0 : \ln \mathrm{RR}_0 \in \left(\ln \widehat{\mathrm{RR}} - u_{\alpha}s_g, \infty\right)\right\}$
$\ln \mathrm{RR} \geq \ln \mathrm{RR}_0$	$\ln \mathrm{RR} < \ln \mathrm{RR}_0$	$\mathcal{CS}_{1-\alpha} = \left\{\ln \mathrm{RR}_0 : \ln \mathrm{RR}_0 \in \left(-\infty, \ln \widehat{\mathrm{RR}} + u_{\alpha}s_g\right)\right\}$

$$\text{p-hodnota} = \begin{cases} 2\Pr(Z_W \geq |z_W||H_0), \text{ ak} & H_1 : \mathrm{RR} \neq \mathrm{RR}_0 \\ \Pr(Z_W \geq z_W|H_0), \text{ ak} & H_1 : \mathrm{RR} > \mathrm{RR}_0 \\ \Pr(Z_W \leq z_W|H_0), \text{ ak} & H_1 : \mathrm{RR} < \mathrm{RR}_0 \end{cases}$$

$$\begin{array}{lll} H_0 & H_1 & \text{hranice } (d,h) \text{ pre } 100(1-\alpha)\% \text{ empirický IS} \\ \mathrm{RR} = \mathrm{RR}_0 & \mathrm{RR} \neq \mathrm{RR}_0 & \mathcal{CS}_{1-\alpha} = \left\{\mathrm{RR}_0 : \mathrm{RR}_0 \in \left(\widehat{\mathrm{RR}} - u_{\alpha/2}s_g, \ln\widehat{\mathrm{RR}} + u_{\alpha/2}s_g\right)\right\} \\ \mathrm{RR} \leq \mathrm{RR}_0 & \mathrm{RR} > \mathrm{RR}_0 & \mathcal{CS}_{1-\alpha} = \left\{\mathrm{RR}_0 : \mathrm{RR}_0 \in \left(\widehat{\mathrm{RR}} - u_{\alpha}s_g, \infty\right)\right\} \\ \mathrm{RR} \geq \mathrm{RR}_0 & \mathrm{RR} < \mathrm{RR}_0 & \mathcal{CS}_{1-\alpha} = \left\{\mathrm{RR}_0 : \mathrm{RR}_0 \in \left(0, \widehat{\mathrm{RR}} + u_{\alpha}s_g\right)\right\} \end{array}$$

$$\text{p-hodnota} = \begin{cases} 2\Pr(Z_W \geq |z_W||H_0), \text{ ak} & H_1 : \ln\mathrm{OR} \neq \ln\mathrm{OR}_0 \\ \Pr(Z_W \geq z_W|H_0), \text{ ak} & H_1 : \ln\mathrm{OR} > \ln\mathrm{OR}_0 \\ \Pr(Z_W \leq z_W|H_0), \text{ ak} & H_1 : \ln\mathrm{OR} < \ln\mathrm{OR}_0 \end{cases}$$

$$\begin{array}{lll} H_0 & H_1 & \text{hranice } (d,h) \text{ pre } 100(1-\alpha)\% \text{ empirický IS} \\ \ln\mathrm{OR} = \ln\mathrm{OR}_0 & \ln\mathrm{OR} \neq \ln\mathrm{OR}_0 & \mathcal{CS}_{1-\alpha} = \left\{\ln\mathrm{OR}_0 : \ln\mathrm{OR}_0 \in \left(\ln\widehat{\mathrm{OR}} - u_{\alpha/2}s_g, \ln\widehat{\mathrm{OR}} + u_{\alpha/2}s_g\right)\right\} \\ \ln\mathrm{OR} \leq \ln\mathrm{OR}_0 & \ln\mathrm{OR} > \ln\mathrm{OR}_0 & \mathcal{CS}_{1-\alpha} = \left\{\ln\mathrm{OR}_0 : \ln\mathrm{OR}_0 \in \left(\ln\widehat{\mathrm{OR}} - u_{\alpha}s_g, \infty\right)\right\} \\ \ln\mathrm{OR} \geq \ln\mathrm{OR}_0 & \ln\mathrm{OR} < \ln\mathrm{OR}_0 & \mathcal{CS}_{1-\alpha} = \left\{\ln\mathrm{OR}_0 : \ln\mathrm{OR}_0 \in \left(-\infty, \ln\widehat{\mathrm{OR}} + u_{\alpha}s_g\right)\right\} \end{array}$$

