2 Model rozdelenia pravdepodobnosti a štatistický model

Pri riešení biomedicínskych problémov pomocou matematickej štatistiky sa vyskytuje stochastický (náhodný) element, ktorým sa nie je možné zaoberať len pomocou základných zákonov aritmetiky. Preto štatistické metódy potrebujú stochastické modely. Vývoj takýchto modelov tvorí *deduktívny*[14] (matematický) aspekt štatistiky. Štatistické problémy sú však *induktívne*[15], pretože vznikajú ako dôsledok pozorovania istých javov v reálnych biomedicínskych situáciách. Tieto javy sú výsledkom nejakého experimentu alebo pozorovania. Otázky, ktoré experiment alebo pozorovanie rieši, sú všeobecnejšieho typu. Pýtajú sa na niečo, čo nie je priamo pozorovateľné, ale je logicky obsiahnuté v dátach. Hovoríme, že usudzujeme („inferujeme“) niečo na základe dát. Na riešenie deduktívnych problémov matematiky často postačuje čiastočne dostupná informácia, aby sme boli schopní vytvoriť novú matematickú vetu. Na riešenie induktívnych problémov štatistiky potrebujeme všetky dostupné dáta. Len tak môžeme vyvodiť závery. Ignorovanie nejakej časti dát nie je akceptovateľné. Pri deduktívnych problémoch je kvalita novej matematickej vety rovnaká ako kvalita jej predchádzajúcich axióm, definícií alebo viet. Pri induktívnych problémoch je stupeň istoty v záveroch väčší ako v samotných dátach. Čím máme viac dát, tým sa kvalita výstupov zväčšuje. Avšak jedno nové pozorovanie môže naše závery zmeniť.

Dáta môžeme jednoducho popísať aj pomocou **charakteristík (parametrov) polohy a variability** a zobraziť ich pomocou **štatistickej grafiky** (pozri kap. Charakteristiky polohy a variability a štatistická grafika), čo je súčastou tzv. **exploratórnej analýzy dát** (EDA). Avšak použitím nejakého modelu sa o dátach dozvieme viac v zjednodušenej podobe, tento model nám navyše umožní interpretáciu výsledkov a uľahčí nám aj komunikáciu medzi dátami a biomedicínskou praxou.

Koncepcia modelu rozdelenia pravdepodobnosti a štatistického modelu predstavuje jeden zo základných piliérov bioštatistiky. Tieto modely sú charakterizované ich parametrami v prípade parametrického modelu, ktorým sa budeme podrobnejšie zaoberať. Jeho parametre slúžia na jednoduchú charakterizáciu dát a zjednodušujú interpretáciu výsledkov. Modely rozlišujeme podľa toho, či sú to modely na diskrétne alebo spojité dáta. **Model rozdelenia pravdepodobnosti** je charakterizovaný funkciou hustoty a distribučnou funkciou na základe presne špecifikovaných parametrov, ktoré je potrebné odhadnúť z dát pomocou funkcie vierohodnosti – ide teda o „model na dáta“[16]. Dáta sú realizáciami nahodnej premennej, o ktorej predpokladáme, že má asymptoticky (pre veľké n) nejaké rozdelenie, napr. normálne, binomické, multinomické, súčinové multinomické alebo Poissonove. Tento asymptotický predpoklad je fundamentálnym základom štatistickej inferencie, ktorá je procesom vyvodenia záverov (na základe dát a z nich vypočítaných odhadov parametrov) prostredníctvom postupu testovania hypotéz, ktorý používa tzv. štatistiky, testovacie štatistiky a ich asymptotické rozdelenia pravdepodobnosti (pozri kap. Testovanie hypotéz). Hypotézy testujeme aj v **štatistických modeloch**, ktoré často predstavujú modely kauzálnej závislosti zavislých premenných na prediktoroch. V týchto modeloch pomocou funkcie vierohodnosti odhadujeme parametre, ktoré zjednodušujú interpretácie výsledkov, či už štatisticky nesignifikantných alebo signifikantných, ale aj biologicky (medicínsky) nevýznamných alebo významných.

Všetky vyššie uvedené postupy sú súčasťou širšieho pojmu **(bio)štatistická analýza**. Bez správnej formulácie a aplikácie modelu rozdelenia pravdepodobnosti a štatistického modelu na dáta by štatistická analýza nebola možná a závery z nej by boli problematické.

Základom každej empirickej štúdie, či už experimentálnej alebo observačnej, je zabezpečiť **dáta** (**dátový súbor**, **realizácie**), ktoré označíme $\mathbf{x}$, procesom, ktorý nazývame **experiment** (**pokus**) alebo **meranie**. V najjednoduchšej podobe $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ (vektor pozorovaní, vektor výberových hodnôt, vektor realizácií), kde n je **rozsah (náhodného) výberu** a x_i sú realizácie vždy

[14] Dedukcia – na základe všeobecne platných záverov hľadáme riešenia nejakého konkrétneho prípadu.

[15] Indukcia – na základe konkrétnych prípadov vyvodzujeme všeobecnejšie platné závery.

[16] Model rozdelenia pravdepodobnosti je možné použiť aj v súvislosti so štatistikou – tu ide o „model pre štatistiku“ a s testovacou štatistikou – tu ide o „model pre testovaciu štatistiku“, ale aj v súvislosti s chybami štatistického modelu – tu ide o „model pre chyby (reziduály)“.

označované malými písmenami (všeobecne ozn. x). (Jednorozmerná) **náhodná premenná** X je funkcia z výberového priestoru $\mathcal{Y}$ (označovaná aj ako priestor elementárnych udalostí alebo množina výsledkov náhodného pokusu) do množiny reálnych čísel $\mathbb{R}$. Pozorovanie x je realizáciou náhodnej premennej X. Analogicky môžeme definovať **k-rozmerný náhodný vektor** $(X_1, X_2, \ldots, X_k)^T$. Príkladom takéhoto vektora je dvojrozmerný náhodný vektor $(X_{1i}, X_{2i})^T$, $i = 1, 2, \ldots, n$, s realizáciami $(x_{1i}, x_{2i})^T$ usporiadanými po riadkoch do matice $n \times 2$, ktorá je v tomto prípade dátovým súborom[17].

Príklad 46 (príklady náhodných premenných) *(1) Chirurg vykoná 100 transplantácií srdca, kde náhodná premenná X je napr. počet úspešne vykonaných transplantácií. (2) 50 bežcov beží maratón, kde náhodná premenná X je napr. čas odbehnutia maratónu v hodinách. (3) Hodíme 30-krát kockou, kde náhodná premenná X je napr. počet hodených šestiek. (4) Na 75 deťoch vo veku 10 rokov zmeriame výšku (v metroch) a hmotnosť (v kilogramoch), kde náhodná premenná X je napr. Rohrerov index ($RI = \frac{\text{hmotnosť v kg}}{(\text{výška v m})^3}$).*

Príklad 47 (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.*

Základný predpoklad je, že nejaké pozorovanie (výberová hodnota, realizácia) x je hodnota (realizácia) **náhodnej premennej** X a naším cieľom je použiť x na vyvodenie záverov o neznámom modeli (rozdelenia pravdepodobnosti alebo štatistickom) $F_*(\cdot)$ premennej X. Naše závery o $F_*(\cdot)$ sú zaťažené neistotou kvôli náhodnosti X, z ktorej pochádzajú x. Cieľom je zabezpečiť,

1. aby stupeň neistoty bol čo možno najmenší, berúc do úvahy náhodnosť X,
2. a aby bol tento stupeň neistoty vyjadrený v našich záveroch.

Ekvivalentne $(x_1, x_2, \ldots, x_k)^T$ je realizácia náhodného vektora $(X_1, X_2, \ldots, X_k)^T$ a naším cieľom je použiť túto realizáciu na vyvodenie záverov o neznámom k-rozmernom modeli $F_*^{(k)}(\cdot)$ náhodného vektora $(X_1, X_2, \ldots, X_k)^T$.

Definícia 14 (štatistická inferencia) *Štatistická inferencia (zriedkavo nazývaná aj štatistická indukcia) je proces vyvodenia záverov na základe dát prostredníctvom testovania hypotéz, modelu rozdelenia pravdepodobnosti alebo štatistického modelu (Cox, 2006). Tento proces je ovplyvnený náhodnými chybami, náhodným výberom, voľbou testovacieho kritéria, štatistického modelu alebo modelu rozdelenia pravdepodobnosti. Výsledkom tohto procesu sú zmysluplné závery aplikované na dobre definované dostatočne všeobecné situácie.*

Podstata vytvárania realizácií x limituje možnosti voľby modelu F_*. Inferencia bude o to presnejšia, o čo lepšie bude vybraná čo najmenšia množina $\mathcal{F}$ tak, aby $F_* \in \mathcal{F}$, kde $\mathcal{F}$ nazývame **množinou modelov** (štatistických modelov a modelov rozdelenia pravdepodobnosti). V mnohých prípadoch predpokladáme, že $X_i, i = 1, 2, \ldots, n$, sú **rovnako rozdelené náhodné premenné** (ozn.

iid; *independently identically distributed*). V tomto prípade hovoríme o **jednoduchom náhodnom výbere** (srs; *simple random sample*) s rozsahom n, kde X je charakterizované rozdelením $F_*(\cdot)$. angl

Príklad 48 (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*** *(Dalgaard, 2008). Ak vyberáme s vrátením, hovoríme o* ***jednoduchom náhodnom výbere s vrátením****. 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 $\mathcal{M} = \{1, 2, \ldots, 100\}$.*

Riešenie aj v R[18]

(a) Spolu máme $\binom{N}{n}$ možných náhodných výberov (kombinácie bez opakovania n-tej triedy z N prvkov množiny $\mathcal{M}$). 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 (kombinácie s opakovaním n-tej triedy z N prvkov množiny $\mathcal{M}$). 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 49 (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*.*

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)
```

Distribučná funkcia F náhodnej premennej X je definovaná ako $F_X(x) = \Pr(X < x)$, kde zápis znamená pravdepodobnosť, že náhodná premenná X nadobúda hodnoty menšie alebo rovné ako nejaké číslo x, ktoré nazývame *kvantil*. Dolný index v F_X sa spravidla vynecháva a píšeme F.

[17] Pre dvojrozmerný náhodný vektor sa často používa ozn. $(X, Y)^T$ a pre jeho realizáciu ozn. $(x, y)^T$.

[18] Detaily o jazyku R pozri napr. v Chambers (2008), Becker a kol. (1988) alebo Matloff (2011).

Náhodná premenná X sa nazýva **diskrétna**, ak jej distribučná fukcia F je schodovitá funkcia. V tomto prípade hovoríme, že ide o **diskrétne rozdelenie** a množina $\mathcal{Y}$, z ktorej pochádzajú X je konečná alebo nanajvýš spočítateľná (má spočítateľne veľa prvkov).

Distribučná funkcia diskrétnej náhodnej premennej X je definovaná ako (Azzalini, 1996)

$$F_X(x) = \Pr(X < x) = \sum_{i:x_i \leq x} \Pr(X = x_i),$$

kde $\sum_{i=1}^{k(\infty)} p_i = 1, \Pr(X = x_i) = p_i = f_X(x_i) = f(x_i), \forall x_i$, sa nazýva **pravdepodobnostná funkcia**. Často zapisujeme $\{x_i, p_i\}_{i=1}^{k(\infty)}$, kde x_i sú realizácie, p_i sú pravdepodobnosti výskytu x_i, $k \in \mathbb{N}^+$, kde $\mathbb{N}^+$ znamená množinu kladných prirodzených čísel[19].

Príklad 50 (diskrétna premenná) *(1) počet úspešne vykonaných transplantácií v SR, (2) počet hodených šestiek, (3) štyri vekové kategórie, do ktorých zaradíme subjekty, (4) pohlavie, (5) počet starších súrodencov (`o.sib.N`; dve kategórie; dáta: `two-samples-means-birth.txt`), (6) vzdelanie matky (`edu.M`, štyri kategórie; dáta: `anova-newborns.txt`).*

Náhodná premenná X sa nazýva **spojitá**, ak jej distribučná fukcia F je absolútne spojitá funkcia. V tomto prípade hovoríme, že ide o **spojité rozdelenie** a množina $\mathcal{Y}$ je nekonečná (nekonečne veľa prvkov).

Distribučná funkcia spojitej náhodnej premennej X je definovaná ako

$$F_X(x) = \int_{-\infty}^{x} f(t)\,dt, f(x) \geq 0,$$

kde $\int_{-\infty}^{\infty} f(x)\,dx = 1$, $f_X(x) = f(x) = \frac{\partial}{\partial x} F_X(x)$ sa nazýva **hustota**.

Príklad 51 (spojitá premenná) *(1) čas odbehnutia maratónu v hodinách; (2) hmotnosť v kilogramoch; (3) výška v centimetroch; (4) Rohrerov index; (5) dĺžkošírkový index lebky vypočítaný ako podiel náhodných premenných najväčšia šírka mozgovne a najväčšia dĺžka mozgovne (`skull.B` a `skull.L`; v mm; dáta: `one-sample-mean-skull-mf.txt`); (6) stranový rozdiel vertikálneho priemeru v strede dĺžky tela kľúčnej kosti na pravej a ľavej strane tela (`length.R` a `length.L`; v mm; dáta: `paired-means-clavicle2.txt`); (7) najväčšia výška mozgovne a morfologická výška tváre (`skull.pH` a `face.H`; v mm; dáta: `one-sample-correlation-skull-mf.txt`).*

Rozmery na živom objekte nemožno merať s absolútnou presnosťou a počet desatinných miest závisí na presnosti merania; ako posledné desatinné miesto by sa malo uvádzať to, na ktorom sa pri opakovanom meraní rovnakého objektu na viac desatinných miest zhodujú všetky merania.

Príklad 52 (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 $\Pr(a \leq X < b) = \Pr(X < b) - \Pr(X < a) = F_X(b) - F_X(a)$, kde $a = \mu - k\sigma$, $b = \mu + k\sigma$, $k = 1, 2, 3$.*

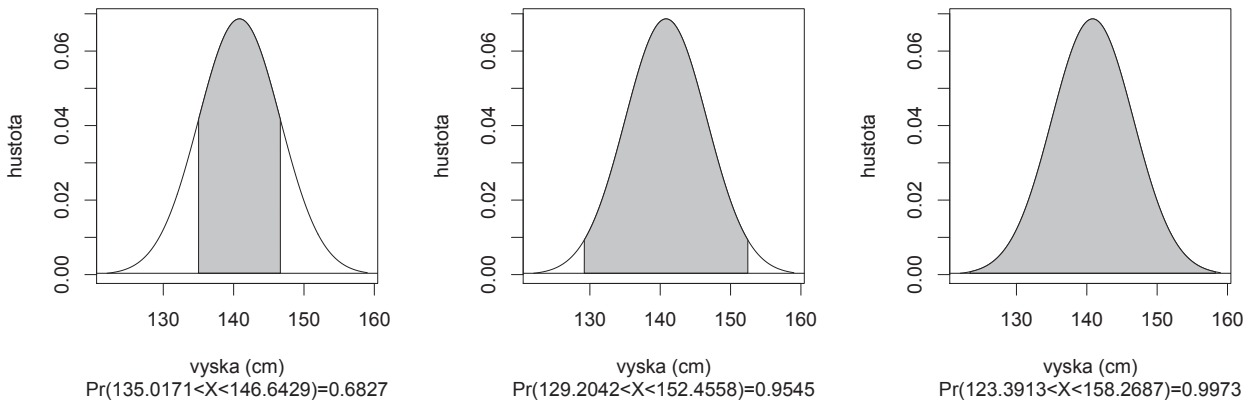
[19] Písmeno $\mathbb{N}$ sa v angličtine nazýva *blackboard bold* N (*double-struck capital* N).

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

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**“).



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

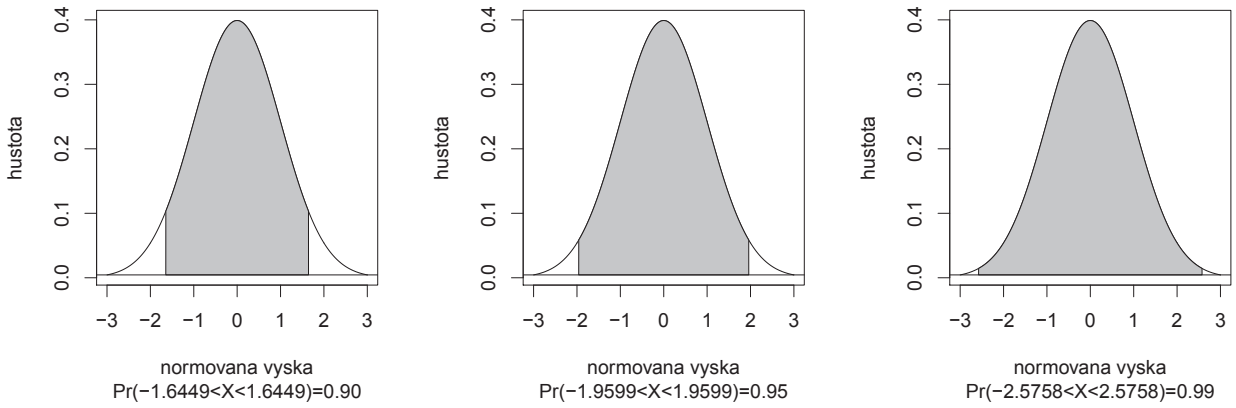
Príklad 53 (normálne rozdelenie) *Majme* $X \sim N(\mu, \sigma^2)$, *kde* $\mu = 150, \sigma^2 = 6.25$. *Vypočítajte* $a = \mu - x_{1-\alpha/2}\sigma$ *a* $b = \mu + x_{1-\alpha/2}\sigma$ *tak, aby* $\Pr(a \leq X \leq b) = 1 - \alpha$, *bola rovná* $0.90, 0.95$ *a* 0.99. *Číslo* $x_{1-\alpha}$ *je kvantil normálneho normovaného rozdelenia, t.j.* $\Pr(Z = \frac{X-\mu}{\sigma} < x_{1-\alpha}) = 1 - \alpha, Z \sim N(0, 1)$.

Riešenie (aj v R); (pozri obrázok 2)
$\Pr\left(\mu - x_{1-\alpha/2}\sigma < X < \mu + x_{1-\alpha/2}\sigma\right) = \Pr(X < \mu + x_{1-\alpha}\sigma) - \Pr(X < \mu - x_{1-\alpha}\sigma) = 1 - \alpha = 0.9$.
Z-transformáciou[20] na normálne normované rozdelenie dostaneme $\Pr(-x_{1-\alpha/2} < Z < x_{1-\alpha/2}) = 0.9$, kde $\frac{\mu - x_{1-\alpha/2}\sigma - \mu}{\sigma} = -x_{1-\alpha/2}$, $\frac{\mu + x_{1-\alpha/2}\sigma - \mu}{\sigma} = x_{1-\alpha/2}$, $x_{1-\alpha} = x_{0.95} = 1.64$, t.j. $90.00\,\%$ dát leží v intervale $\mu \pm 1.64\sigma$.

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

$\Pr(a < X < b) = 0.95$. Potom $x_{0.975} = 1.96$, t.j. 95.00 % dát leží v intervale $\mu \pm 1.96\sigma$.
$\Pr(a < X < b) = 0.99$. Potom $x_{0.995} = 2.58$, t.j. 99.00 % dát leží v intervale $\mu \pm 2.58\sigma$.

```
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 $\Pr\left(x_{\alpha/2} < Z < x_{1-\alpha/2}\right) = \Phi\left(x_{1-\alpha/2}\right) - \Phi\left(x_{\alpha/2}\right) = 1 - \alpha$, 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 54 (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?*

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 55 (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?*

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 pravdepodobnosti, 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 56 (binomické rozdelenie) *Predpokladajme, že pohlavie novorodencov (mužské alebo ženské) sa správa podľa modelu binomického rozdelenia s parametrami p (pravdepodobnosť výskytu chlapcov) a N (rozsah náhodného výberu), ozn. $Bin(N, p)$, kde $N = 1113$, $p = 0.52$, t.j. rodí sa o niečo viac chlapcov než dievčat (dáta: `two-samples-probabilities-sexratio.txt`). Aká je pravdepodobnosť, že sa narodí 700 a viac dievčat?*

Príklad 57 (binomické rozdelenie) *Predpokladajme, že $\Pr(vír) = 0.533 = p_1$ je pravdepodobnosť výskytu dermatoglyfického vzoru vír na palci pravej ruky mužov českej populácie a $\Pr(ostatné) = 0.467 = p_2$ 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 $N = 300$.*

$\mathcal{F}$ môže byť nejaká množina **distribučných funkcií** identifikovateľných pomocou parametra $\boldsymbol{\theta}$ z parametrického priestoru $\boldsymbol{\Theta} \in \mathbb{R}^k$ ($\mathbb{R}$ znamená množinu reálnych čísel[21]), čo môžeme formálne zapísať ako (Azzalini, 1996)

$$\mathcal{F} = \left\{ F(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k \right\},$$

kde pre každé fixné $\boldsymbol{\theta}$ je $F(\cdot; \boldsymbol{\theta})$ distribučnou funkciou, ktorej nosič je $\mathcal{Y}_{\boldsymbol{\theta}} \subseteq \mathbb{R}^n$. **Nosič** $\mathcal{Y}_{\boldsymbol{\theta}}$ je najmenšia množina, na ktorej je hustota definovaná. Výberový priestor je množina $\mathcal{Y}$ všetkých možných hodnôt x, ktoré charakterizujeme modelom. Formálne $\mathcal{Y} = \cup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{Y}_{\boldsymbol{\theta}}$. Často však $\mathcal{Y}_{\boldsymbol{\theta}}$ je rovnaký pre všetky $\boldsymbol{\theta}$, a preto koinciduje s $\mathcal{Y}$.

Príklad 58 (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.*

[21] $\mathbb{R}^k$ znamená k-rozmernú množinu reálnych čísel (k je dĺžka vektora parametrov $\boldsymbol{\theta} \in \boldsymbol{\Theta}$), kde ak $k = 1$, potom $\theta \in \Theta$ je číslo (skalár). $\mathbb{R}^n$ znamená n-rozmernú množinu reálnych čísel (n je rozsah náhodného výberu). $\mathbb{R}^+$ znamená množinu kladných reálnych čísel.

Príklad 59 (parametre) *Príkladmi parametrov môžu byť: (1) stredná hodnota μ a (2) rozptyl σ^2 dĺžky lebky (`skull-L`, v mm) egyptskej stredovekej mužskej populácie; (3) rozdiel medzi strednými hodnotami $\mu_1 - \mu_2$ dĺžky lebky mužskej a ženskej; (4) podiel rozptylov σ_1^2/σ_2^2 hodnôt dĺžky lebky u mužov a u žien (dáta: `two-samples-means-skull.txt`); (5) korelačný koeficient ρ medzi dĺžkou dolnej končatiny (`lowex.L`, v mm) a dĺžkou trupu (`tru.L`, v mm); (6) rozdiel korelačných koeficientov $\rho_1 - \rho_2$ medzi dĺžkou dolnej končatiny a dĺžkou trupu u mužov a u žien (dáta: `two-samples-correlations-trunk.txt`); (7) pravdepodobnosť p výskytu mužov (`sex`; m − muž, f − žena); (8) pravdepodobnosť výskytu popôrodných zmien p na ženských panvových kostiach u Afričaniek, ako aj rozdiel pravdepodobností $p_1 - p_2$ výskytu výrazných popôrodných zmien na panvových kostiach Afričaniek a Inuitiek (dáta: `more-samples-probabilities-pubis.txt`); (9) rozdiel pravdepodobností $p_1 - p_2$ sexuálnej orientácie na opačné pohlavie (`sexor` − sexuálna orientácia; `op` − výlučne na opačné pohlavie, `sa` − minimálne/občas na rovnaké pohlavie) u mužov a žien (dáta: `anova-head.txt`).*

Čítanie označení. 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“.

Tri typy priestorov. Výberový priestor $\mathcal{Y}$ súvisí s náhodnou premennou a jej realizáciou, nosič $\mathcal{Y}_{\boldsymbol{\theta}}$ súvisí s hodnotami, na ktorých je definovaná hustota rozdelenia pravdepodobnosti a parametrický priestor $\boldsymbol{\Theta}$ súvisí s parametrom $\boldsymbol{\theta}$.

Príklad 60 (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 $\mathcal{Y} = \{0, 1, \ldots, N\}$.*

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 $X \sim N(Np, Npq)$; príklady pozri v tabuľke 2.

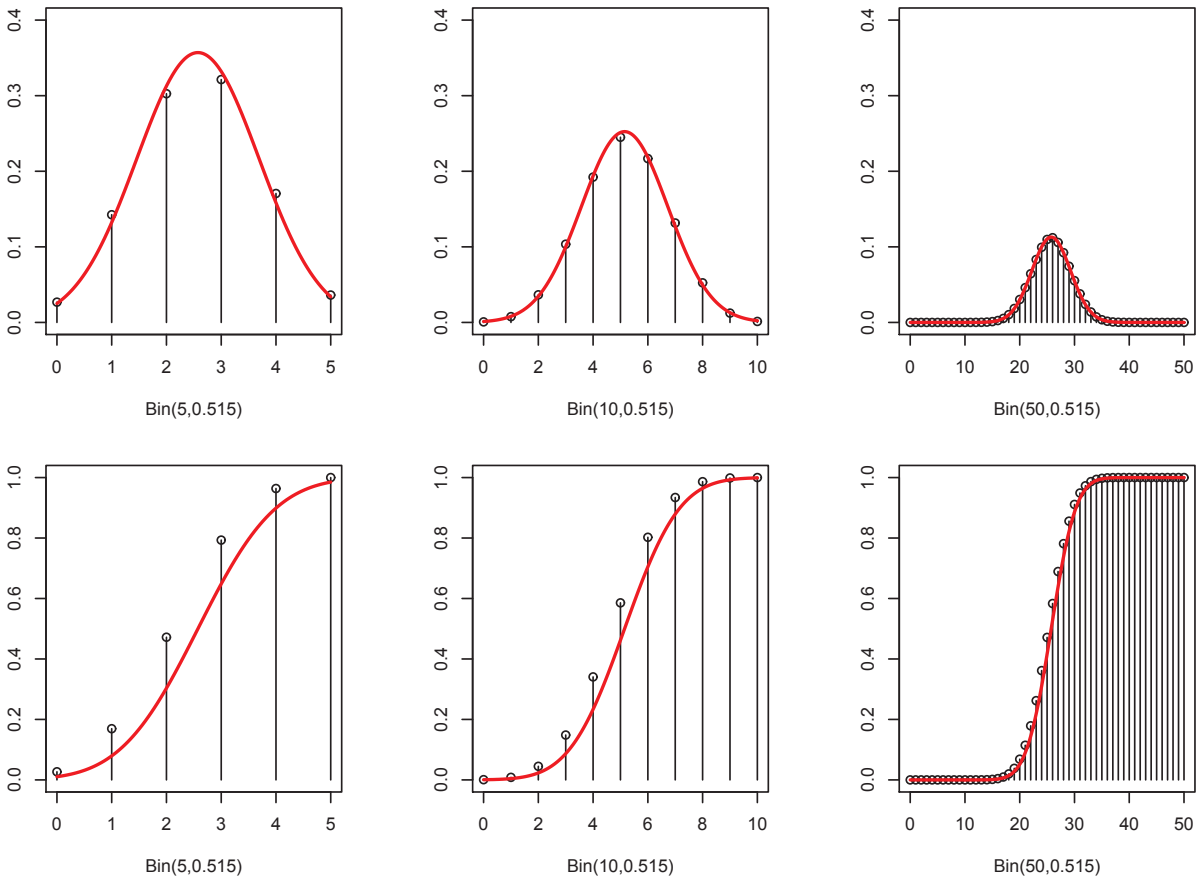
Tabuľka 2: Príklady minimálnych N pre fixované p potrebných na aproximáciu

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 61 (aproximácia binomického rozdelenia normálnym) *Nech* $\Pr(muž) = 0.515$ *znamená pravdepodobnosť výskytu mužov v populácii a* $\Pr(žena) = 0.485$ *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 $N(Np, Npq)$.*