$$\text{p-hodnota} = \begin{cases} 2\Pr(Z_W \geq |z_W||H_0), \text{ ak} & H_1 : \mathrm{OR} \neq \mathrm{OR}_0 \\ \Pr(Z_W \geq z_W|H_0), \text{ ak} & H_1 : \mathrm{OR} > \mathrm{OR}_0 \\ \Pr(Z_W \leq z_W|H_0), \text{ ak} & H_1 : \mathrm{OR} < \mathrm{OR}_0 \end{cases}$$

$$\begin{array}{lll} H_0 & H_1 & \text{hranice } (d,h) \text{ pre } 100(1-\alpha)\% \text{ empirický IS} \\ \mathrm{OR} = \mathrm{OR}_0 & \mathrm{OR} \neq \mathrm{OR}_0 & \mathcal{CS}_{1-\alpha} = \left\{\mathrm{OR}_0 : \mathrm{OR}_0 \in \left(\widehat{\mathrm{OR}} - u_{\alpha/2}s_g, \widehat{\mathrm{OR}} + u_{\alpha/2}s_g\right)\right\} \\ \mathrm{OR} \leq \mathrm{OR}_0 & \mathrm{OR} > \mathrm{OR}_0 & \mathcal{CS}_{1-\alpha} = \left\{\mathrm{OR}_0 : \mathrm{OR}_0 \in \left(\widehat{\mathrm{OR}} - u_{\alpha}s_g, \infty\right)\right\} \\ \mathrm{OR} \geq \mathrm{OR}_0 & \mathrm{OR} < \mathrm{OR}_0 & \mathcal{CS}_{1-\alpha} = \left\{\mathrm{OR}_0 : \mathrm{OR}_0 \in \left(0, \widehat{\mathrm{OR}} + u_{\alpha}s_g\right)\right\} \end{array}$$

Príklad 76 (maximálne vierohodné odhady; nádor prsníka) *Majme početnosti subjektov X_1, ktoré majú rozšírené metastázy nádoru prsníka, kde $X_1 \sim Bin(N_1, p_1)$ a početnosti subjektov X_2, ktoré majú lokalizované metastázy nádoru prsníka, kde $X_2 \sim Bin(N_2, p_2)$.*
(1a) Aplikujte funkciu vierohodnosti $L(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2)$, kde $\boldsymbol{\theta} = (p_1, p_2)^T$ na dáta v tabuľke a vypočítajte $\widehat{\theta}$.
(1b) Nakreslite funkciu vierohodnosti ako funkciu p_1 a p_2 [superpozícia `contour()` a `image()`].
(2a) Aplikujte funkciu vierohodnosti $L(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2)$, kde $\boldsymbol{\theta} = (\theta, \eta)^T$, logaritmus pomeru šancí $\theta = \ln\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$ a rušivý parameter $\eta = \ln\frac{p_2}{1-p_2}$ na dáta v tabuľke a vypočítajte $\widehat{\theta}$.
(2b) Nakreslite funkciu vierohodnosti ako funkciu θ a η [superpozícia `contour()` a `image()`].
(2c) Vypočítajte vierohodnostný 95% DIS pre θ pomocou metodiky 15% cut-off štandardizovanej profilovej funkcie vierohodnosti. DIS dokreslite do jedného obrázka k profilovej funkcii vierohosnoti v jej 15% cut-off. DÚ

metastázy	rozšírené	lokalizované	spolu
áno	5	1	6
nie	10	9	19
spolu	15	10	25

Príklad 77 (jednovýberový test stredných hodnôt) *Toto nie je DÚ, len informácia o dátach.*