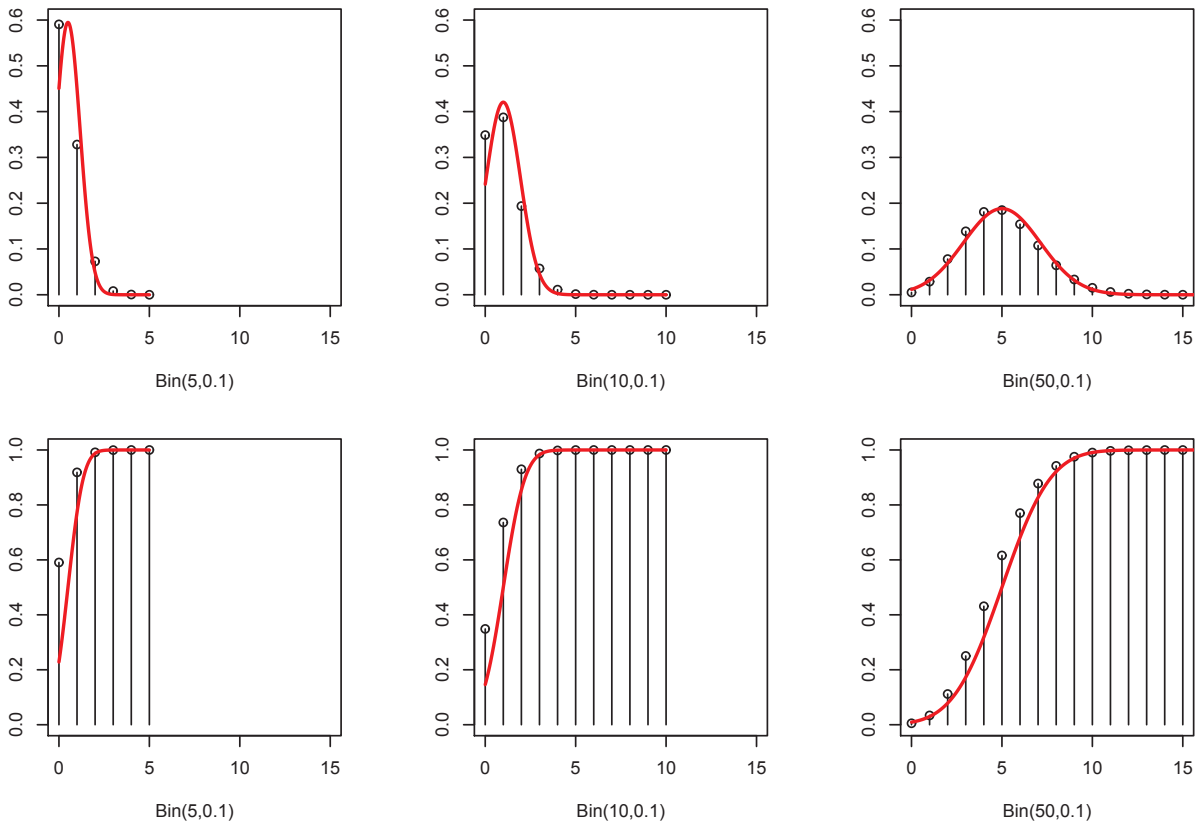
Riešenie (aj v R) (pozri obrázok 3 a 4)
Aproximácia znamená „približné vyjadrenie“, t.j. buď nejaké rozdelenie aproximujeme iným (majúcim isté výhody oproti tomu, ktoré aproximujeme), alebo aproximujeme dáta nejakým rozdelením (ktoré popisuje dáta pomocou ľahko interpretovateľných parametrov).
(a) $E[X] = Np = 5 \times 0.515 = 2.575, E[Y] = 5 \times 0.485 = 2.425$,
$\Pr(X \leq 3) = \sum_{k \leq 3} \binom{5}{k} 0.515^k 0.485^{5-k} = 0.793$,
$\Pr(X \leq 3) = 0.648, N(5 \times 0.515, 5 \times 0.515 \times 0.485)$.
(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
```



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)

Z príkladu 61 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



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)

pravdepodobnosti p blížiace sa jednotke alebo nule sú potrebné väčšie početnosti ako pre pravdepodobnosti p blízke hodnote 0.5.

Distribučné funkcie patriace do množiny $\mathcal{F}$ sú distribučnými funkciami diskrétnych alebo spojitých náhodných premenných. Potom $\mathcal{F}$ môže byť definovaná ako množina **pravdepodobnostných funkcií** alebo **funkcií hustoty** a model môžeme formálne zapísať ako (Casella a Berger, 2002)

$$\mathcal{F} = \left\{ f(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k \right\}$$

pre nejakú funkciu hustoty f. Vektor $\boldsymbol{\theta}$ sa nazýva **parameter**, množina $\boldsymbol{\Theta}$ **parametrický priestor** a $\mathcal{F}$ **parametrický model**. Keďže prvky $\mathcal{F}$ sú asociované s prvkami $\boldsymbol{\Theta}$, existuje $\boldsymbol{\theta}_* \in \boldsymbol{\Theta}$ asociovaná s F_* a nazýva sa **skutočná hodnota parametra**. A štatistická inferencia je práve o $\boldsymbol{\theta}_*$. Paramertický štatistický model môže byť charakterizovaný aj pomocou k-rozmerného vektora parametrov $\boldsymbol{\theta}$, preto sa v označení používa $\mathbb{R}^k$, kde pre skalár θ bude $k = 1$.

Ak však

$$\mathcal{F} = \{\text{množina všetkých hustôt funkcií jednej premennej}\},$$

ide o **neparametrický model** (Wasserman, 2006).

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 $\boldsymbol{\theta} = (\mu, \sigma^2)^T$. Hustota tohto rozdelenia má tvar $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$.

Príklad 62 (normálne rozdelenie) *Predpokladajme, že náhodná premenná X má asymptoticky (pre veľké n) normálne rozdelenie so strednou hodnotou $E[X] = \mu$ a rozptylom $Var[X] = \sigma^2$, čo zapisujeme ako $X \sim N(\mu, \sigma^2)$. Príkladmi takýchto náhodných premenných sú: (1) dĺžka pravej kľúčnej kosti u mužov (`length.R`; dáta: `paired-means-clavicle2.txt`); (2) šírka lebky u žien (`skull.B`; dáta: `one-sample-mean-skull-mf.txt`).*

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 $\boldsymbol{\theta} = (0, 1)^T$. Hustota tohto rozdelenia má tvar $\phi(x) = f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}$.

Príklad 63 (štandardizované normálne rozdelenie) *Predpokladáme, že náhodná premenná X šírka lebky u mužov (`skull.B`; dáta: `one-sample-mean-skull-mf.txt`) má asymptoticky normálne rozdelenie so strednou hodnotou μ a rozptylom σ^2, čo zapisujeme ako $X \sim N(\mu, \sigma^2)$. Keď od X odpočítame jej strednú hodnotu μ a tento rozdiel vydelíme odmocninou z rozptylu $\sigma = \sqrt{\sigma^2}$, dostaneme náhodnú premennú Z, ktorá má asymptoticky normálne rozdelenie so strednou hodnotou $\mu = 0$ a rozptylom $\sigma^2 = 1$, čo zapisujeme ako $Z \sim N(0, 1)$.*

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 (Casella a Berger, 2002)

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

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

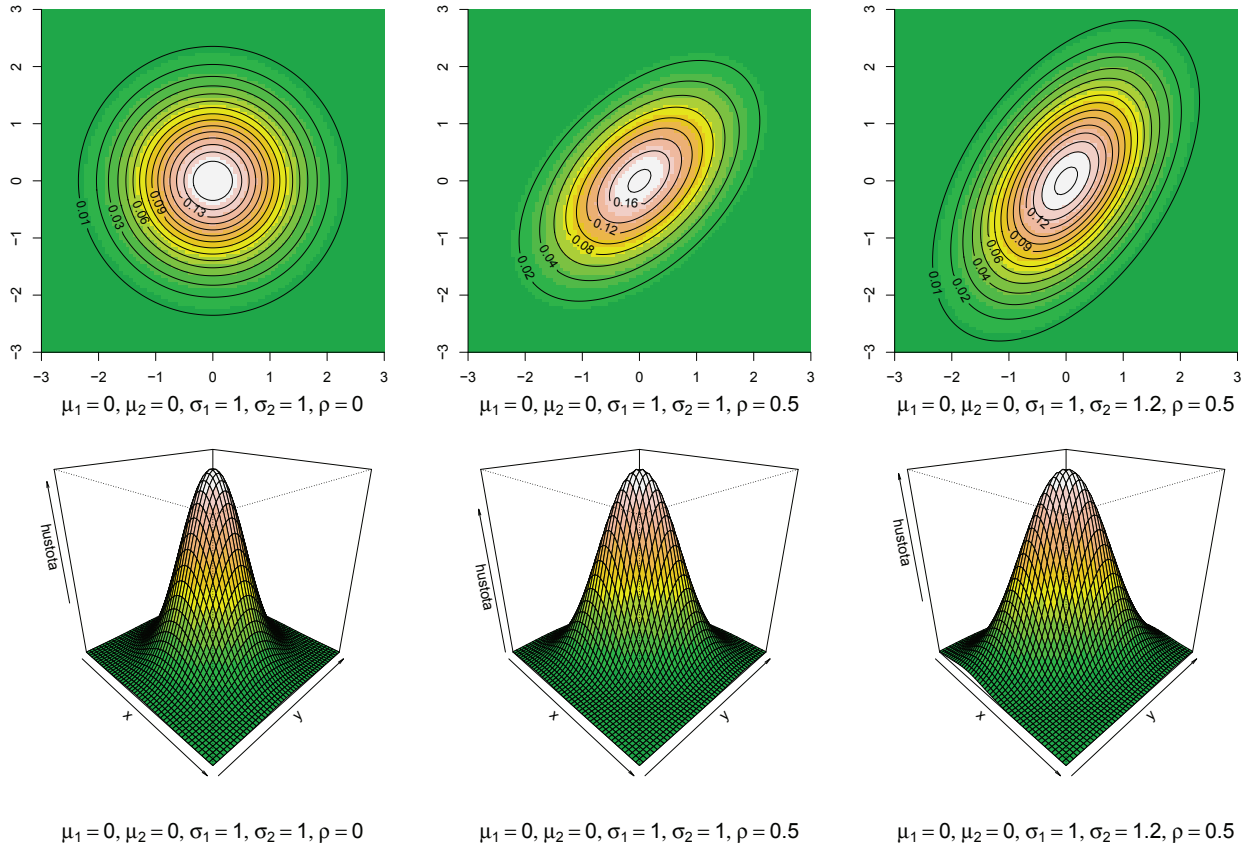
$$-\frac{1}{2} \begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix},$$

marginálne rozdelenia sú $X \sim N(\mu_1, \sigma_1^2)$ a $Y \sim N(\mu_2, \sigma_2^2)$, ρ je koeficient korelácie (detaily maticovej algebry pozri v (Gentle, 2007). Margináne rozdelenie je rozdelenie marginálnej náhodnej premennej, tu X nezávisle na Y a naopak Y nezávisle na X.

Z vyššie uvedeného textu 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.

Príklad 64 (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*

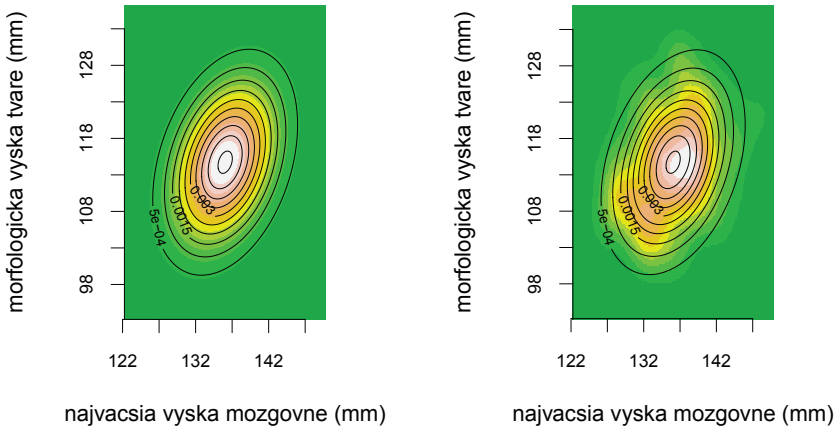
(a) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1,\ \rho = 0$*;*
(b) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1,\ \rho = 0.5$*;*
(c) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1.2,\ \rho = 0.5$.
Vzorové riešenie pozri na obrázku 5.



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 65 (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-mf.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 $\widehat{\boldsymbol{\theta}} = (\overline{x}_1, \overline{x}_2, s_1^2, s_2^2, r)^T$ odhadnutý z dát, kde r je Pearsonov korelačný koeficient. Riešenie pozri na obrázku 6.*



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)

Štandardizované dvojrozmerné normálne rozdelenie. Náhodný vektor $(X, Y)^T$ má dvojrozmerné normálne rozdelenie

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

s hustotou (Bickel a Doksum, 2006)

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

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

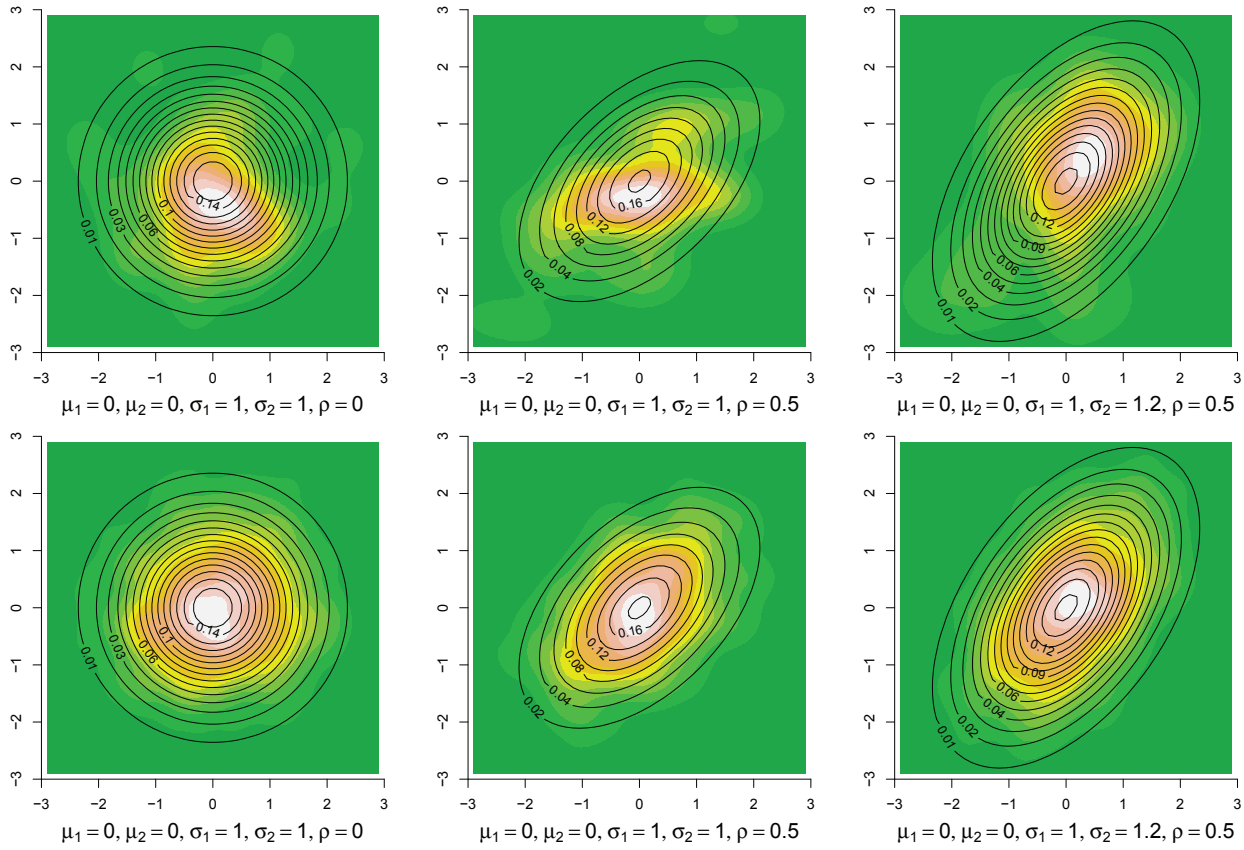
$$-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix},$$

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

Príklad 66 (štandardizované dvojrozmerné normálne rozdelenie) *Nech náhodnou premennou $X \sim N(\mu_1, \sigma_1^2)$ je najväčšia výška mozgovne u mužov (*`skull.pH`*; v mm) a náhodnou premennou $Y \sim N(\mu_2, \sigma_2^2)$ je morfologická výška tváre u mužov (*`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*

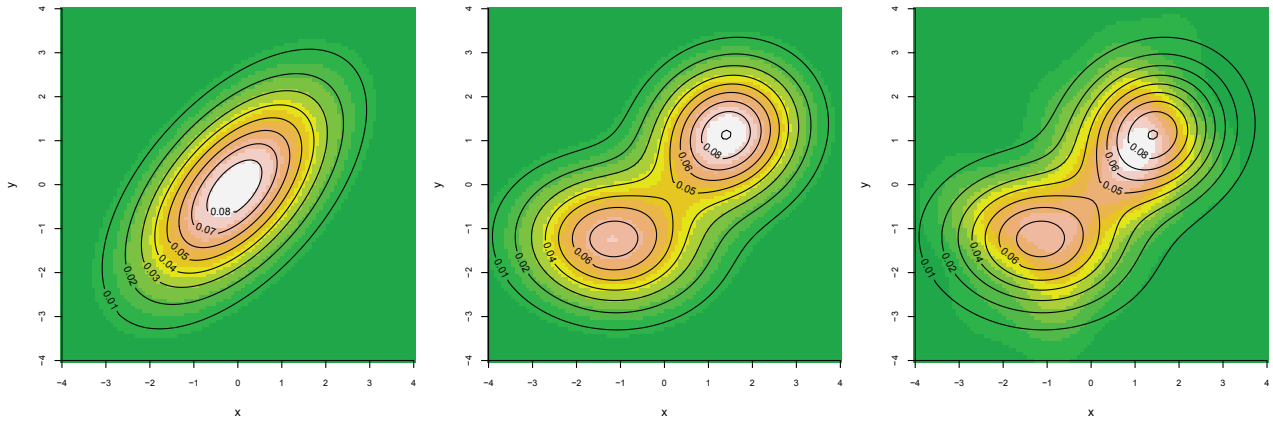
hodnotu μ_2 a tento rozdiel vydelíme odmocninou z rozptylu σ_2, dostaneme náhodnú premennú Z_Y, ktorá má asymptoticky normálne rozdelenie so strednou hodnotou $\mu_2 = 0$ a rozptylom $\sigma_2^2 = 1$, čo zapisujeme ako $Z_Y \sim N(0,1)$. Potom $(Z_X, Z_Y)^T$ má štandardizované dvojrozmerné normálne rozdelenie $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ s parametrami $\boldsymbol{\mu} = (0,0)^T$ a $\sigma_1^2 = 1$, $\sigma_2^2 = 1$ a ρ sú parametre kovariančnej matice $\boldsymbol{\Sigma}$ (dáta: `one-sample-correlation-skull-mf.txt`).

Príklad 67 (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 $n = 1000$)

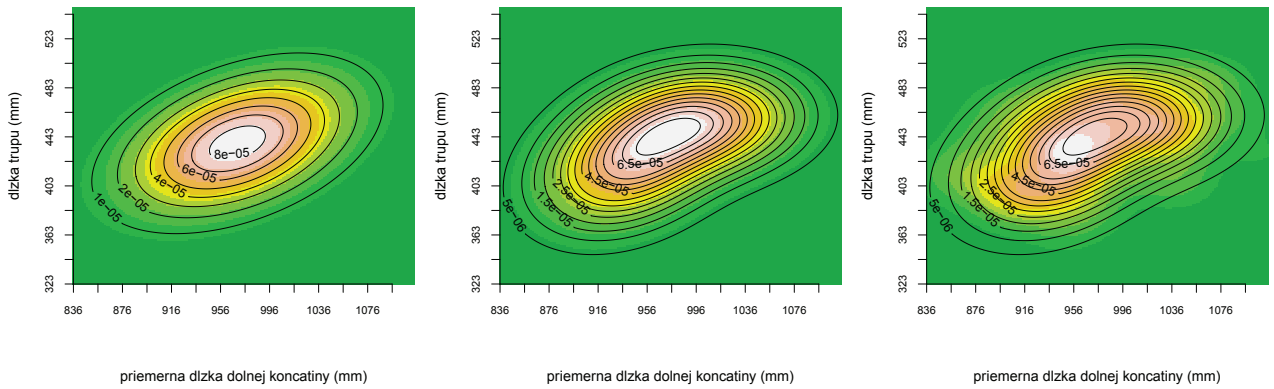
Príklad 68 (zmes dvoch dvojrozmerných normálnych rozdelení) *Simuláciu pseudonáhodných čísel zo zmesi dvoch normálnych rozdelení* $pN_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-p)N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *môžeme v* R *urobiť použitím jedného z alternatínych postupov z príkladu 67. 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()`*. Teoretickým rozdelením v tomto prípade bude* $N_2\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)$*.*
(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) $\widehat{\boldsymbol{\theta}} = (\overline{x}_{11}, \overline{x}_{12}, s^2_{11}, s^2_{12}, r_1, \overline{x}_{21}, \overline{x}_{22}, s^2_{21}, s^2_{22}, r_2)^T$*,* $n_1 = n_2 = 50$ *a* $p = 0.5$ *(odhady pochádzajú z nasimulovaných dát).*
(2) $\widehat{\boldsymbol{\theta}} = (\overline{x}_1, \overline{x}_2, s^2_1, s^2_2, r)^T$ *a* $n_1 = n_2 = 50$ *(odhady pochádzajú zo spoločného výberu nasimulovaných dát).*
Vzorové riešenie pozri na obrázku 8.



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) – simulačná štúdia

Príklad 69 (zmes dvoch dvojrozmerných normálnych rozdelení) *Nech* $(X_1, Y_1)^T$ *pochádza z rozdelenia* $N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$*, 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 rozdelenia* $N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$*, kde* X_2 *je 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* $\boldsymbol{\theta} = (\mu_{11}, \mu_{12}, \sigma^2_{11}, \sigma^2_{12}, \rho_1, \mu_{21}, \mu_{22}, \sigma^2_{21}, \sigma^2_{22}, \rho_2)^T$ *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 realizácií (1) pomocou funkcie `image()` *a superponujte ho 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) $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_{11}, \widehat{\mu}_{12}, \widehat{\sigma}^2_{11}, \widehat{\sigma}^2_{12}, \widehat{\rho}_1, \widehat{\mu}_{21}, \widehat{\mu}_{22}, \widehat{\sigma}^2_{21}, \widehat{\sigma}^2_{22}, \widehat{\rho}_2)^T$ *a* $p = n_1/(n_1+n_2)$*; parametre sú odhadnuté z dát.*
(2) $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}^2_1, \widehat{\sigma}^2_2, \widehat{\rho})^T$*; parametre sú odhadnuté zo spoločného výberu.*
Vzorové riešenie pozri na obrázku 9 (dáta `two-samples-correlations-trunk.txt`*).*

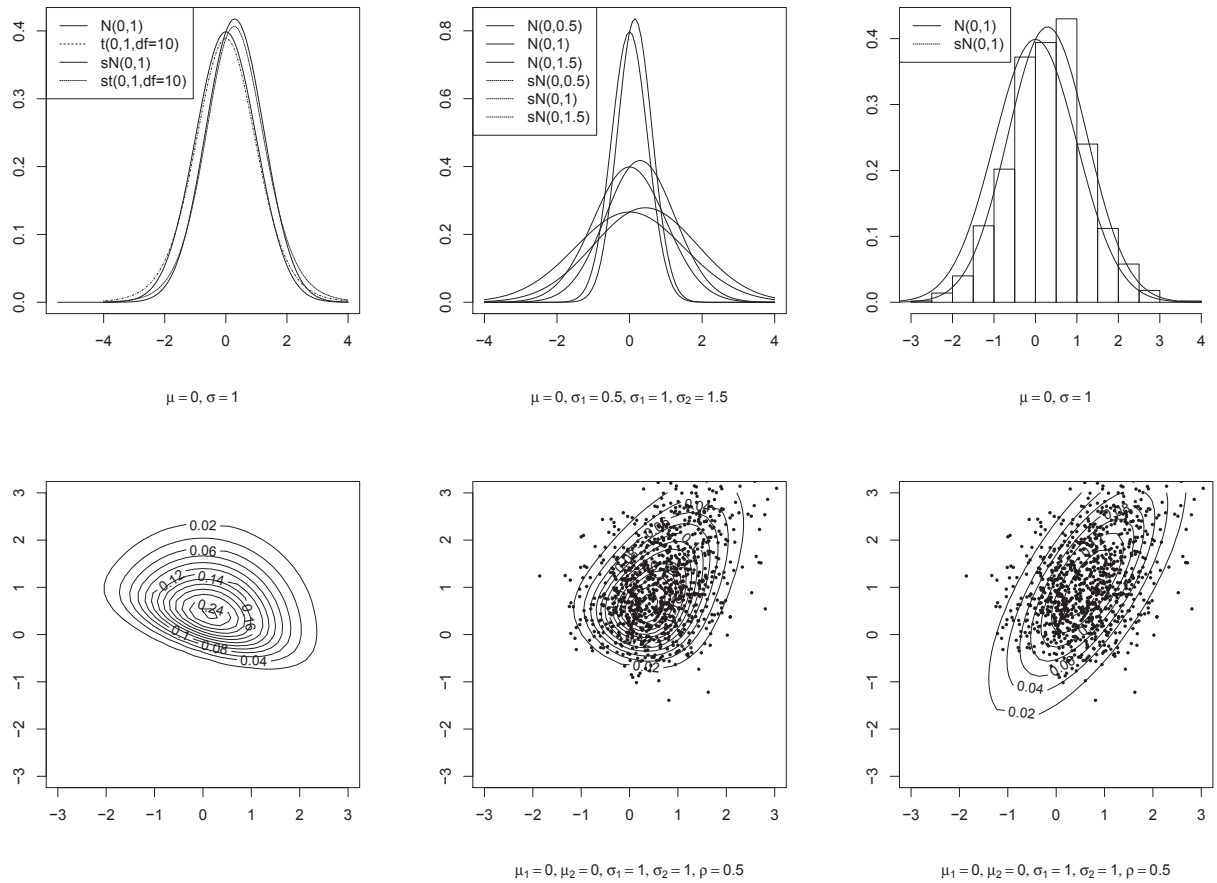


Obr. 9: Spoločná hustota dvojrozmerného normálneho rozdelenia (vľavo), hustota zmesi dvoch dvojrozmerných normálnych rozdelení (uprostred) a dvojrozmerný jadrový odhad superponovaný hustotou zmesi dvoch dvojrozmerných normálnych rozdelení (vpravo) – reálne dáta

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 môže byť napr. zošikmená jedna alebo obe premenné (príklad zošikmenia oboch premenných zľava pozri na obrázku 10, dolný riadok).

Binomické rozdelenie. Majme nezávislé identické *Bernoulliho* pokusy s odpoveďami $X_i = 1$ (udalosť nastala) alebo $X_i = 0$ (udalosť nenastala) pre $i = 1, 2, ..., N$, kde N je počet nezávislých pokusov. Pravdepodobnosť nastatia udalosti pre každý pokus $\Pr(X_i = 1) = p$, pravdepodobnosť neúspechu pre každý pokus $\Pr(X_i = 0) = 1 - p$. Počet nastatí udalosti $X = \sum_{i=1}^{N} X_i$, pravdepodobnosť nastatia udalosti je p. Náhodná premenná X má binomické rozdelenie s parametrami N a p, t.j. $X \sim Bin(N, p)$, kde $\theta = p$. Pravdepodobnosť, že X je rovné nejakému číslu $x = n$ (x je realizácia X) zapisujeme ako $\Pr(X = x) = \binom{N}{x} p^x (1-p)^{N-x}$, pre $x = 0, 1, 2, \ldots, N$ (Christensen, 1997). Stredná hodnota náhodnej premennej X je definovaná ako $E[X] = \sum_{x=0}^{N} x \binom{N}{x} p^x (1-p)^{N-x} = Np$ a rozptyl $Var[X] = \sum_{x=0}^{N} (x - Np)^2 \binom{N}{x} p^x (1-p)^{N-x} = Np(1-p)$.

Strieška v označení $\widehat{p}$ sa všeobecne používa na označenie odhadov parametrov rozdelení pravdepodobnosti. Tieto odhady sa počítajú z dát. Pre spojité náhodné premenné označujeme rozsah



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

náhodného výberu n, ale pri binomickom (a iných diskrétnych) rozdeleniach máme početnosti dve – počet úspechov a rozsah náhodného výberu – preto pre počet úspechov rezervujeme ozn. n a rozsah náhodného výberu N. Počet úspechov ozn. aj $x = n$, ak ide o realizáciu náhodnej premennej X.

Ekvivalentne môžeme rozdelenie náhodnej premennej pochádzajúcej z binomického rozdelenia zapísať ako $\mathbf{X} \sim Bin\,(N, p, 1-p)$, kde $\mathbf{X} = (X_1, X_2)^T$, $\boldsymbol{\theta} = (p, 1-p)^T$, X_1 je počet úspechov, $X_2 = N - X_1$ je počet neúspechov, $X_1 \sim Bin\,(N, p)$, $X_2 \sim Bin\,(N, 1-p)$. Potom $E[X_1] = Np$, $E[X_2] = N\,(1-p)$, $Var[X_2] = Np\,(1-p) = Var[X_1]$ nezávisle na p, $Cov\,[X_1, X_2] = -Np\,(1-p)$ a $Cor\,[X_1, X_2] = -1$. (Vysvetlenia označení pozri v poznámke pod čiarou[22]). Realizácie náhodných premenných X_1 a X_2 budeme označovať ako n_1 a n_2. Tento typ označenia je vhodnejší z dôvodu zovšeobecnenia binomického rozdelenia na viacrozmerné rozdelenia, kde $\mathbf{n} = (n_1, n_2)^T$ a $\mathbf{p} = (p_1, p_2)^T$, $p_1 = p$ a $p_2 = 1 - p$ (pre porovnanie pozri ozn. v práci Verzani, 2005). Potom $\boldsymbol{\theta} = \mathbf{p}$.

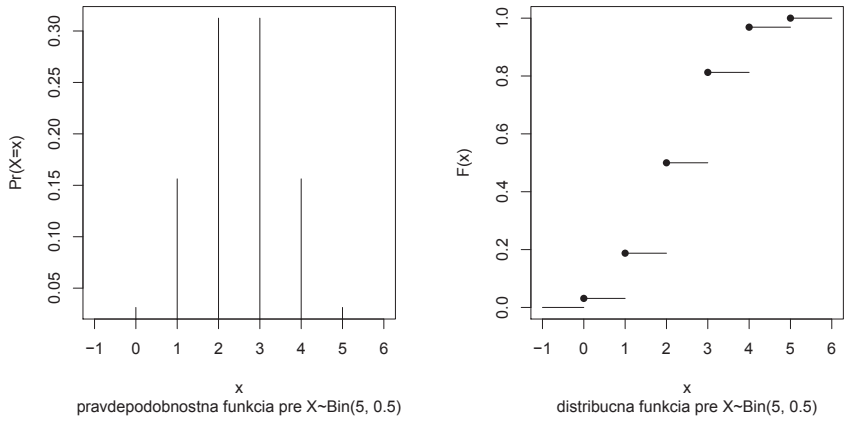
Príklad 70 (binomické rozdelenie, binomický experiment) *Experiment pozostávajúci z fixného počtu Bernouliho experimentov (ozn. N) sa nazýva binomický experiment. Pravdepodobnosť úspechu ozn. p, pravdepodobnosť neúspechu $q = 1 - p$. Náhodná premenná X je počet pozorovaných úspechov počas experimentu. 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*

[22] $E[X]$ je označenie pre strednú hodnotu náhodnej premennej X, $Var[X]$ pre rozptyl, $Cov[X, Y]$ pre kovarianciu dvoch premenných X a Y a $Cor[X, Y]$ pre korelačný koeficient.

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 v R (pozri obrázok 11)

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



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

Multinomické rozdelenie. Nech N je počet nezávislých identických pokusov a v každom z nich môže nastať $J \geq 2$ navzájom disjunktných udalostí s možnými odpoveďami $X_{ij} = 1$ (udalosť nastala) alebo $X_{ij} = 0$ (udalosť nenastala), kde $i = 1, 2, \ldots, N$ a $j = 1, 2, \ldots, J$. Potom $X_j = \sum_{i=1}^{N} X_{ij}$. Pravdepodobnosť nastatia j-tej udalosti v i-tom pokuse $\Pr(X_{ij} = 1) = p_j$, $\sum_{j=1}^{J} p_j = 1$. Náhodná premenná $\mathbf{X} = (X_1, X_2, \ldots, X_J)^T$ má (J-rozmerné) multinomické rozdelenie s parametrami N a $\mathbf{p}$, t.j. $\mathbf{X} \sim Mult_J(N, \mathbf{p})$, kde $\boldsymbol{\theta} = \mathbf{p}$ a $\mathbf{p} = (p_1, p_2, \ldots, p_J)^T$. Pravdepodobnosť, že X_j je rovné nejakému číslu n_j zapisujeme ako (Casella a Berger, 2002)

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_J = x_J) = \frac{N!}{x_1! x_2! \ldots x_J!} p_1^{x_1} p_2^{x_2} \ldots p_J^{x_J} = \frac{N!}{\prod_j x_j!} \prod_{j=1}^{J} p_j^{x_j},$$

kde $N = \sum_{j=1}^{J} X_j$, $X_j \geq 0$ a $x_j = n_j$ sú realizácie X_j. Potom $\mathbf{n} = (n_1, n_2, \ldots, n_J)^T$. Pre marginálne rozdelenie píšeme $X_j \sim Bin(N, p_j)$, kde stredná hodnota $E[X_j] = Np_j$, rozptyl $Var[X_j] = Np_j(1-p_j)$, kovariancia $Cov[X_i, X_j] = -Np_ip_j$, korelačný koeficient $Cor[X_i, X_j] = (-p_ip_j)/\sqrt{p_i(1-p_i)p_j(1-p_j)}$. Stredná hodnota $E[\mathbf{X}] = N\mathbf{p}$ a kovariančná matica $Var[\mathbf{X}] = N\left(\mathbf{D_p} - \mathbf{pp}^T\right)$, kde $\mathbf{D_p} = \mathrm{diag}(\mathbf{p})$ a

$$\left(\mathbf{D_p} - \mathbf{pp}^T\right)_{ij} = \begin{cases} p_i(1-p_i) & \text{ak } i = j \\ -p_ip_j & \text{ak } i \neq j \end{cases}.$$

Disjunktnosť znamená, že v i-tom pokuse mohla udalosť nastať len raz, t.j. výsledkom takéhoto pokusu môže byť vektor napr. $(1, 0, \ldots, 0)^T$ alebo $(0, 1, \ldots, 0)^T$, kde okrem jednej jednotky na j-tom mieste máme vždy $J - 1$ núl na ostatných miestach. Kedže sumácia $\mathbf{p}$ nám dáva jednotku, $Var[\mathbf{X}]$ je singulárna matica. Matica diag($\mathbf{p}$) má diagonálu v podobe $\mathbf{p}$ a mimodiagonálne prvky rovné 0. Ak $J = 2$, potom $Bin(N, p) \approx Mult_2(N, \mathbf{p})$, t.j. multinomické rozdelenie je zovšeobecnením binomickeho rozdelenia.

Príklad 71 (multinomické rozdelenie; príklady) *Príklady premenných, o ktorých predpokladáme, že majú multinomické rozdelenie:*
(1) farba dúhovky – hodnotená podľa škály R. Martina (Martin, 1914/1928) a kategorizovaná do štyroch kategórií: hnedá, hnedozelená, melírovaná a modrá (dáta: `multinom-iris-color.txt`*);*
(2) zakončenie troch hlavných dlaňových línií – kategorizované do troch kategórií: vysoké, stredné a nízke (dáta: `multinom-palmar-lines.txt`*);*
(3) priľahlosť ušného laloka – podľa priľahlosti k hlave kategorizovaná do troch kategórií: priľahlý, stredne priľahlý, odstávajúci (dáta: `multinom-earlobe.txt`*);*
(4) krvná skupina – kategorizovaná v AB0 systéme do štyroch kategórií: skupina 0, A, B a AB (dáta: `multinom-blood-groups.txt`*).*

Príklad 72 (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 pozri v tabuľke 3. Vypočítajte $Var[X_1]$, $Var[X_3]$, $Cov[X_1, X_3]$, $Cor[X_1, X_3]$ a očakávané početnosti $Np_j, j = 1, 2, \ldots, 8$.*

Tabuľka 3: Kontingenčná tabuľka 2×3 pravdepodobností p_j pre dva socioekonomické statusy, dve politické príslušnosti a dve politické filozofie (multinomické rozdelenie)

	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

Riešenie
$X = (X_1, X_2, \ldots, X_8) \sim Mult_J(N, \mathbf{p})$, 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 3 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 početnosti pre každú bunku tabuľky sú (všeobecne nemusia byť) celé čísla, pozri tabuľku 4.

Súčinové multinomické rozdelenie. Nech N_k je počet nezávislých identických pokusov a v každom z nich môže nastať $J \geq 2$ navzájom disjunktných udalostí s možnými odpoveďami $X_{kji} = 1$ (udalosť nastala) alebo $X_{kji} = 0$ (udalosť nenastala), kde $i = 1, 2, \ldots, N_k$, $k = 1, 2, \ldots, K$

Tabuľka 4: Kontingenčná tabuľka 2×3 očakávaných početností Np_j pre dva socioekonomické statusy, dve politické príslušnosti a dve politické filozofie (multinomické rozdelenie)

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

a $j = 1, 2, \ldots, J$. Nech $X_{kj} = \sum_{i=1}^{N_k} X_{kji}$ a $\sum_{k=1}^{K} N_k = N$. Pravdepodobnosť nastatia (kj)-tej udalosti v i-tom pokuse $\Pr(X_{kji} = 1) = p_{kj}$, $\sum_{k=1}^{K} \sum_{j=1}^{J} p_{kj} = 1$. Náhodná premenná $\mathbf{X}_k = (X_{k1}, X_{k2}, \ldots, X_{kJ})^T$ má (J-rozmerné) multinomické rozdelenie s parametrami N_k a $\mathbf{p}_k$, t.j. $Mult_J(N_k, \mathbf{p}_k)$, kde $\boldsymbol{\theta}_k = \mathbf{p}_k$ a $\mathbf{p}_k = (p_{k1}, p_{k2}, \ldots, p_{kJ})^T$. Realizácie náhodnej premennej $\mathbf{X}_k$ označujeme ako $\mathbf{x}_k$. Potom $x_{kj} = n_{kj}$ a naviac $\mathbf{n}_k = (n_{k1}, n_{k2}, \ldots, n_{kJ})^T$. Nech $\mathbf{X}_k$ sú nezávislé, potom platí (Christensen, 1997)

$$\Pr(X_{kj} = x_{kj}, \forall k, j; j = 1, 2, \ldots, J; k = 1, 2, \ldots, K) = \prod_{k=1}^{K} \Pr(X_{kj} = x_{kj}, \forall j; j = 1, 2, \ldots, J)$$

a $\sum_{j=1}^{J} X_{kj} = N_k$ pre $\forall k$. Ďalej platí

$$\Pr(X_{kj} = x_{kj}, \forall j) = \left(N_k! / \prod_{j=1}^{J} x_{kj}! \right) \prod_{j=1}^{J} p_{kj}^{x_{kj}}.$$

Z toho vyplýva, že

$$\Pr(X_{kj} = x_{kj}, \forall k, j; j = 1, 2, \ldots, J; k = 1, 2, \ldots, K) = \prod_{k=1}^{K} \left(\left(N_k! / \prod_{j=1}^{J} x_{kj}! \right) \prod_{j=1}^{J} p_{kj}^{x_{kj}} \right).$$

Očakávané hodnoty $N_k p_{kj}$, rozptyly $Var[X_{kj}]$ a kovariancie $Cov[X_{kj}]$ a korelácie $Cor[X_{kj}]$ vnútri nejakého $\mathbf{X}_k$ vieme vypočítať. Kovariancie medzi rôznymi $\mathbf{X}_k$, napr. $Cov[\mathbf{X}_1, \mathbf{X}_2]$, sú nuly kvôli nezávislosti jednotlivých $\mathbf{X}_k$. Potom $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K)^T$ má súčinové multinomické rozdelenie s parametrami $\boldsymbol{\theta}_k = \mathbf{p}_k, k = 1, 2, \ldots, K$.

Príklad 73 (súčinové multinomické rozdelenie; príklady) *Príklady premenných, o ktorých predpokladáme, že majú súčinové multinomické rozdelenie:*
(1) farba dúhovky – hodnotená podľa škály R. Martina (Martin, 1914/1928) a kategorizovaná do štyroch kategórií (hnedá, hnedozelená, melírovaná a modrá) zisťovaná súčasne s výskytom (prítomnosť, neprítomnosť) radiálnych útvarov v štruktúre dúhovky (dáta: `multinom-iris-color.txt`*);*
(2) zakončenie hlavných dlaňových línií – kategorizované do troch kategórií (vysoké, stredné a nízke) zisťované súčasne s odtieňmi farby vlasov (svetlá, stredná a tmavá) na základe štandardnej Fischer-Sallerovej stupnice odtieňov (dáta: `multinom-palmar-lines.txt`*);*
(3) priľahlosť ušného laloka – podľa priľahlosti k hlave kategorizovaná do troch kategórií (priľahlý, stredne priľahlý, odstávajúci) zisťovaná u mužov a u žien (dáta: `multinom-earlobe.txt`*);*
(4) krvná skupina – kategorizovaná v AB0 systéme do štyroch kategórií (0, A, B a AB) zisťovaná v Košiciach a v Prahe (dáta: `multinom-blood-groups.txt`*).*

Príklad 74 (súčinové multinomické rozdelenie) *Majme dáta z príkladu 72 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 $Cor\,[X_{11}, X_{23}]$.*

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 3 prepísať na súčinovo-multinomický model, teda podmienené pravdepodobnosti $p_{j|i}$ budú dané socioekonomickým statusom i (pozri tabuľku 5).

Tabuľka 5: Kontingenčná tabuľka 2×3 pravdepodobností $p_{j|i}$ pre dva socioekonomické statusy, dve politické príslušnosti a dve politické filozofie (súčinové multinomické rozdelenie)

	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 $N_2 = 20$ pozri očakávané početnosti v tabuľke 6.

Tabuľka 6: Kontingenčná tabuľka 2×3 očakáných početností $N_i p_{j|i}$ pre dva socioekonomické statusy, dve politické príslušnosti a dve politické filozofie (súčinové multinomické rozdelenie)

	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 75 (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 $p_j, j = 1, 2, \ldots, 9$, pozri v tabuľke 7. $\mathbf{X} = (X_1, X_2, \ldots, X_9)^T \sim Mult_9(N, \mathbf{p})$. 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?*

Tabuľka 7: Kontingenčná tabuľka 3×3 pravdepodobností p_j pre tri farby vlasov a tri farby očí (multinomické rozdelenie)

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 veľké a pravdepodobnosti výskytu udalostí p nie príliš malé. V opačnom prípade je vhodné Poissonovo rozdelenie (Agresti, 2002).

Poissonovo rozdelenie. Poissonovo rozdelenie je limitným pripadom Binomického rozdelenie $Bin(N, p)$, kde $N \to \infty, p \to 0$, teda $Np \to \lambda$. Ak je X náhodná premenná s Poissonovým rozdelením a parametrom $\theta = \lambda$, $X \sim Poiss(\lambda)$, potom (Casella a Berger, 2002)

$$\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, \ldots.$$

Realizáciu náhodnej premennej X označujeme ako $x = n$. $E[X] = \lambda$ a $Var[X] = \lambda$. Tomu korešponduje

$$\binom{N}{x} p^x (1-p)^{N-x} = \left[(Np)^x (1-p)^N / x!\right] (1-p)^{-x} \frac{N!}{(N-x)! N^x},$$

ak $N \to \infty, p \to 0$ a $Np \to \lambda \Rightarrow$ $(Np)^x \to \lambda^x$, $(1-p)^N \to e^{-\lambda}$, $(1-p)^{-x} \to 1$, $\frac{N!}{(N-x)!N^x} \to 1$.

Znak „$\to$“ čítame ako „konverguje“ (do nekonečna, k nule, k λ a pod.). V prípade, že za „$\to$“ je znak nekonečna, napr. $N \to \infty$, potom čítame celý výraz ako „pre (dostatočne) veľké N“ alebo „N konverguje do nekonečna“.

Príklad 76 (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 77 (Poissonovo rozdelenie; príklady) *Príklady premenných, o ktorých predpokladáme, že majú Poissonove rozdelenie:*
(1) početnosť absencie palmárneho trirádia d (Okajima a Iwayanagi, 1986);
(2) incidencia os japonicum (os zygomaticum bipartitum/tripartitum), variety lícnej kosti, kedy bežne celistvá kosť je aj v dospelosti rozdelená švami na dve až tri časti, pozri Hauser, De Stefano a kol. (1989, s. 222) a Dodo (1974);
(3) celková perioperačná mortalita, t.j. mortalita v súvislosti s výkonom chirurgických operačných zákrokov (Bainbridge a kol., 2012);
(4) výskyt ďalšieho infarktu myokardu do 30 dní po operácii u pacientov, ktorí boli liečení na ochorenie koronárnych tepien chirurgickým zásahom pomocou by-passu, kde operácia prebehla klasickým spôsobom so zastavením srdca a umelou cirkuláciou (Cheng a kol., 2005);
(5) výskyt ďalšieho infarktu myokardu do 30 dní po operácii u pacientov, ktorí boli liečení na ochorenie koronárnych tepien chirurgickým zásahom pomocou by-passu, kde operácia prebehla moderným postupom na tlčúcom srdci (off-pump coronary artery bypass; Cheng a kol. (2005);
(6) výskyt bilaterálnej agenézy horných laterálnych rezákov 12 *a* 22 *(Alt a kol., 2013).*

Príklad 78 (Poissonovo rozdelenie; pruské armádne jednotky) *Nech početnosti Pruských armádnych jednotiek, v ktorých nastalo n úmrtí zapríčinených kopnutím koňom za rok (von 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 8). Vypočítajte očakávané početnosti, za predpokladu $X \sim Poiss(\lambda)$, kde $\lambda = \frac{\sum_n n m_n}{\sum_n m_n}$.*

Tabuľka 8: Pozorované a očakávané početnosti m_n (zaokrúhlené na nula desatinných miest) Pruských armádnych jednotiek, v ktorých nastalo n úmrtí zapríčinených kopnutím koňom

n	0	1	2	3	4	5+
pozorované m_n	109	65	22	3	1	0
očakávané m_n	109	66	20	4	1	0

Príklad 79 (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 nezávislé poissonovské náhodné premenné s parametrami $(\lambda_1, \lambda_2, \lambda_3)$ a $X_1 + X_2 + X_3 \sim Poiss(\lambda_1 + \lambda_2 + \lambda_3)$.*

Zovšeobecnením príkladu 79 dostaneme

$$X_1 + X_2 + \ldots + X_J \sim Poiss\,(\lambda_1 + \lambda_2 + \ldots + \lambda_J)$$

Poissonovo vs. multinomické rozdelenie. Dá sa ukázať, že

$$(X_1 + X_2 + \ldots + X_J)\,|N \sim Mult_J\,(N, p_1, p_2, \ldots, p_J)\,,$$

kde $N = \sum_j X_j$ a $p_j = \lambda_j / \sum_j \lambda_j, j = 1, 2, \ldots, J$. Ak $X_j, j = 1, 2, \ldots, J$ sú nezávislé, $X_j \sim Poiss\,(\lambda_j)$, kde $E(X_j) = \lambda_j$, potom podmienená pravdepodobnosť, že všetky $X_j = x_j$ za podmienky $N = \sum_j X_j$ sa rovná

$$\begin{aligned}\Pr\left[(X_1 = x_1, X_2 = x_2, \ldots, X_J = x_J)|\sum_j X_j = N\right] &= \frac{\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_J = x_J)}{\Pr(\sum_j X_j = N)} \\ &= \frac{\prod_j \lambda_j^{x_j} e^{-\lambda_j}/x_j!}{\lambda^N e^{-\lambda}/N!} = \frac{N! e^{-\lambda} \prod_j \lambda_j^{x_j}}{\prod_j \lambda^x e^{-\lambda} x_j!} \\ &= \frac{N!}{\prod_j x_j!} \prod_j \left(\frac{\lambda_j}{\lambda}\right)^{x_j}, \text{ kde } p_j = \frac{\lambda_j}{\lambda}.\end{aligned}$$

Toto podmienené rozdelenie sa často používa pri log-lineárnych modeloch. Ak máme početnosti X_j pochádzajúce z Poissonovho rozdelenia, potrebujeme ich celkovú sumu N („grand total“). Teda potrebujeme podmiené rozdelenie pri danom N, čo je v podstate multinomické rozdelenie.

Overdispersion a underdispersion. V praxi často variabilita presahuje variabilitu danú binomickým či Poissonovým modelom alebo je variabilita menšia ako variabilita daná binomickým či Poissonovým modelom. Prepokladáme, že každý človek má rovnakú pravdepodobnosť úmrtia pri dopravnej nehode budúci týždeň. Realistickejšie však tieto pravdepodobnosti varírujú napr. podľa času stráveného šoférovaním, závisia od toho, či osoba má zapnuté pásy, závisia od geografickej polohy a pod. angl

V prípade *overdispersion*, teda v prípade, keď rozptyl presahuje strednú hodnotu, je realistickejšie nahradiť Poissonovo rozdelenie **negatívne binomickým rozdelením** (Agresti, 2002). Ak máme binomické (prip. multinomické) rozdelenie, tiež môže nastať prípad *overdispersion*, pretože skutočné rozdelenie je zmes rôznych binomických rozdelení s parametrami varírujúcimi kvôli nenameraným (mätúcim) premenným. angl angl

Negatívne binomické rozdelenie. Majme nezávislé identické *Bernoulliho* pokusy s odpoveďami $X_i = 1$ (udalosť nastala) alebo $X_i = 0$ (udalosť nenastala) pre $i = 1, 2, \ldots$. Pravdepodobnosť nastatia udalosti pre každý pokus $\Pr(X_i = 1) = p$, pravdepodobnosť neúspechu pre každý pokus $\Pr(X_i = 0) = 1 - p$. Nech X je počet úspechov pred k-tym neúspechom. Potom $\Pr(X = x) = \binom{x+k-1}{x} p^x (1-p)^k$ má negatívne binomické rozdelenie s $E[X] = k\frac{p}{1-p}$ a $Var[X] = k\frac{p}{(1-p)^2}$, ozn. $X \sim Negbin(k, p)$. Poissonovo rozdelenie je limitným prípadom negatívne binomického rozdelenia, kde $k \to \infty$, $p \to 0$ a fixovaným $kp = \lambda$.

Príklad 80 (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 $Bin(N, p)$.*

Overdispersion v binomickom modeli. Nech $X_i \sim Bin(N, p_i)$ a nech $X = X_I$ je náhodne zvolené angl

z X_i, kde náhodný index $I = i$ má pravdepodobnosť $1/m$. Budeme teda mať zmes binomických rozdelení s marginálnou pravdepodobnosťou

$$\Pr(X = x) = E[\Pr(X_I = x|I)] = \frac{1}{m}\sum_{i=1}^{m}\Pr(X_i = x) = \frac{1}{m}\sum_{i=1}^{m}\binom{N}{x}p_i^x(1-p_i)^x,$$

$$E[X] = E[E[X_I|I]] = \frac{1}{m}\sum_{i=1}^{m}E[X_i] = \frac{N}{m}\sum_{i=1}^{m}p_i = N\pi,$$

kde $\pi = \sum_{i=1}^{m} p_i/m$ a

$$\begin{aligned} Var[X] &= E[Var[X_I|I]] + Var[E[X_I|I]] = \frac{1}{m}\sum_{i=1}^{m}Var[X_i] + \frac{1}{m}\sum_{i=1}^{m}E^2[X_i] - \left(\sum_{i=1}^{m}E[X_i]/m\right)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}Np_i(1-p_i) + \frac{1}{m}\sum_{i=1}^{m}(Np_i)^2 - (N\pi)^2 = N\pi(1-\pi) + N(N-1)\sigma_p^2, \end{aligned}$$

kde σ_p^2 je definovaná ako pri *underdispersion*. Z toho vyplýva, že pri náhodnej voľbe z rôznych individuálnych pravdepodobností p_i je rozptyl väčší ako rozptyl za platnosti binomického modelu.

angl

Príklad 81 (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 početnosti chlapcov (pozri tabuľku 9). 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. $X \sim Bin(N, \pi)$.*

Tabuľka 9: Pozorované početnosti rodín m_n s n chlapcami

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

Riešenie

Tabuľka 10: Očakávané početnosti rodín m_n (zaokrúhlené na nula desatinných miest) s n chlapcami (binomické rozdelenie)

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 (pozri tabuľku 9) a vypočítané (teoretické) m_n (pozri tabuľku 10) zistíme, že pozorované poukazujú na *overdispersion*, t.j. máme väčšie početnosti rodín s malým a veľkým množstvom chlapcov v porovnaní s teroretickými početnosťami. angl

Underdispersion v binomickom modeli. Nech $X_1, X_2, \ldots, X_N$ sú nezávislé binomické pokusy angl

s pravdepodobnosťami $p_1, p_2, \ldots, p_N$. Nech $X = \sum_{i=1}^N X_i$, potom $E[X] = \sum_{i=1}^N p_i = N\pi$, kde $\pi = \frac{1}{N}\sum_{i=1}^N p_i$, ale

$$\begin{aligned} Var[X] &= \sum_{i=1}^N Var[X_i] = \sum_{i=1}^N p_i(1-p_i) = \sum_{i=1}^N p_i - \sum_{i=1}^N p_i^2 \\ &= \sum_{i=1}^N p_i - \frac{\left(\sum_{i=1}^N p_i\right)^2}{N} - \left(\sum_{i=1}^N p_i^2 - \frac{\left(\sum_{i=1}^N p_i\right)^2}{N}\right) \\ &= N\pi - N\pi^2 - N\sigma_p^2 = N\pi(1-\pi) - N\sigma_p^2, \end{aligned}$$

kde $\sigma_p^2 = \frac{1}{N}\left(\sum_{i=1}^N p_i^2 - \frac{\left(\sum_{i=1}^N p_i\right)^2}{N}\right)$ je rozptyl medzi p_i. Z toho vyplýva, že pri rôznych individuálnych pravdepodobnostiach p_i je rozptyl menší ako rozptyl za platnosti binomického modelu.

Overdispersion v Poissonovom modeli. Predpokladajme, že náhodná premenná X má rozptyl $Var[X]$ podmienený strednou hodnotou $E[X] = \mu$, kde μ je náhodná premenná so strednou hodnotou $E[\mu]$ a rozptylom $Var[\mu] = \sigma_\mu^2$. Teda pre jednotlivé subjekty μ, charakterizujúca napr. nehodu, varíruje. Hoci počet nehôd na subjekt má rozdelenie $Poiss(\mu)$, marginálne rozdelenie bude charaterizované *overdispersion*, a teda angl

$$E[X_\mu] = E[E[X_\mu|\mu]] = E[\mu] \text{ a } Var[X_\mu] = E[Var[X_\mu|\mu]] + Var[E[X_\mu|\mu]] = E[\mu] + \sigma_\mu^2,$$

čo poukazuje na väčšiu variabilitu v porovnaní s Poissonovým modelom. Za predpokladu, že μ má **gama rozdelenie**, môžeme ľahko spočítať marginálne pravdepodobnosti, t.j. ak X_μ má Poissonovo rozdelenie so strednou hodnotou μ, μ má hustotu $f(\mu) = \frac{1}{\Gamma(\alpha)}\mu^{\alpha-1}\lambda^\alpha e^{-\lambda\mu}$. Náhodná premenná X predstavujúca zmes X_μ má strednú hodnotu $E[X] = E[\mu] = \frac{\alpha}{\lambda}$ a rozptyl $Var[X] = E[\mu] + Var[\mu] = \frac{\alpha}{\lambda} + \frac{\alpha}{\lambda^2}$. Marginálna pravdepodobnosť pre $x = 0, 1, \ldots,$ je potom rovná

$$\begin{aligned} \Pr(X = x) &= E[\Pr(X_\mu = x|\mu)] = E[e^{-\mu}\frac{\mu^x}{x!}] = \frac{\lambda^\alpha}{\Gamma(\alpha)x!}\int e^{-\mu}\mu^x\mu^{\alpha-1}e^{-\lambda\mu}d\mu \\ &= \frac{\lambda^\alpha\Gamma(x+\alpha)}{(\lambda+1)^{x+\alpha}\Gamma(\alpha)x!} = \binom{x+\alpha-1}{\alpha-1}\left(\frac{\lambda}{\lambda+1}\right)^\alpha\left(1-\frac{\lambda}{\lambda+1}\right)^x, \end{aligned}$$

kde $(x+\alpha-1)! = \Gamma(x+\alpha)$. Ide teda o negatívne binomické rozdelenie, kde X je počet neúspechov (úrazov, zlyhaní) zaznamenaných po α úspechoch, pravdepodobnosť úspechu $\pi = \frac{\lambda}{\lambda+1}$ a pomer zlyhaní $\mu = \frac{1-\pi}{\pi}\alpha$.

angl

Príklad 82 (overdispersion v Poissonovom modeli) *Majme početnosti robotníkov m_n s n úrazmi v továrni (pozri v tabuľku 11; Greenwood a Yule (1920)). 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. $X \sim Poiss(\lambda)$.*

Tabuľka 11: Pozorované početnosti robotníkov m_n s n úrazmi v továrni

n	0	1	2	3	4	≥ 5
pozorované m_n	447	132	42	21	3	2

Riešenie

Keď porovnáme pozorované m_n (pozri tabuľku 11) a vypočítané (teoretické, očakávané) m_n (pozri tabuľku 12) zistíme, že pozorované poukazujú na *overdispersion*, t.j. máme viac robotníkov bez úrazu ako aj viac robotníkov s väčším množstvom úrazov v porovnaní s teoretickými početnosťami.

Tabuľka 12: Očakávané početnosti robotníkov m_n (zaokrúhlené na nula desatinných miest) s n úrazmi v továrni (Poissonovo rozdelenie)

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

2.1 *Simulačný experiment ako nástroj štúdia teoretických vlastností modelov

Monte Carlo (MC) experiment. Pojem MC metóda je známy od 40. rokov dvadsiateho storočia a zaviedli ho fyzici pracujúci na projekte o jadrových zbraniach v Los Alamos National Laboratory, menovite Stanislav Ulam, Enrico Fermi, John von Neuman a Nicholas Metropolis. MC je odvodené od kasína v Monaku, kde Ulamov strýko hrával hazardné hry. Použitie tzv. **náhodnosti a opakovateľnosti MC procesu** je analogické aktivitám v kasíne. Prvýkrát metódu použil Enrico Fermi v roku 1930 na výpočet vlastností novoobjaveného neutrónu. MC experiment bol použitý tiež v 50. rokoch dvadsiateho storočia počas vývoja vodíkovej bomby. U.S. Air Force bola v tom čase hlavnou organizáciou zodpovednou za financovanie a rozšírenie informácie o MC metódach, ktoré si začali hľadať cestu k mnohým ďalším aplikáciám, najprv vo fyzike, neskôr v chémii a nakoniec aj v matematike a štatistike. V štatistike sa MC metódy používajú na študovanie asymptotických vlastností odhadov a testovacích štatistík (príp. štatistických modelov) a zisťovanie ich správania sa za kontrolovaných podmienok (pozri napr. Rizzo, 2007; Gentle, 2009). Jediný predpoklad **dobrej simulácie pseudonáhodných čísel** používaný v MC metódach je **dostatočná náhodnosť v širšom slova zmysle** (Suess a Trumbo, 2010).

Simulačný experiment (mnohonásobne opakované náhodné výbery) musí spĺňať nasledovné tri kritériá (Robert a Casella, 2010)

1. **relevantnosť** – vygenerované (simulované) dáta musia byť generované na základe relevantných pravidiel, napr. kombinácia minulých skúseností a súčasných dát, *hypotetického pravdepodobnostného* alebo *štatistického modelu*, ktorý chceme študovať a pod.;

2. **stabilita** – *centrálna limitná veta* (CLV; použiteľná aj pre n menšie ako 100) a *dva zákony veľkých čísel* (použiteľné pre n väčšie ako 100 alebo 1000) garantujú, že ak je plán simulačnej štúdie správny a simulácia má dostatočne veľa opakovaní, dostaneme stabilný výsledok namiesto náhodného šumu;

3. **diagnostika** – pomocou rôznych numerických a grafických metód môžeme rozlíšiť signál od šumu, napr. porovnanie numerického a analytického riešenia, porovnanie viacerých podobných modelov, použitie dostatočného množstva opakovaní a pod.

Veta 1 (CLV) *Nech $X_i,\ i = 1, 2, \ldots, n$, sú iid náhodné premenné s rovnakou strednou hodnotou $E[X_i] = \mu \in \mathbb{R}$ a konečným rozptylom $Var[X_i] = \sigma^2 < \infty$. Potom $\frac{1}{\sigma\sqrt{n}}(\sum_{i=1}^{n} X_i - n\mu)$ má limitne (pre dostatočne veľké n) štandardizované normálne rozdelenie (t.j. konverguje v distribúcii k štandardizovanému normálnemu rozdeleniu $N(0, 1)$; Wasserman (2006).*

Veta 2 (Slabý zákon veľkých čísel) *Nech $X_i,\ i = 1, 2, \ldots, n$, sú iid náhodné premenné s rovnakou strednou hodnotou $E[X_i] = \mu \in \mathbb{R}$ a konečným rozptylom $Var[X_i] = \sigma^2 < \infty$. Nech $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Potom pre každé $\epsilon > 0$ platí $\lim_{n\to\infty} \Pr(|\overline{X}_n - \mu| < \epsilon) = 1$, t.j. pre dostatočne veľké n platí, že $\overline{X}_n$ konverguje v pravdepodobnosti k μ (Lehmann, 1999).*

Veta 3 (Silný zákon veľkých čísel) *Nech X_i, $i = 1, 2, \ldots, n$, sú iid náhodné premenné s rovnakou strednou hodnotou $E[X_i] = \mu \in \mathbb{R}$ a konečným rozptylom $Var[X_i] = \sigma^2 < \infty$. Nech $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Potom pre každé $\epsilon > 0$ platí $\Pr(\lim_{n\to\infty} |\overline{X}_n - \mu| < \epsilon) = 1$, t.j. pre dostatočne veľké n platí, že $\overline{X}_n$ konverguje skoro iste k μ (Lehmann, 1999).*

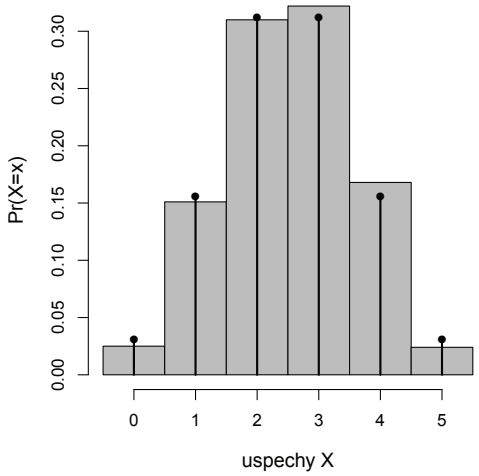
CLV a zákony veľkých čísel nám v praxi zabezpečia použitie nejakého modelu rozdelenia pravdepodobnosti na reálne dáta za predpokladu, že máme dostatočne veľký rozsah náhodného výberu (DasGupta, 2008). Dôsledky použitia modelu rozdelenia pravdepodobnosti na reálne dáta s malým rozsahom siahajú od nesprávneho použitia štatistického modelu po nesprávne použitie štatistického testu, čoho dôsledkom je nedôveryhodná interpretácia výsledkov štatistickej analýzy.

Príklad 83 (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.*

Riešenie (pozri tabuľku 13 a obrázok 12)

Tabuľka 13: Simulované a teoretické relatívne početnosti úspechov

relatívne početnosti/n	0	1	2	3	4	5
simulované	0.025	0.151	0.310	0.322	0.168	0.024
teoretické	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 84 (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 (aj v R)
Príklad hovorí o tom, ako dobre normálne rozdelenie aproximuje binomické pri rozsahu $N = 100$, čo je dôležité pri testovaní hypotéz.

$E[X_N] = Np = 50$, $\sqrt{Var[X_N]} = \sqrt{Np(1-p)} = 5$.
Ak $Y_N = X_N/N$, potom $\Pr(|Y_N - 1/2| < \epsilon) = 0.236$, kde $\epsilon = 0.02$. $\Pr(0.48 < Y_{100} < 0.52) = \Pr(48 < X_{100} < 52) = \Pr(48.5 < X_{100} < 51.5) = \Pr(\frac{48.5-50}{5} < Z_{100} < \frac{51.5-50}{5})$, kde $X_{100} \sim N(50, 5)$ s použitím úpravy na spojitosť.

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

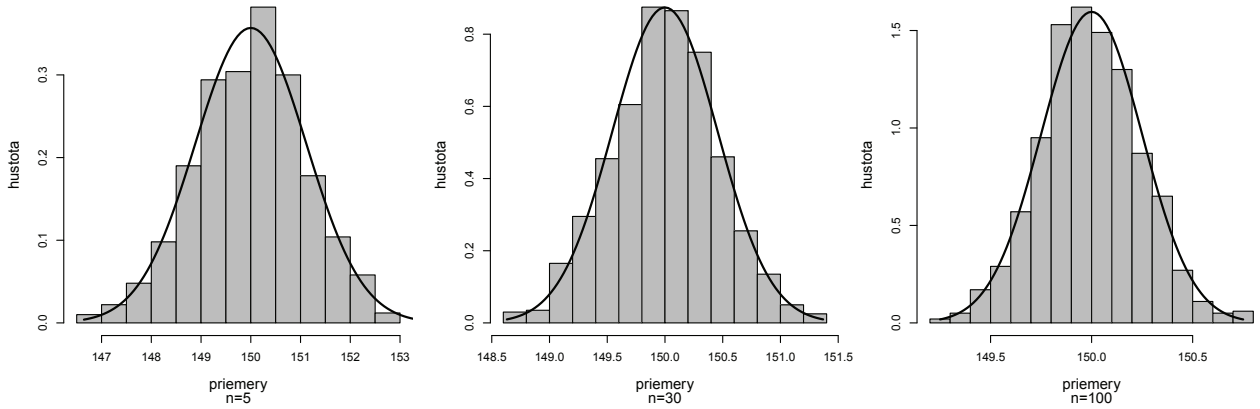
Výsledky sa zhodujú na tri desatinné miesta. Všeobecne platí $X_M \sim N(M/2, M/4)$ a $Y_M = X_M/M \sim N(1/2, 1/(4M))$.

Príklad 85 (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 (1)* $n = 5$*, (2)* $n = 30$ *a (3)* $n = 100$*. Pre každú simuláciu* X *vypočítajte aritmetické priemery* $\overline{x}_m$*,* $m = 1, 2, \ldots, M$*, kde* $M = 1000$*. Superponujte ich histogram v relatívnej škále s teoretickou krivkou hustoty pre* $\overline{X}_n$*. Vypočítajte* $\Pr(\overline{X}_n > 151)$ *pre* $n = 30$ *zo simulovaných dát a porovnajte tento výsledok s teoretickou (očakávanou) pravdepodobnosťou.*

Riešenie (aj v R) (pozri obrázok 13)
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.
$\Pr(\overline{X}_n > 151) = \Pr(\frac{\overline{X}_n - 150}{\sqrt{6.25}/\sqrt{n}} > \frac{151-150}{\sqrt{6.25}/\sqrt{n}}) \approx \Phi(2.190890) = 0.01422987$, kde $n = 30$.