Hodnotený súbor: Z archívnych materiálov (Schmidt 1888) máme k dispozícii pôvodné kraniometrické údaje o dĺžke a šírke lebky zo starovekej egyptskej populácie. Súčasne máme k dispozícii priemerné hodnoty oboch rozmerov, hodnoty smerodajnej odchýlky a počty prípadov vzorky z novovekej egyptskej populácie (dĺžka lebky: $\overline{x}_{\mathrm{m}} = 177.568$ mm, $\overline{x}_{\mathrm{f}} = 171.962$ mm; s_{m}=7.526 mm, s_{f}=7.052 mm; n_{m}=88, $n_{\mathrm{f}} = 52$ a šírka lebky: $\overline{x}_{\mathrm{m}} = 136.402$ mm, $\overline{x}_{\mathrm{f}} = 131.038$ mm; $s_{\mathrm{m}} = 6.411$ mm, $s_{\mathrm{f}} = 5.361$ mm; $n_{\mathrm{m}} = 87$, $n_{\mathrm{f}} = 52$).

Súbor dát: `one-sample-mean-skull-mf.txt`

Popis premenných:
`id` – poradové číslo;
`pop` – populácie (egant – egyptská staroveká);
`sex` – pohlavie (m – muž, f – žena);
`skull.L` – najväčšia dĺžka mozgovne (mm), t.j. priama vzdialenosť kraniometrických bodov *glabella* a *opisthocranion*;
`skull.B` – najväčšia šírka mozgovne (mm), t.j. vzdialenosť oboch kraniometrických bodov *euryon*.

Biologické súvislosti: Brachycefalizácia, resp. debrachycefalizácia (t.j. relatívne skracovanie či predlžovanie lebky), patrí medzi prejavy sekulárneho trendu. Tieto zmeny lebky/hlavy korelujú so zmenami kostí končatín a dávajú sa do súvislostí so zmenami vonkajších životných podmienok i genetického zloženia populácie. Napriek tomu, že pomer šírky a dĺžky lebky závisí od oboch rozmerov, ukazuje sa, že zmeny v tvare lebky ovplyvňujú predovšetkým zmeny v jej šírke.

Ciele:
(A) zistiť, či sa dĺžka lebky starovekej egyptskej populácie líši v strednej hodnote od novovekej egyptskej populácie (zvlášť u mužov a u žien);
(B) zistiť, či sa šírka lebky starovekej egyptskej populácie líši v strednej hodnote od novovekej egyptskej populácie (zvlášť u mužov a u žien).

Príklad 78 (dvojvýberový test stredných hodnôt) *Toto nie je DÚ, len informácia o dátach.*

Hodnotený súbor: Máme k dispozícii údaje o pôrodnej hmotnosti prvorodených a druhorodených chlapcov, novorodencov narodených v krajskej nemocnici v priebehu jedného roka (Alánová 2008). Novorodencov narodených vo vyššom poradí sme z tohto porovnania vylúčili.

Súbor dát: `two-samples-means-birth.txt`

Popis premenných:
`o.sib.N` – počet starších súrodencov (0 – žiadny, 1 – jeden);
`birth.W` – pôrodná hmotnosť (g).

Biologické súvislosti: Z niektorých štúdií vyplýva, že medzi prvorodenými a druhorodenými novorodencami môžu byť rozdiely v pôrodnej hmotnosti. Prvorodení by potom mali mať nižšiu pôrodnú hmotnosť než deti narodené ako druhé v poradí.

Ciele:
(A) zistiť, či sa pôrodná hmotnosť prvorodených a druhorodených chlapcov z jednej pôrodnice a sezóny v priemere líši.