```
M <- 1000; n30 <- 30
X30 <- matrix(0,M,n30)
for (i in 1:M) X30[i,] <- rnorm(n30,150,sqrt(6.25))
x.bar30 <- rowMeans(X30)
# pravdepodobnosti
mean(x.bar30>151) # 0.014238
1-pnorm((151-150)/sqrt(6.25/n30)) # 0.01422987
# obrazok pre n=30
windows(5,5)
par(mar=c(5,4.5,1,1))
hist(x.bar30,probability=TRUE,col="gray",main="",
     ylab="hustota",xlab="priemery",sub="n=30",cex.lab=1.2,cex.sub=1.2)
xmin <- 150-3*sqrt(6.25/n30)
xmax <- 150+3*sqrt(6.25/n30)
curve(dnorm(x,150,sqrt(6.25/n30)),from=xmin,to=xmax,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).



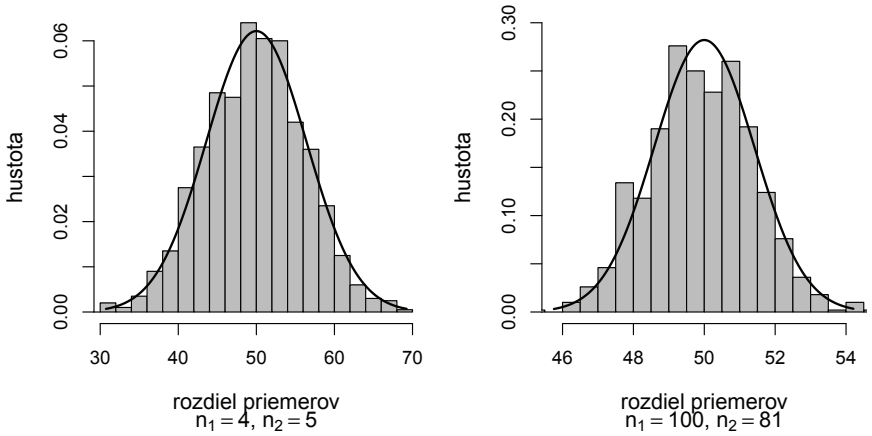
Obr. 13: Histogram vygenerovaných priemerov superponovaný teoretickou krivkou hustoty $\overline{X}_n$

Príklad 86 (normálne rozdelenie, simulačná štúdia) *Nech (a)* $X \sim N(\mu, \sigma^2)$, $\mu = 0$, $\sigma^2 = 1$, *(b)* $X \sim Exp(\lambda)$, $\lambda = 1/3$, *(c)* $X \sim Unif(\min, \max), \min = 0, \max = 1$, *(d)* $X \sim [pN(0,1) + (1-p)N(0,10)]$, *kde* $p = 0.9$. *Použite* R *na simuláciu pseudonáhodných čísel z daných rozdelení – rozsahy náhodných výberov* $n = 2, 5, 20, 50, 100$ *a* 500. *Pre každú simuláciu* X *vypočítajte aritmetické priemery* $\overline{x}_m$, $m = 1, 2, \ldots, M$, *kde* $M = 1000$. *Zobrazte ich do histogramu v relatívnej škále a superponujte ho s teoretickou krivkou hustoty* $N(\mu, \sigma^2/n)$ *prislúchajúcej danej simulácii.*

Príklad 86 slúži na zistenie vlastností rozdelenia výberového 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 87 (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* $\overline{X}_{n_1} - \overline{Y}_{n_2}$.

Príklad 87 slúži na zistenie vlastností rozdelenia rozdielu dvoch výberových priemerov pri rôznych situáciách. 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 14).



Obr. 14: Histogram vygenerovaných rozdielov priemerov superponovaný teoretickou krivkou hustoty rozdelenia $\overline{X}_{n_1} - \overline{Y}_{n_2}$

2.2 *Štatistika

Definícia 15 (štatistika) *Ľubovoľná funkcia* $T(\cdot): \mathcal{Y} \to \mathbb{R}^r$, *pre nejaké* $r \in \mathbb{N}^+$ *náhodného výberu* $(X_1, X_2, \ldots, X_n)^T$, *kde funkcia* $T = T(X_1, X_2, \ldots, X_n)$ *nezávislá na* θ *sa nazýva* ***štatistika*** *a hodnota* $t = T(x_1, x_2, \ldots, x_n)$ *korešpondujúca realizáciám* $x_1, x_2, \ldots, x_n$ *sa nazýva* ***realizácia štatistiky*** *(****pozorovaná hodnota štatistiky****).*

Príklad 88 (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:* $T_1 = \sum_{i=1}^n X_i \in \mathbb{R}$, $T_2 = \sum_{i=1}^n X_i^2 \in \mathbb{R}^+ \cup \{0\}$, $T_3 = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2) \in \mathbb{R}^2$.

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.

Príklad 89 (štatistika; príklady) *(a) Vypočítajte štatistiku $T_1 = \sum_{i=1}^{n} X_i$ pre realizácie náhodnej premennej X najväčšia dĺžka mozgovne (`skull.L`; mm; dáta: `one-sample-mean-skull-mf.txt`). V tomto prípade ide o čitateľ aritmetického priemeru $\overline{x}$. (b) vypočítajte štatistiku $T_2 = \sum_{i=1}^{n} X_i^2$ pre realizácie náhodnej premennej $X = X_1 - X_2$, kde X predstavuje stranový rozdiel vertikálneho priemeru v strede dĺžky tela kľúčnej kosti na pravej aj ľavej strane tela (`length.R` a `length.L`; v mm; dáta: `paired-means-clavicle2.txt`). V tomto prípade ide o sumu štvorcov stranových rozdielov.*

Definícia 16 (postačujúca štatistika) *Pre štatistický model $\mathcal{F}$ je štatistika $T(x)$* ***postačujúca*** *pre parameter θ, ak má rovnakú hodnotu pre dva body rôzne x_1 a x_2 z výberového priestoru $\mathcal{Y}$ iba ak tieto body majú ekvivalentné funkcie vierohodnosti (Azzalini, 1996), t.j. pre*

$$\forall x_1, x_2 \in \mathcal{Y} : T(x_1) = T(x_2) \Rightarrow L(\theta, x_1) \approx L(\theta, x_2) \text{ pre všetky } \theta \in \Theta.$$

Detaily o funkcii vierohodnosti pozri v kapitole 2.3 *Funkcia vierohodnosti. Aj napriek tomu, že nebijektívna transformácia obsahuje „všetku informáciu o dátach“, treba mať na zreteli, že to súvisí s voľbou modelu, t.j. ak sa zmení model, štatistika už nemusí byť postačujúca (Bickel a Doksum, 2006). Preto je vo ľba modelu veľmi dôležitá. Anděl (2011) uvádza inú definíciu postačujúcej štatistiky.

Definícia 17 (postačujúca štatistika) *Štatistika $T(\mathbf{X})$ sa nazýva* ***postačujúca štatistika*** *pre parameter θ, ak podmienené rozdelenie vektora $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ pri danom $T(\mathbf{X})$ nezávisí na θ.*

Predchádzajúca definícia hovorí o tom, že ak $T(\mathbf{X})$ je postačujúca štatistika pre θ, potom každá inferencia o θ závisí na $\mathbf{X}$ len cez hodnotu $T(\mathbf{X})$, t.j. ak $\mathbf{x}$ a $\mathbf{y}$ sú dve realizácie a $T(\mathbf{x}) = T(\mathbf{y})$, potom inferencia o θ bude rovnaká nezávisle na tom, či sme pozorovali $\mathbf{X} = \mathbf{x}$ alebo $\mathbf{Y} = \mathbf{y}$.

Veta 4 (postačujúca štatistika) *Ak $f_1(\mathbf{x}|\theta)$ je združená hustota $\mathbf{X}$ a $f_2(t|\theta)$ je hustota $T(\mathbf{X})$, potom $T(\mathbf{X})$ je postačujúca štatistika pre θ, ak pre všetky $\mathbf{x}$ je podiel $f_1(\mathbf{x}|\theta)/f_2(t|\theta)$ konštanta nezávislá na θ (Casella a Berger, 2002).*

Príklad 90 (postačujúca štatistika binomického rozdelenia) *Nech X_i, $i = 1, 2, \ldots, N$, sú iid Bernoulliho pokusy a $X = \sum_{i=1}^{N} X_i$. Potom $X \sim Bin(N, p)$. Ukážte, že $T(\mathbf{X}) = \sum_{i=1}^{N} X_i$ je postačujúca štatistika pre p.*

Riešenie
$f_1(\mathbf{x}|p)/f_2(t|p) = \prod_{i=1}^{N} p^{x_i}(1-p)^{1-x_i} / \left(\binom{N}{t} p^t (1-p)^{N-t}\right) = 1/\binom{N}{\sum x_i}$. Tento podiel je nezávislý na p, t.j. súčet jednotiek obsahuje všetku informáciu o p, ktorá je v dátach.

Príklad 91 (postačujúca štatistika normálneho rozdelenia) *Nech $X_i \sim N(\mu, \sigma^2)$, kde $i = 1, 2, \ldots, n$, sú iid premenné a σ^2 poznáme. Ukážte, že $T(\mathbf{X}) = \sum_{i=1}^{N} X_i/n = \overline{X}$ je postačujúca štatistika pre μ.*

Riešenie
$f_1(\mathbf{x}|\mu)/f_2(t|\mu) = \prod_{i=1}^{N} f(x_i,\mu)/f_2(t|\mu)$, kde $f(x_i,\mu)$ je hustota normálneho rozdelenia a $f_2(t|\mu) = (2\pi\sigma^2/n)^{-1/2}\exp(-n(\overline{x}-\mu)^2/(2\sigma^2))$. Po viacerých algebraických úpravách dostaneme

$$f_1(\mathbf{x}|\mu)/f_2(t|\mu) = n^{-1/2}(2\pi\sigma^2)^{-(n-1)/2}\exp(-\sum_{i=1}^{n}(x_i-\overline{x})^2/(2\sigma^2)).$$

Tento podiel je nezávislý na μ, t.j. $\overline{X}$ obsahuje všetku informáciu o μ, ktorá je v dátach.

Dá sa ukázať, že ak $\mathrm{X}_i \sim N(\mu,\sigma^2)$, kde $i = 1,2,\ldots,n$, sú iid premenné, potom $T_1(\mathbf{X}) = (\overline{X}, \sum_{i=1}^{n} X_i^2)^T$ a $T_2(\mathbf{X}) = (\overline{X}, \sum_{i=1}^{n}(X_i-\overline{X})^2)^T$ sú postačujúce štatistiky $\boldsymbol{\theta} = (\mu,\sigma^2)^T$.

Špeciálnym prípadom štatistiky je **testovacia štatistika**, ktorá má kľúčovú úlohu pri testovaní hypotéz.

Príklad 92 (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 $p = 0.1, 0.5, 0.9$, 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.*

Príklad 92 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.

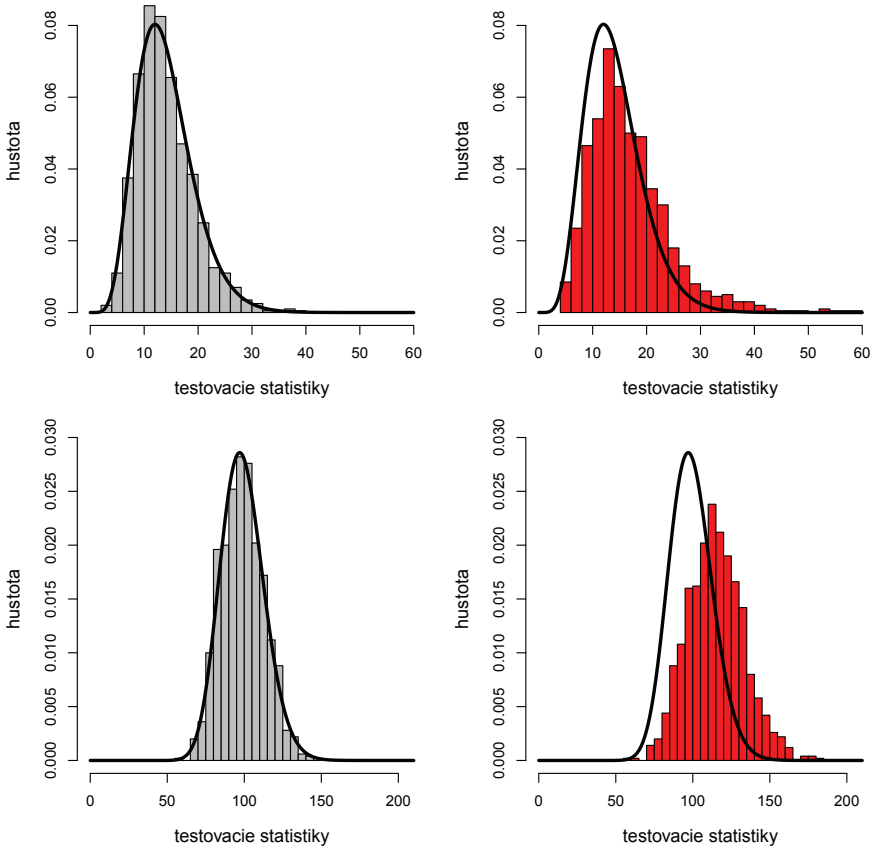
Príklad 93 (testovacia štatistika, simulačná štúdia) *Na základe simulačnej štúdie preverte, že ak (a) $X \sim N(\mu,\sigma^2)$, kde $\mu = 0, \sigma^2 = 1$ a (b) $X \sim [(1-p)N(\mu,\sigma^2) + pN(\mu,\sigma_1^2)]$, kde $p = 0.05$ a $\sigma_1^2 = 2$, potom testovacia štatistika $F = \frac{(n-1)S^2}{\sigma^2}$ má asymptoticky χ^2_{n-1} rozdelenie s $n-1$ stupňami voľnosti. Použité 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.*

Riešenie
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. Vieme, že $E[S^2] = \sigma^2 = 1$ a $Var[S^2] = 2\sigma^4/(n-1) = 2/(n-1)$, $E[F] = n-1$ a $Var[F] = 2(n-1)$, t.j. chceme, aby sa výsledky simulačnej štúdie priblížili týmto teoretickým hodnotám (pozri tabuľku 14).

Tabuľka 14: Teoretické hodnoty stredných hodnôt a rozptylov S^2 a F a ich odhadny zo simulačnej štúdie pri $n = 15$ a $n = 100$

odhady počítané pri simulácii	$E[S^2]$	$Var[S^2]$	$E[F]$	$Var[F]$
teoretické hodnoty, $n = 15$	1.0000	0.1429	14.0000	28.0000
$N(\mu,\sigma^2)$, $n = 15$	1.0003	0.1458	14.0039	28.5763
$X \sim [(1-p)N(\mu,\sigma^2) + pN(\mu,\sigma_1^2)]$, $n = 15$	1.2015	0.2943	16.8213	57.6880
teoretické hodnoty, $n = 15$	1.0000	0.0202	99.0000	198.0000
$N(\mu,\sigma^2)$, $n = 100$	0.9985	0.0198	98.8552	193.7022
$X \sim [(1-p)N(\mu,\sigma^2) + pN(\mu,\sigma_1^2)]$, $n = 100$	1.1596	0.0342	114.8005	335.5359

Pri dostatočne veľkom počte opakovaní vidíme zhodu medzi teoretickým a simulovaným rozdelením F, len ak ide o dáta z normálneho rozdelenia (pozri obrázok 15).



Obr. 15: Histogramy vygenerovaných testovacích štatistík v relatívnej škále superponované s teoretickými krivkami hustoty F; $X \sim N(0,1)$ (ľavý stĺpec) a $X \sim [(1-p)N(\mu,\sigma^2) + pN(\mu,\sigma_1^2)]$ (pravý stĺpec); $n = 15$ (horný riadok), $n = 100$ (dolný riadok)

2.3 *Funkcia vierohodnosti

Funkcia vierohodnosti je najpoužívanejšou funkciou v štatistike. Sumarizuje informácie dostupné z dát v podobe logaritmu, prvej a druhej derivácie. Používa sa nielen pri odhadovaní parametrov rozdelení pravdepodobnosti, ale aj pri testovaní hypotéz a štatistickom modelovaní.

Majme štatistický model $\mathcal{F} = \left\{ f(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k \right\}$. Nech $k = 1$. Ak už bolo x pozorované, hodnota funkcie hustoty $f(\theta, \mathbf{x})$ závisí len od θ. Táto funkcia nám dáva pravdepodobnosť (hustotu) pozorovaní, a priori vo vzťahu k experimentu, ktoré sme predtým pozorovali. Ak chceme porovnať alebo zoradiť $\theta_1, \theta_2 \in \Theta$ podľa dôležitosti, použijeme podiel $f(\mathbf{x}, \theta_1)/f(\mathbf{x}, \theta_2)$, ak existuje. Tento podiel sa nezmení, ak čitateľ a menovateľ vynásobíme nejakou konštantou c nezávislou na θ. Dá sa preto povedať, že $f(\theta, \mathbf{x})$ je vhodná na porovnanie prvkov množiny Θ až na multiplikatívnu konštantu c (Pawitan, 2001).

Definícia 18 (funkcia vierohodnosti) *Pre štatistický model $\mathcal{F}$, na základe ktorého predpokladáme, že $x \in \mathbb{R}$ boli pozorované, použijeme pojem* ***vierohodnosť*** *(**vierohodnostná funkcia**) pre funkciu $\Theta \to \mathbb{R}^+ \cup \{0\}$ definovanú ako (Cox, 2006)*

$$L(\theta, \mathbf{x}) = c(\mathbf{x}) f(\mathbf{x}, \theta),$$

kde $c \in \mathbb{R}$ je nezávislá na θ a $f(\mathbf{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$.

Zápis $L(\theta, \mathbf{x})$ neindikuje závislosť na $\mathbf{x}$, a preto sa často používa zápis $L(\theta|\mathbf{x})$; podobné platí aj pre hustotu, t.j. často sa používa $f(x_i|\theta)$. Tiež je jedno, či píšeme c alebo $c(\mathbf{x})$, keďže vierohodnosť je funkciou θ. V skutočnosti $L(\theta|\mathbf{x})$ definuje triedu funkcií, ktorej prvky sa odlišujú vďaka multiplikatívnej konštante c. Z vyššie uvedenej definície vyplýva, že dva body asociované s proporcionálnymi hustotami determinujú rovnakú vierohodnosť, t.j. ich vierohodnosti sú ekvivalentné. Vierohodnosť ale nie je pravdepodobnosťou. Vierohodnosť je nezáporná a vo väčšine prípadov pozitívna pre všetky θ, preto môžeme definovať **prirodzený logarimus funkcie vierohodnosti** ako (Brazzale a kol., 2007)

$$\ln(L(\theta|\mathbf{x})) = l(\theta|\mathbf{x}) = \ln c + \ln(f(\mathbf{x}|\theta)),$$

kde $l(\theta|\mathbf{x}) = -\infty$ ak $L(\theta|\mathbf{x}) = 0$. V zmysle c ide teda o triedu funkcií.

Definícia 19 (slabý princíp vierohodnosti) *Pre štatistický model $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta\}$ dva body $x_1, x_2 \in \mathbb{R}$, kde platí $L(\theta, x_1) \approx L(\theta, x_2)$, musia viesť k rovnakému záveru (inferenčnému záveru).*

Z definície 19 vyplýva, že všetka informácia o θ je obsiahnutá vo funkcii vierohodnosti pre realizácie x (z nejakého experimentu). Dve funkcie vierohodnosti parametra θ obsahujú rovnakú informáciu o θ, ak sú proporcionálne pre nejaké x (z rovnakých alebo rozdielnych experimentov).

Definícia 20 (silný princíp vierohodnosti) *Majme x_1 z modelu $\mathcal{F}_1 = \{f_1(\cdot;\theta) : \theta \in \Theta\}$ a x_2 z modelu $\mathcal{F}_2 = \{f_2(\cdot;\theta) : \theta \in \Theta\}$, kde $L_1(\theta, x_1) \approx L_2(\theta, x_2)$. Potom body $x_1, x_2 \in \mathbb{R}$ musia viesť k rovnakému záveru (inferenčnému záveru).*

Príklad 94 (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 (Azzalini, 1996)*

$$L_1(\theta|x_1, x_2) = \binom{N}{x_1}\theta^{x_1}(1-\theta)^{x_2}, x_1 = 1, 2, \ldots, N; x_2 = N - x_1,$$

a

$$L_2(\theta|x_1, x_2) = \binom{x_1 + x_2 - 1}{x_1}\theta^{x_1}(1-\theta)^{x_2}; x_1 = 1, 2, \ldots; x_2 = 1, 2, \ldots,$$

kde jadro funkcie vierohodnosti pre oba prípady bude rovné $L_1(\theta|x_1, x_2) = L_2(\theta|x_1, x_2) = \theta^{x_1}(1-\theta)^{x_2}$.

Časť funkcie vierohodnoti zahŕňajúca parameter sa nazýva **jadro** (**kernel**). Kedže maximalizácia funkcie je vo vzťahu k parametru, zvyšok (nejaká konštanta) nezavislý na parametri je pri maximalizácii nepotrebný. Jadro funkcie vierohodnosti je často značené rovnako ako samotná funkcia vierohodnosti (ku ktorej je proporcionálne).

Štatistická teória je kompromis rôznych logicky korektných požiadaviek použitých v kombinácii smerujúci k praktickým potrebám. V praxi slabý princíp vierohodnosti platí takmer vždy, ale silný princíp vierohodnosti iba niekedy (Cox a Donnelly, 2011).

Definícia 21 (maximálne vierohodný odhad) *Maximálne vierohodný odhad parametra θ, ozn. $\widehat{\theta}_{ML} = \widehat{\theta}$ (ozn. ML často neuvádzame a nahrádzame ho slovným spojením „MLE θ je rovné $\widehat{\theta}$“ alebo skrátene „MLE $\widehat{\theta}$“) je taká hodnota parametra θ, ktorá maximalizuje funkciu vierohodnosti $L(\theta|x)$; pozri Cox (2006); Lehmann a Casella (1998).*

Ak $X \sim N(\mu, \sigma^2)$, potom maximálne vierohodnými odhadmi parametrov μ a σ^2 sú $\widehat{\mu} = \overline{x}$ a $\widehat{\sigma}^2 = \frac{(n-1)}{n}s^2$. Ak $X \sim Bin(N, p)$, potom maximálne vierohodný odhad p je $\widehat{p} = x/N$.

Definícia 22 (funkcia vierohodnosti binomického rozdelenia a jej jadro) *Nech X je náhodná premenná, ktorá má binomické rozdelenie s parametrami N a p, t.j. $X \sim Bin(N,p)$. Realizácia X je $x = n$. Potom jadro funkcie vierohodnosti má tvar $L(p|x) = p^x (1-p)^{N-x}$ a jeho logaritmus je rovný $l(p|x) = x \ln p + (N-x) \ln (1-p)$; pozri (Bickel a Doksum, 2006). Binomický koeficient, kombinačné čislo $\binom{N}{x}$, nepíšeme, lebo ho pri maximalizácii nepotrebujeme.*

Definícia 23 (funkcia vierohodnosti multinomického rozdelenia a jej jadro) *Nech $\mathbf{X}$ je náhodná premenná, ktorá má (J-rozmerné) multinomické rozdelenie s parametrami N a $\mathbf{p}$, t.j. $\mathbf{X} \sim Mult_J(N, \mathbf{p})$, kde $\mathbf{X} = (X_1, X_2, \ldots, X_J)^T$. Realizácia X_j je $x_j = n_j$. Funkcia vierohodnosti je proporcionálna ku jadru vierohodnosti $L(\mathbf{p}|\mathbf{x}) = \prod_{j=1}^{J} p_j^{x_j}$ a jej logaritmus $l(\mathbf{p}|\mathbf{x}) = \sum_{j=1}^{J} x_j \ln p_j$ (Casella a Berger, 2002). Konštantu $N!/\prod_j x_j!$ nepíšeme, lebo ju pri maximalizácii nepotrebujeme.*

Definícia 24 (funkcia vierohodnosti Poissonovho rozdelenia a jej jadro) *Nech X je náhodná premenná, ktorá má Poissonovo rozdelenie s parametrom λ, t.j. $X \sim Poiss(\lambda)$. Potom jadro vierohodnoti (Casella a Berger, 2002)*

$$L(\lambda|\mathbf{x}) = \lambda^{\sum_{i=1}^{N} x_i} e^{-N\lambda},$$

a jeho logaritmus $l(\lambda|\mathbf{x}) = \sum_{i=1}^{N} x_i \ln \lambda - N\lambda$. Menovateľ funkcie vierohodnosti $x_1!x_2! \ldots x_N!$ nepíšeme, lebo ho pri maximalizácii nepotrebujeme.

Príklad 95 (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).*

Riešenie
Logaritmus funkcie vierohodnosti $l(\lambda|\mathbf{x}) = \sum_{i=1}^{N} x_i \ln \lambda - N\lambda$, potom $\frac{\partial l(\lambda|\mathbf{x})}{\partial \lambda} = \frac{N\overline{x}}{\lambda} - N$, z čoho vyplýva, že $\widehat{\lambda} = \overline{x}$. Teda ak $N = 5$, 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 jadra funkcie vierohodnosti ako $l(\lambda|\mathbf{x}) = -\lambda \sum_n m_n + \sum_n n m_n \ln \lambda$. Maximálne vierohodný odhad $\widehat{\lambda} = \frac{\sum_n n m_n}{\sum_n m_n}$.

Maximálne vierohodný odhad zjednodušuje pohľad na funkciu vierohodnosti, pretože číslo predstavujúce odhad je jednoduchšie ako funkcia. Všeobecne však jedno číslo (odhad parametra) nie je postačujúce na to, aby reprezentovalo funkciu vierohodnosti. Ak je funkcia vierohodnosti dobre **aproximovaná nejakou kvadratickou funkciou**, potom potrebujeme na jej opis najmenej dve charakteristiky – **polohu maxima** a **zakrivenie v maxime**. Presnejšie potrebujeme aproximáciu logaritmu funkcie vierodnosti **okolo maximálne vierohodného odhadu θ polohy maxima** nejakou kvadratickou funkciou. V tomto prípade nazývame funkciu vierohodnosti *regulárnou*. Prvú deriváciu logaritmu funkcie vierohodnosti nazývame **skóre funkcia** a označujeme ju ako $S(\theta) = \frac{\partial}{\partial \theta} l(\theta|\mathbf{x})$. Z tejto rovnosti je zrejmé, že maximálne vierohodný odhad $\widehat{\theta}$ je riešením **vierohodnostných** (**skóre**) **rovníc** $S(\theta) = 0$. V maxime je druhá derivácie logaritmu funkcie vierohodnosti záporná a zakrivenie v bode $\widehat{\theta}$ bude rovné **Fisherovej miere informácie** $\mathcal{I}(\widehat{\theta})$, kde $\mathcal{I}(\widehat{\theta}) = -\frac{\partial^2}{\partial \theta^2} l(\theta|\mathbf{x})|_{\theta=\widehat{\theta}}$. Veľké zakrivenie zodpovedá strmému a úzkemu vrcholu, čo indikuje menšiu neistotu o θ. $\mathcal{I}(\widehat{\theta})$ nazývame **pozorovaná Fisherova miera informácie**. **Maximálne vierohodný odhad rozptylu** odhadu

parametra θ potom môžeme definovať ako $\widehat{Var[\widehat{\theta}]} = 1/\mathcal{I}(\widehat{\theta})$. **Očakávaná Fisherova miera informácie** je definovaná ako $I(\theta) = E[S^2(\theta)] = Var[S(\theta)] = -E[\frac{\partial}{\partial\theta}S(\theta)]$. Keďže $X_i, i = 1, 2, \ldots, n$, sú nezávislé, potom platí $I(\theta) = ni(\theta)$, kde $i(\theta)$ je Fisherova miera informácie jedného pozorovania.

Príklad 96 ($\mathcal{I}(\widehat{p})$ a rozptyl pre p; $X \sim Bin(N, p)$) *Z funkcie vierohodnosti odvoďte pozorovanú Fisherovu mieru informácie $\mathcal{I}(\widehat{p})$ a rozptyl $\widehat{Var[\widehat{p}]}$.*

Riešenie
$l(p|x) = x \ln p + (N - x) \ln(1 - p) = N\widehat{p} \ln p + N(1 - \widehat{p}) \ln(1 - p)$, potom
$\frac{\partial^2}{\partial p^2} l(p|x) = -(N\widehat{p})/p^2 - N(1 - \widehat{p})/(1 - p)^2$.
Ak dosadíme $p = \widehat{p}$, dostaneme
$\frac{\partial^2}{\partial p^2} l(p|x)|_{p=\widehat{p}} = -(N\widehat{p})/\widehat{p}^2 - N(1 - \widehat{p})/(1 - \widehat{p})^2 = [-N(1 - \widehat{p}) - N\widehat{p}]/[\widehat{p}(1 - \widehat{p})]$
$= -N/[\widehat{p}(1 - \widehat{p})]$, z čoho vyplýva, že

$$\frac{1}{\mathcal{I}(\widehat{p})} = \widehat{Var[\widehat{p}]} = \frac{\widehat{p}(1 - \widehat{p})}{N}.$$

Príklad 97 ($\mathcal{I}(\widehat{\lambda})$ 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.*

Riešenie
$\frac{\partial^2}{\partial\lambda^2} l(\lambda|\mathbf{x}) = -\frac{N\overline{x}}{\lambda^2}$, z čoho po dosadení $\lambda = \widehat{\lambda}$ dostaneme $\frac{1}{\mathcal{I}(\widehat{\lambda})} = \widehat{Var[\widehat{\lambda}]} = \frac{\overline{x}}{N}$, ktorý je rovný 0.56.

Ako vhodný spôsob aproximácie logaritmu funkcie vierohodnosti $l(\theta|\mathbf{x})$ a $l(\boldsymbol{\theta}|\mathbf{x})$ a nejakej funkcie $g(\theta)$ a $g(\boldsymbol{\theta})$, sa javí jednorozmerný a mnohorozmerný **Taylorov rozvoj r-tého rádu**. Táto aproximácia je dostatočne dobrá z hľadiska konvergencie ku funkcii, ktorú aproximujeme.

Definícia 25 (Taylorov rozvoj r-tého rádu) *Ak existujú r-té derivácie funkcie $g(x)$, ozn. $g^{(r)}(x) = \frac{\partial^r}{\partial x^r} g(x)$, potom definujeme Taylorov rozvoj r-tého rádu pre nejakú konštantu a nasledovne (Casella a Berger, 2002)*

$$T_r(x) = \sum_{j=0}^{r} \frac{g^{(j)}(a)}{j!}(x - a)^j.$$

Pri praktických aplikáciách budeme predpokladať, že zvyšok $g(x) - T_r(x)$ bude rýchlo konvergovať k nule (pozri nasledujúcu vetu). Explicitná forma zvyšku nebude dôležitá a budeme ho zanedbávať, pretože nás bude zaujímať iba samotná aproximácia. Jedna z možných podôb zvyšku je nasledovná

$$g(x) - T_r(x) = \int_a^x \frac{g^{(r+1)}(t)}{r!}(x - t)^r dt.$$

Veta 5 (Taylorova veta) *Ak derivácia $g^{(r)}(a) = \frac{\partial r}{\partial x^r} g(x)|_{x=a}$ existuje, potom (Casella a Berger, 2002)*

$$\lim_{x \to a} \frac{g(x) - T_r(x)}{(x - a)^r} = 0.$$

V štatistických aplikáciách Taylorovej vety budeme používať Taylorov rozvoj prvého alebo druhého rádu. Rovnako budeme používať aj mnohorozmerné rozšírenie Taylorovej vety.

Majme **kvadratickú aproximáciu logaritmu funkcie vierohodnosti** pomocou Taylorovho rozvoja druhého rádu okolo $\widehat{\theta}$ definovanú ako

$$l(\theta|\mathbf{x}) \approx l(\widehat{\theta}|\mathbf{x}) + S(\widehat{\theta})(\theta - \widehat{\theta}) - \frac{1}{2}\mathcal{I}(\widehat{\theta})(\theta - \widehat{\theta})^2,$$

z ktorej dostaneme **aproximáciu logaritmu relatívnej (štandardizovanej) funkcie vierohodnosti**

$$\ln \mathcal{L}(\theta|\mathbf{x}) = \ln \frac{L(\theta|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})} = l(\theta|\mathbf{x}) - l(\widehat{\theta}|\mathbf{x}) \approx -\frac{1}{2}\mathcal{I}(\widehat{\theta})(\theta - \widehat{\theta})^2.$$

Posledná rovnosť predstavuje kvadratickú aproximáciu normalizovaného logaritmu funkcie vierohodnosti okolo $\widehat{\theta}$. Na porovnanie skutočnej funkcie vierohodnosti a jej kvadratickej aproximácie tieto dve funkcie nakreslíme do jedného obrázka. Pri zobrazovaní fixujeme maximum logaritmu funkcie vierohodnosti do nuly a rozsah stanovíme od -4 do 0.

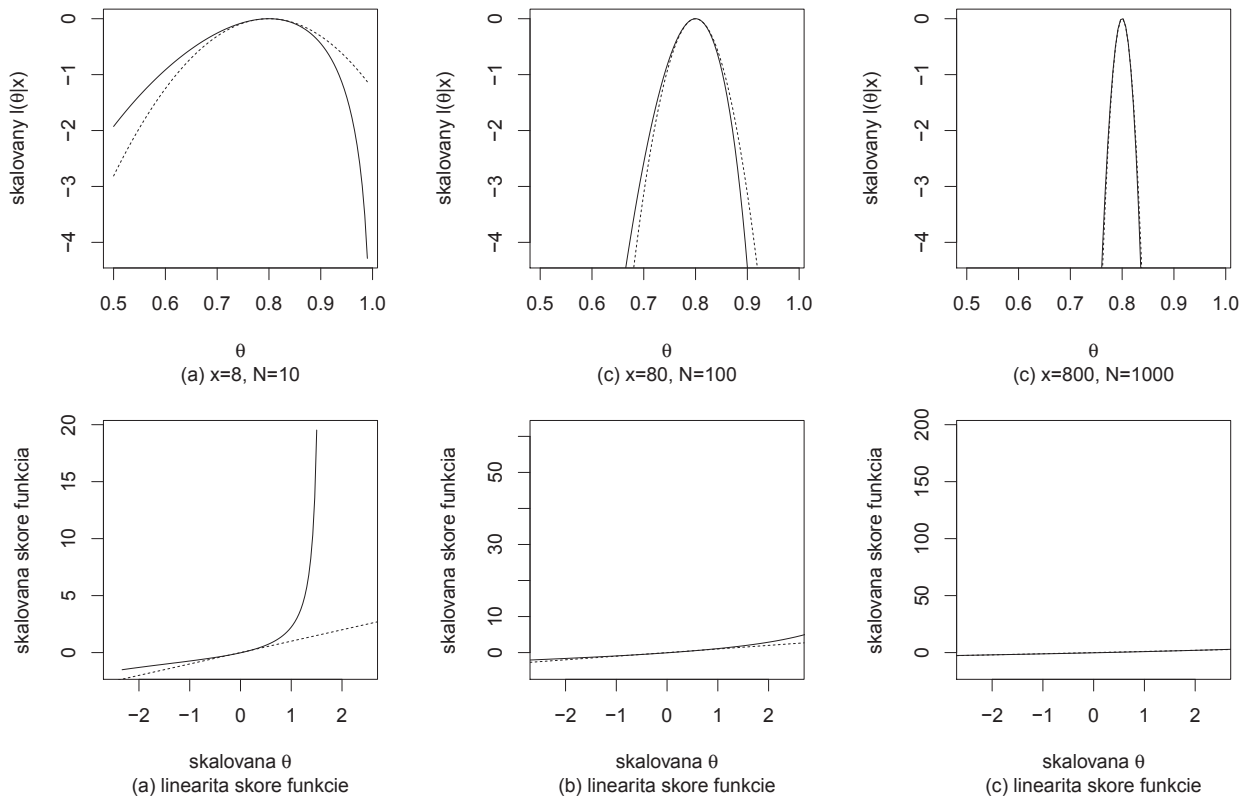
Praktické pravidlo – dostatočne regulárna funkcia vierohodnosti indikuje normalitu X. Alternatívne môžeme zobrať deriváciu kvadratickej aproximácie, kde dostaneme $S(\theta) \approx -\mathcal{I}(\widehat{\theta})(\theta - \widehat{\theta})$ alebo $-\mathcal{I}^{-1/2}(\widehat{\theta})S(\theta) \approx \mathcal{I}^{1/2}(\widehat{\theta})(\theta - \widehat{\theta})$. Posledná rovnosť nie je závislá na mierke θ. Potom kvadratickú aproximáciu môžeme preveriť graficky zobrazením $-\mathcal{I}^{-1/2}(\widehat{\theta})S(\theta)$ oproti $\mathcal{I}^{1/2}(\widehat{\theta})(\theta - \widehat{\theta})$. Za platnosti kvadratickej aproximácie je grafom priamka s jednotkovým sklonom. Pre normálne rozdelené dáta to musí platiť presne. Keďže každá funkcia je lokálne lineárna, je potrebné preveriť, na akom intervale linearitu očakávame. V ideálnom prípade $\mathcal{I}^{1/2}(\widehat{\theta})(\theta - \widehat{\theta}) \sim N(0,1)$, preto kontrolu urobíme aspoň na intervale $\langle -2, 2\rangle$.

Príklad 98 (kvadratická aproximácia funkcie vierohodnosti) *(1) Nakreslite škálovaný logaritmus funkcie vierohodnosti binomického rozdelenia. Na x-ovej osi bude p a na y-ovej osi* $\ln \mathcal{L}(p|\mathbf{x}) = l(p|\mathbf{x}) - \max(l(p|\mathbf{x}))$. *Porovnajte* $\ln \mathcal{L}(p|\mathbf{x})$ *s kvadratickou aproximáciou vypočítanou pomocou Taylorovho rozvoja* $\ln \mathcal{L}(p|\mathbf{x}) = \ln(\frac{L(p|\mathbf{x})}{L(\widehat{p}|\mathbf{x})}) \approx -\frac{1}{2}\mathcal{I}(\widehat{p})(p - \widehat{p})^2$. *(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 16.*

Ak je funkcia vierohodnosti viacrozmerná, je problém ju zobraziť. Ak je $\boldsymbol{\theta}$ dvojrozmerný vektor, potom môžeme $L(\boldsymbol{\theta}|\mathbf{x})$ zobraziť ako kontúrový graf alebo perspektívny trojrozmerný graf v podobe plochy. Nech $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T$. Za predpokladu diferencovateľnosti $L(\boldsymbol{\theta}|\mathbf{x})$ je **skóre funkcia** definovaná ako vektor $S(\boldsymbol{\theta})$, ktorého jednotlivé členy $S(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_i} l(\boldsymbol{\theta}|\mathbf{x}), i = 1, 2, \ldots, k$, a maximálne vierohodný odhad θ_i je riešením vierohodnostných rovníc $S(\boldsymbol{\theta}) = 0$. **Pozorovaná Fisherova informačná matica** druhých derivácií $l(\boldsymbol{\theta}|\mathbf{x})$ má tvar $\mathcal{I}_{ij}(\widehat{\boldsymbol{\theta}})$ je rovný $-\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\boldsymbol{\theta}|\mathbf{x})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$. Maximálne vierohodný odhad kovariančnej matice $\widehat{Var[\widehat{\boldsymbol{\theta}}]} = \mathcal{I}^{-1}(\widehat{\boldsymbol{\theta}})$, t.j. odhad kovariančnej matice $\widehat{Var[\widehat{\boldsymbol{\theta}}]}$ je rovný inverzii pozorovanej Fisherovej informačnej matice $\mathcal{I}(\widehat{\boldsymbol{\theta}})$. **Očakávaná Fisherova informačná matica** je definovaná ako $I(\boldsymbol{\theta}) = E[S(\boldsymbol{\theta})(S(\boldsymbol{\theta}))^T] = Var[S(\boldsymbol{\theta})] = -E[\frac{\partial}{\partial \boldsymbol{\theta}} S(\boldsymbol{\theta})]$. Keďže $X_i, i = 1, 2, \ldots, n$, sú nezávislé, potom platí $I(\boldsymbol{\theta}) = ni(\boldsymbol{\theta})$, kde $i(\boldsymbol{\theta})$ je Fisherova miera informácie jedného pozorovania.

Kvadratická aproximácia logaritmu funkcie vierohodnosti pomocou Taylorovho rozvoja druhého rádu okolo $\widehat{\boldsymbol{\theta}}$ je definovaná ako

$$l(\boldsymbol{\theta}|\mathbf{x}) \approx l(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + S(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T \mathcal{I}(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}).$$



Obr. 16: 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

Pre normálne rozdelené X platí

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \ln \frac{L(\boldsymbol{\theta}|\mathbf{x})}{L(\widehat{\boldsymbol{\theta}}|\mathbf{x})} = l(\boldsymbol{\theta}|\mathbf{x}) - l(\widehat{\boldsymbol{\theta}}|\mathbf{x}) \approx -\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T \mathcal{I}(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}),$$

t.j. funkcia vierohodnosti a jej kvadratická aproximácia sú identické.

Definícia 26 (funkcia vierohodnosti normálneho rozdelenia) *Nech $X_1, X_2, \ldots, X_n$ sú nezávislé rovnako rozdelené premenné, $X_i \sim N(\mu, \sigma^2)$, kde $i = 1, 2, \ldots, n$, $\boldsymbol{\theta} = (\mu, \sigma^2)^T \in \Theta = \mathbb{R} \times \mathbb{R}^+$. Vďaka nezávislosti X_i dostaneme*

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right)\right) \end{aligned}$$

a korešpondujúci logaritmus

$$l(\boldsymbol{\theta}|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right).$$

Príklad 99 ($\mathcal{I}(\widehat{\boldsymbol{\theta}})$ pre vektor $\boldsymbol{\theta} = (\mu, \sigma^2)^T$; $X_i \sim N(\mu, \sigma^2)$) *Nech $X_i \sim N(\mu, \sigma^2)$, kde $i = 1, 2, \ldots, n$. Čomu je rovná pozorovaná Fisherova informačná matica $\mathcal{I}(\widehat{\boldsymbol{\theta}})$, kde $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma}^2)^T$?*

Riešenie
Logaritmus funkcie vierohodnosti má tvar

$$l(\boldsymbol{\theta}|\mathbf{x}) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Derivácie funkcie vierohodnosti v μ a σ^2 budú nasledovné

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

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

Potom

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

Príklad 100 ($\mathcal{I}(\widehat{\mathbf{p}})$ 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 $\widehat{Var[\widehat{\mathbf{p}}]}$.*

Riešenie
Označme $p_J = 1 - \sum_{j=1}^{J-1} p_j$ a predefinujme J-rozmerný vektor $\mathbf{p}$ na $(J-1)$-rozmerný vektor $\mathbf{p} = (p_1, p_2, \ldots, p_{J-1})^T$. Potom $l(\mathbf{p}|\mathbf{x}) = \sum_{j=1}^{J-1} n_j \ln p_j + n_J \ln(1 - \sum_{j=1}^{J-1} p_j)$ a $\frac{\partial}{\partial p_j} l(\mathbf{p}|\mathbf{x}) = \frac{n_j}{p_j} - \frac{n_J}{p_J}$, ktoré tvoria $S(\mathbf{p})$. Fisherovu informačnú maticu dostaneme nasledovne

$$\mathcal{I}(\mathbf{p}) = -\frac{\partial}{\partial\mathbf{p}}S(\mathbf{p}) = \operatorname{diag}\left(\frac{n_1}{p_1^2}, \frac{n_2}{p_2^2}, \ldots, \frac{n_{J-1}}{p_{J-1}^2}\right) + \frac{n_J}{p_J^2}\mathbf{1}\mathbf{1}^T,$$

kde $\mathbf{1}$ je $(J-1)$-rozmerný vektor jednotiek. Potom pozorovaná Fisherova informačná matica bude rovná

$$\mathcal{I}(\widehat{\mathbf{p}}) = N\left(\operatorname{diag}\left(\frac{1}{\widehat{p}_1}, \frac{1}{\widehat{p}_2}, \ldots, \frac{1}{\widehat{p}_{J-1}}\right) + \frac{\mathbf{1}\mathbf{1}^T}{\widehat{p}_J}\right) = N\begin{pmatrix} \frac{1}{\widehat{p}_1} + \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_J} & \cdots & \frac{1}{\widehat{p}_J} \\ \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_2} + \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_J} & \cdots & \frac{1}{\widehat{p}_J} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_J} & \cdots & \frac{1}{\widehat{p}_J} & \frac{1}{\widehat{p}_{J-1}} + \frac{1}{\widehat{p}_J} \end{pmatrix}$$

a

$$\widehat{Var[\widehat{\mathbf{p}}]} = \mathcal{I}^{-1}(\widehat{\mathbf{p}}) = \frac{1}{N}\left(\operatorname{diag}(\widehat{\mathbf{p}}) - \widehat{\mathbf{p}}\widehat{\mathbf{p}}^T\right) = \frac{1}{N}\begin{pmatrix} \widehat{p}_1(1-\widehat{p}_1) & -\widehat{p}_1\widehat{p}_2 & \ldots & -\widehat{p}_1\widehat{p}_{J-1} \\ -\widehat{p}_2\widehat{p}_1 & \widehat{p}_2(1-\widehat{p}_2) & \ldots & -\widehat{p}_2\widehat{p}_{J-1} \\ \vdots & \vdots & \vdots & \vdots \\ -\widehat{p}_{J-1}\widehat{p}_1 & -\widehat{p}_{J-1}\widehat{p}_2 & \ldots & \widehat{p}_{J-1}(1-\widehat{p}_{J-1}) \end{pmatrix}.$$

Ak do $\widehat{Var[\widehat{\mathbf{p}}]}$ pridáme jeden riadok a jeden stĺpec zodpovedajúce $\widehat{p}_J$, dostaneme singulárnu kovariančnú maticu J-rozmerného vektora $\widehat{\mathbf{p}}$.

Profilová vierohodnosť. Aj napriek tomu, že funkcia vierohodnosti je často viacrozmerná, je jednoduchšie ju zobrazovať pre každý parameter θ_i zvlášť alebo pre nejakú podmnožinu parametrov vektora $\boldsymbol{\theta}$. Napr. pri modeli normálneho rozdelenia nás zaujíma len stredná hodnota μ, pričom rozptyl σ^2 je tzv. rušivý parameter (potrebný kvôli adaptácii modelu na variabilitu v dátach). Potrebujeme teda metódu, ktorá koncentruje vierohodnosť len na parameter záujmu eliminovaním rušivého parametra. Vierohodnostný prístup na elimináciu rušivého parametra pozostáva zo substitúcie jeho maximálne vierohodným odhadom v každej fixovanej hodnote parametra záujmu. Výsledkom je **profilová vierohodnostná funkcia**.

Zakrivenie profilovej vierohodnosti. Zakrivenie profilovej funkcie vierohodnosti súvisí s Fisherovou informačnou maticou. Ak napr. parametrom záujmu je θ_1 z vektora $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$, potrebujeme $\mathcal{I}(\widehat{\boldsymbol{\theta}})$ a jej inverziu $\mathcal{I}^{-1}(\widehat{\boldsymbol{\theta}})$ v nasledovnom tvare

$$\mathcal{I}(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}, \mathcal{I}^{-1}(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}.$$

Potom zakrivenie profilovej funkcie vierohodnosti v $\widehat{\theta}_1$ nie je I_{11} ale $(I^{11})^{-1}$, kde $(I^{11})^{-1}$ je vo všeobecnosti menšie ako I_{11}. Interpretácia je nasledovná – informačné číslo I_{11} je zakrivením profilovej funkcie vierohodnosti v $\widehat{\theta}_1$, kde o θ_2 sa predpokladá, že je známe v pozorovanom odhade $\widehat{\theta}_2$; avšak $(I^{11})^{-1}$ je zakrivením profilovej funkcie vierohodnosti v $\widehat{\theta}_1$, ktoré berie do úvahy, že θ_2 je neznáme. Z toho je potom zrejmé, že $(I^{11})^{-1}$ je menšie ako I_{11}. Na základe vyššie uvedeného môžeme kvadraticky aproximovať logaritmus profilovej funkcie vierohodnosti použitím $\widehat{\theta}_i$ a $(I^{ii})^{-1}$, kde

$$\mathcal{L}(\theta_i|\mathbf{x}) = \ln \frac{L(\theta_i|\mathbf{x})}{L(\widehat{\theta}_i|\mathbf{x})} = l(\theta_i|\mathbf{x}) - l(\widehat{\theta}_i|\mathbf{x}) \approx -\frac{1}{2}(I^{ii})^{-1}(\theta_i - \widehat{\theta}_i)^2.$$

Podobným spôsobom je možné kvadraticky aproximovať aj plochu vierohodnosti.

Invariantnosť maximálne vierohodného odhadu. Invariantnosť maximálne vierohodného odhadu znamená, že ak $\widehat{\theta}$ je maximálne vierohodný odhad θ (t.j. $\widehat{\theta} = \widehat{\theta}_{ML}$) a $g(\theta)$ je funkciou θ, potom $g(\widehat{\theta})$ je tiež maximálne vierohodný odhad $g(\theta)$. Maximálne vierohodný odhad rozptylu $g(\theta)$ potom môžeme definovať ako $\widehat{Var[g(\widehat{\theta})]} = [\frac{\partial}{\partial\theta} g(\theta)|_{\theta=\widehat{\theta}}]^2 / \mathcal{I}(\widehat{\theta})$. V prípade vektora $\boldsymbol{\theta}$ je $\widehat{Var[\mathbf{g}(\widehat{\boldsymbol{\theta}})]} = \Delta^T \mathcal{I}^{-1}(\widehat{\boldsymbol{\theta}})\Delta$, kde $\Delta = \frac{\partial}{\partial\boldsymbol{\theta}} \mathbf{g}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$. Tento vzorec vychádza z **delta metódy**, ktorá je postavená na Taylorovom rozvoji prvého rádu (t.j. jeho lineárnej zložky) okolo bodu $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, kde

$$\mathbf{g}(\widehat{\boldsymbol{\theta}}) \approx \mathbf{g}(\boldsymbol{\theta}) + \sum_{i=1}^{k} \left(\frac{\partial}{\partial\theta_i} \mathbf{g}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right) (\widehat{\theta}_i - \theta_i) = \mathbf{g}(\boldsymbol{\theta}) + \Delta^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

potom

$$\widehat{Var[\mathbf{g}(\widehat{\boldsymbol{\theta}})]} \approx \sum_{i=1}^{k} \left(\frac{\partial}{\partial\theta_i} \mathbf{g}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right)^2 \widehat{\sigma}_i^2 + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left(\frac{\partial}{\partial\theta_i} \mathbf{g}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right) \left(\frac{\partial}{\partial\theta_j} \mathbf{g}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right) \widehat{\sigma}_{ij} = \Delta^T \widehat{\boldsymbol{\Sigma}} \Delta,$$

kde $\widehat{\boldsymbol{\Sigma}} = \widehat{Var[\widehat{\boldsymbol{\theta}}]}$, $\widehat{\sigma}_i^2 = \widehat{Var[\widehat{\theta}_i]}$, $\widehat{\sigma}_{ij} = \widehat{Cov[\widehat{\theta}_i, \widehat{\theta}_j]}$ a $i \neq j; i, j = 1, 2, \ldots, k$. Δ je matica $k \times k_1$, kde derivovanie prebieha po zložkách, t.j. (i, j)-ty element Δ je rovný $\frac{\partial g_j(\boldsymbol{\theta})}{\partial\theta_i}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, $\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), \ldots, g_{k_1}(\boldsymbol{\theta}))^T$, $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, k_1 \leq k$. V praxi sa často vyskytuje situácia $k \neq 1$ a $k_1 = 1$, napr. $\boldsymbol{\theta} = (p_1, p_2)^T$ a $g(\boldsymbol{\theta}) = \frac{p_1}{p_2}$. Ak $k = k_1 = 1$, potom

$$\widehat{Var[g_1(\widehat{\theta}_1)]} = \widehat{Var[g(\widehat{\theta})]} = \left(\frac{\partial}{\partial\theta} g(\theta)|_{\theta=\widehat{\theta}} \right)^2 \widehat{\sigma}^2,$$

kde $\widehat{\sigma}^2 = \widehat{Var[\widehat{\theta}]}$.

Príklad 101 (profilová vierohodnosť; normálne rozdelenie) *Profilová funkcia vierohodnosti pre μ je rovná $L(\mu|\mathbf{x}) = c\exp\left(-n\widehat{\sigma}^2_\mu/(2\widehat{\sigma}^2)\right)$, kde c je nejaká konštanta a $\widehat{\sigma}^2_\mu = \frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2$, t.j. ide o rez $L((\mu,\sigma^2)^T|\mathbf{x})$ v bode $\sigma^2 = \widehat{\sigma}^2$. Profilová funkcia vierohodnosti pre σ^2 je rovná $L(\sigma^2|\mathbf{x}) = c\,(\sigma^2)^{-n/2}\exp\left(-n\widehat{\sigma}^2/(2\sigma^2)\right)$, kde c je nejaká konštanta a $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2$.*

Príklad 102 (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).*

2.4 *Maximalizácia funkcie vierohodnosti

Maximálne vierohodný odhad $\widehat{\theta}$ je možné vypočítať pomocou metód numerickej optimalizácie (pozri napr. Horová a Zelinka, 2008) aplikovaných na funkciu vierohodnosti $L(\theta|\mathbf{x})$ alebo jej logaritmus $l(\theta|\mathbf{x})$.

Newtonova (Newton-Raphsonova) metóda (metóda dotyčníc) je pomenovaná po Isaacovi Newtonovi (1643−1727) a Josephovi Raphsonovi (1648−1715). Majme kvadratickú aproximáciu logaritmu funkcie vierohodnosti pomocou Taylorovho rozvoja druhého rádu okolo nejakého bodu θ_0 definovanú ako

$$l(\theta|\mathbf{x}) \approx l(\theta_0|\mathbf{x}) + S(\theta_0)(\theta-\theta_0) - \frac{1}{2}\mathcal{I}(\theta_0)(\theta-\theta_0)^2$$

alebo lineárnu aproximácu skóre funkcie pomocou Taylorovho rozvoja prvého rádu

$$S(\theta) \approx S(\theta_0) - \mathcal{I}(\theta_0)(\theta-\theta_0).$$

Z tejto aproximácie môžeme odvodiť nasledovnú iteračnú funkciu

$$\theta_0 + \frac{S(\theta_0)}{\mathcal{I}(\theta_0)}.$$

Postup je nasledovný:

1. inicializácia metódy použítím vhodne zvoleného štartovacieho parametra $\theta^{(0)}$, pre ktorý platí $\mathcal{I}(\theta^{(0)}) \neq 0$,

2. iterácia rovnosti

$$\theta^{(i)} = \theta^{(i-1)} + \frac{S(\theta^{(i-1)})}{\mathcal{I}(\theta^{(i-1)})},$$

kde $\mathcal{I}(\theta^{(i-1)}) \neq 0$, pre $i = 1,2,\ldots$, pokiaľ nebude $|l(\theta^{(i)}|\mathbf{x}) - l(\theta^{(i-1)}|\mathbf{x})| < \epsilon$, kde ϵ je vhodne zvolené malé číslo (prahová hodnota).

Newton-Raphsonova má jednoduchú geometrickú interpretáciu – bod $\theta^{(i)}$ je priesečník dotyčnice ku grafu skóre funkcie $S(\cdot)$ v bode $[\theta^{(i-1)}, S(\theta^{(i-1)})]$ s x-ovou osou. Táto metóda konverguje kvadraticky, t.j. počet správnych cifier odhadu sa v každom iteračnom kroku zdvojnásobí (rád metódy je

rovný dvom). Konvergencia k lokálnemu extrému však nie je zaručená – metóda môže konvergovať k lokálnemu minimu alebo divergovať, ak začneme iterácie v $\theta^{(0)}$ z konvexnej časti $l(\theta|\mathbf{x})$. Aj keď je funkcia konkávna, konvergencia k lokálnemu maximu nie je zaručená (metóda nerozlišuje lokálne maximum a minimum, pretože rieši rovnosť $S(\theta) = 0$). Ak je funkcia záujmu multimodálna, nemôžeme očakávať, že metóda bude konvergovať ku globálnemu extrému. Ak je $S(\theta)$ dvakrát diferencovateľná, prvá a druhá derivácia nemenia znamienko na intervale $\langle\theta_D, \theta_H\rangle$, funkcia $S(\theta)$ má jednoduchý koreň ($S'(\theta) \neq 0$ pre každé $\theta \in \langle\theta_D, \theta_H\rangle$) a štartovací bod $\theta^{(0)}$ je ten z krajných bodov θ_D, θ_H, v ktorom je znamienko $S(\cdot)$ rovnaké ako znamienko jej druhej derivácie na intervale $\langle\theta_D, \theta_H\rangle$, potom metóda konverguje.

Newton-Raphsonova metóda je implementovaná v R vo funkcii `optimize(f,interval,maximum=FALSE,tol,...)`, kde vstupným argumentom je buď funkcia vierohodnosti alebo jej logaritmus (`f`), štartovací interval (`interval`) a zvolená prahová hodnota (`tol`). Vo všeobecnosti nie je potrebné do funkcie pridávať argument obsahujúci prvú deriváciu funkcie záujmu, pretože je vypočítaná numericky. V prípade zadania argumentu `hessian=TRUE`, vo výsledkoch sa objaví aj $-\mathcal{I}(\widehat{\theta})$.

Alternatívnymi, avšak pomalšími, metódami sú **metóda zlatého rezu** a **metóda sukcesívnej parabolickej interpolácie**. V prvej z nich, kde derivácia funkcie záujmu nie je potrebná, sa interval $\left\langle\theta_D^{(i-1)}, \theta_H^{(i-1)}\right\rangle$, v ktorom leží maximum $l(\theta|\mathbf{x})$, v každom kroku delí v pomere zlatého rezu. Interval sa teda zužuje o $(3-\sqrt{5})/2 \doteq 0.382$ (t.j. komplement zlatého rezu) jeho dĺžky, pričom deliaci bod intervalu je $\theta_{zr}^{(i-1)}$. Metóda zlatého rezu konverguje lineárne (rád metódy je rovný jednej), t.j. chyba sa zmenšuje v každom iteračnom kroku $(1-(3-\sqrt{5})/2)$-krát. Druhou metódou sa dopĺňa vo funkcii `optimize()` prvá v čase, keď je interval $\left\langle\theta_D^{(i-1)}, \theta_H^{(i-1)}\right\rangle$ príliž úzky. Bodmi $\theta_D^{(i-1)}, \theta_{zr}^{(i-1)}, \theta_H^{(i-1)}$ je preložená parabola (kvadratický interpolačný polynóm) a jej maximum bude novým bodom $\theta^{(i)}$. Prednastavené je hľadanie minima (`maximum=FALSE`) v štartovacom intervale (`interval`), čo môže byť zmenené na maximum (`maximum=TRUE`). Vo výstupoch funkcie `optimize()` bude $\widehat{\theta}$ (`minimum` alebo `maximum`) a $l(\widehat{\theta}|\mathbf{x})$ (`objective`).

Majme kvadratickú aproximáciu logaritmu funkcie vierohodnosti pomocou Taylorovho rozvoja druhého rádu okolo nejakého bodu $\boldsymbol{\theta}_0$ definovanú ako

$$l(\boldsymbol{\theta}|\mathbf{x}) \approx l(\boldsymbol{\theta}_0|\mathbf{x}) + S(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathcal{I}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$$

alebo lineárnu aproximáciu skóre funkcie pomocou Taylorovho rozvoja prvého rádu

$$S(\boldsymbol{\theta}) \approx S(\boldsymbol{\theta}_0) - \mathcal{I}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0).$$

Z tejto aproximácie môžeme odvodiť nasledovnú iteračnú funkciu

$$\boldsymbol{\theta}_0 + (\mathcal{I}(\boldsymbol{\theta}_0))^{-1}S(\boldsymbol{\theta}_0).$$

Postup je nasledovný:

1. inicializácia metódy použítím vhodne zvoleného štartovacieho parametra $\boldsymbol{\theta}^{(0)}$, pre ktorý platí $\mathcal{I}(\boldsymbol{\theta}^{(0)}) \neq \mathbf{0}$,

2. iterácia rovnosti
$$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)} + (\mathcal{I}(\boldsymbol{\theta}^{(i-1)}))^{-1}S(\boldsymbol{\theta}^{(i-1)}),$$
$\mathcal{I}(\boldsymbol{\theta}^{(i-1)}) \neq \mathbf{0}$, pre $i = 1, 2, \ldots$, pokiaľ nebude $|l(\boldsymbol{\theta}^{(i)}|\mathbf{x}) - l(\boldsymbol{\theta}^{(i-1}|\mathbf{x})| < \epsilon$, kde ϵ je vhodne zvolené malé číslo (prahová hodnota). Vo všeobecnosti sa $-\mathcal{I}(\boldsymbol{\theta}^{(i-1)}) = l''(\boldsymbol{\theta}^{(i-1)}|\mathbf{x})$ nazýva *hesián*.

Namiesto $(\mathcal{I}(\boldsymbol{\theta}^{(i-1)}))^{-1}$ je lepšie použiť riešenie systému rovníc $(\mathcal{I}(\boldsymbol{\theta}^{(i-1)}))\mathbf{z}_{i-1} = S(\boldsymbol{\theta}^{(i-1)})$ pre nejaké $\mathbf{z}$ a potom $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)} + \mathbf{z}_{i-1}$. V niektorých štatistických modeloch (napr. v logistickom regresnom modeli) sa namiesto $\mathcal{I}(\boldsymbol{\theta}^{(i-1)})$ používa $I(\boldsymbol{\theta}^{(i-1)})$, ktorá má často jednoduchší tvar. Potom hovoríme

o **Fisherovej skóringovej metóde**. Ak namiesto $\mathcal{I}(\boldsymbol{\theta}^{(i-1)})$ použijeme jej pozitívne definitnú aproximáciu počítanú pomocou sukcesívne počítaných gradientov, hovoríme o **quasi Newtonovej metóde** (v angličtine nazývanej aj *variable metric method*). Fisherova skóringová metóda je potom vlastne quasi Newtonova metóda. Gradient nemusí byť špecifikovaný ako funkcia, ale môže byť počítaný numericky, napr. *centrálnou rozdielovou aproximáciou*, ako

$$\frac{\partial}{\partial\theta_i}l(\boldsymbol{\theta}|\mathbf{x}) \approx \frac{l(\boldsymbol{\theta}+\epsilon\mathbf{e}_i|\mathbf{x}) - l(\boldsymbol{\theta}-\epsilon\mathbf{e}_i|\mathbf{x})}{2\epsilon}, \text{ kde } i = 1, 2, \ldots, k,$$

i-ty komponent bazálneho vektora $\mathbf{e}_i$ obsahuje jednotku a na ostatných miestach sú nuly a ϵ je malé číslo. Derivácia v i-tom smere je potom pre $\epsilon \to 0$ nahradená jej konečnou aproximáciou. Quasi Newtonova metóda je implementovaná v R vo funkcii `optim(par,fn,gr,method,control,hessian =FALSE,...)`, kde gradient môže byť špecifikovaný voliteľným argumentom `gr` (prednastavená je metóda spomenutá vyššie[23]). Populárnou metódou aproximácie hesiánu je **Broyden-Fletcher-Goldfarb-Shannova (BFGS) métoda**, kde

$$l''(\boldsymbol{\theta}^{(i)}|\mathbf{x}) \approx l''(\boldsymbol{\theta}^{(i-1)}|\mathbf{x}) + \frac{\mathbf{y}^{(i-1)}(\mathbf{y}^{(i-1)})^T}{(\mathbf{y}^{(i-1)})^T\mathbf{s}^{(i-1)}} - \frac{l''(\boldsymbol{\theta}^{(i-1)}|\mathbf{x})\mathbf{s}^{(i-1)}(l''(\boldsymbol{\theta}^{(i-1)}|\mathbf{x})\mathbf{s}^{(i-1)})^T}{(\mathbf{s}^{(i-1)})^T l''(\boldsymbol{\theta}^{(i-1)}|\mathbf{x})\mathbf{s}^{(i-1)}},$$

kde $\mathbf{y}^{(i-1)} = S(\boldsymbol{\theta}^{(i)}) - S(\boldsymbol{\theta}^{(i-1)})$ a $\mathbf{s}^{(i-1)} = \boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}^{(i-1)} = (\mathcal{I}(\boldsymbol{\theta}^{(i-1)}))^{-1}S(\boldsymbol{\theta}^{(i-1)})$. BFGS metóda je implementovaná vo funkcii `optim()` pri nastavení argumentu `method="BFGS"`. Strata kvadratickej konvergencie oproti Newtonovej metóde je spôsobená aproximáciou hesiánu. BFGS metóda patrí do tzv. *Broydenovej triedy*, kde je hodnota $l''(\boldsymbol{\theta}^{(i)}|\mathbf{x})$ modifikovaná štvrtým členom, ktorý je pri BFGS nulový (pozri Givens a Hoeting, 2005).

Vo funkcii `optim()` je prednastavená **Nelder-Meadova metóda** (nazývaná aj **metóda simplexov**; argument `method="Nelder-Mead"`), ktorá nepoužíva gradient. Je robustná na nespojité funkcie, ale konverguje pomaly. Je vytvorená na základe myšlienky „preskokov“ cez trojuholníky. V každom kroku majme trojuholník definovaný troma bodmi $\boldsymbol{\theta}_1^{(i-1)}$, $\boldsymbol{\theta}_2^{(i-1)}$, $\boldsymbol{\theta}_3^{(i-1)}$, kde platí $l(\boldsymbol{\theta}_1^{(i-1)}|\mathbf{x}) < l(\boldsymbol{\theta}_2^{(i-1)}|\mathbf{x}) < l(\boldsymbol{\theta}_3^{(i-1)}|\mathbf{x})$ a snažíme sa $\boldsymbol{\theta}_1^{(i-1)}$ nahradiť „lepším“ bodom $\boldsymbol{\theta}_1^{(i)}$, pre ktorý platí $l(\boldsymbol{\theta}_1^{(i)}|\mathbf{x}) > l(\boldsymbol{\theta}_1^{(i-1)}|\mathbf{x})$. Ak je tak možné urobiť, nový bod definujeme pomocou stredovej súmernosti a extrapolácie ako

$$\boldsymbol{\theta}_1^{(i)} = \boldsymbol{\theta}_{23}^{(i-1)} + 2\left(\boldsymbol{\theta}_{23}^{(i-1)} - \boldsymbol{\theta}_1^{(i-1)}\right),$$

kde $\boldsymbol{\theta}_{23}^{(i-1)} = \frac{\boldsymbol{\theta}_2^{(i-1)}+\boldsymbol{\theta}_3^{(i-1)}}{2}$. Z vyššie uvedeného vyplýva, že bod $\boldsymbol{\theta}_1^{(i-1)}$zobrazujeme cez stred úsečky s krajnými bodmi $\boldsymbol{\theta}_2^{(i-1)}$ a $\boldsymbol{\theta}_3^{(i-1)}$ do bodu $\boldsymbol{\theta}_{23}^{(i-1)} + \left(\boldsymbol{\theta}_{23}^{(i-1)} - \boldsymbol{\theta}_1^{(i-1)}\right)$ a potom postupujeme po tejto polpriamke ešte ďalej do bodu $\boldsymbol{\theta}_1^{(i)}$. Ak $l(\boldsymbol{\theta}_1^{(i)}|\mathbf{x}) > l(\boldsymbol{\theta}_1^{(i-1)}|\mathbf{x})$, potom trojuholník nahradíme novým trojuholníkom definovaným bodmi $\boldsymbol{\theta}_1^{(i)}$, $\boldsymbol{\theta}_2^{(i-1)}$, $\boldsymbol{\theta}_3^{(i-1)}$.

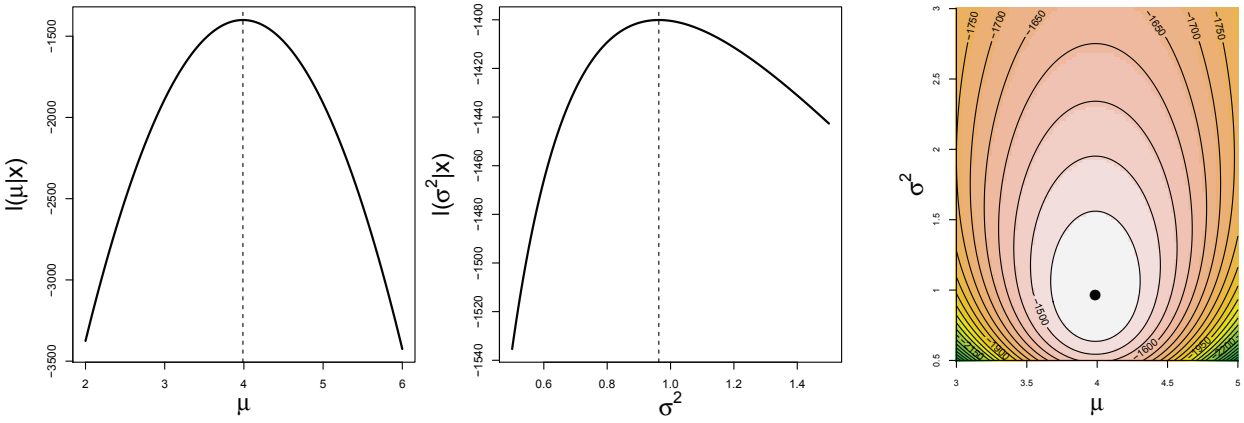
Pri maximalizácii logaritmu funkcie vierohodnosti je vo vstupe funkcie `optim()` potrebné nastaviť `hessian=TRUE` a argument `control=list(fnscale=-1)`, ktorý robí maximalizáciu namiesto prednastavenej minimalizácie (v argumente `control` je možné nastaviť aj maximálne množstvo iterácií pomocou `maxit`). Argument `par` predstavuje štartovaciu hodnotu parametra. Vo výstupoch funkcie `optim()` bude $\widehat{\boldsymbol{\theta}}$ (`par`), $l(\widehat{\boldsymbol{\theta}}|\mathbf{x})$ (`value`) a $l''(\widehat{\boldsymbol{\theta}}|\mathbf{x})$ (`hessian`). Ak je vo výstupe `convergence` rovné nule, potom maximalizácia skonvergovala (k lokálnemu maximu, ktoré nemusí byť globálne).

Príklad 103 (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 $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ a preverte, či je maximálne vierohodný odhad $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ dostatočne blízko*

[23]Centrálna rozdielová aproximácia hrá kľúčovú úlohu pri aproximácii hesiánu, ktorý sa používa v maximálnej vierohodnosti na odhad rozptylu.

k jeho skutočnej hodnote. (c) Nakreslite graf $l((\mu,\sigma^2)^T|\mathbf{x})$ použitím funkcie `image()` *a superponujte ho s kontúrovým grafom použitím funkcie* `contour()`*. Zvýraznite polohu maxima.*

Riešenie (pozri obrázok 17)
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 $\mu_1 = 4, \sigma \in (0.5, 1.5)$;
$l((\mu,\sigma^2)^T|\mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}$, kde $\mu \in (2,6)$ a $\sigma \in (0.5, 1.5)$.
Výsledky simulácie: $\widehat{\mu} = 4.019708$ a $\widehat{\sigma}^2 = 1.000038$.



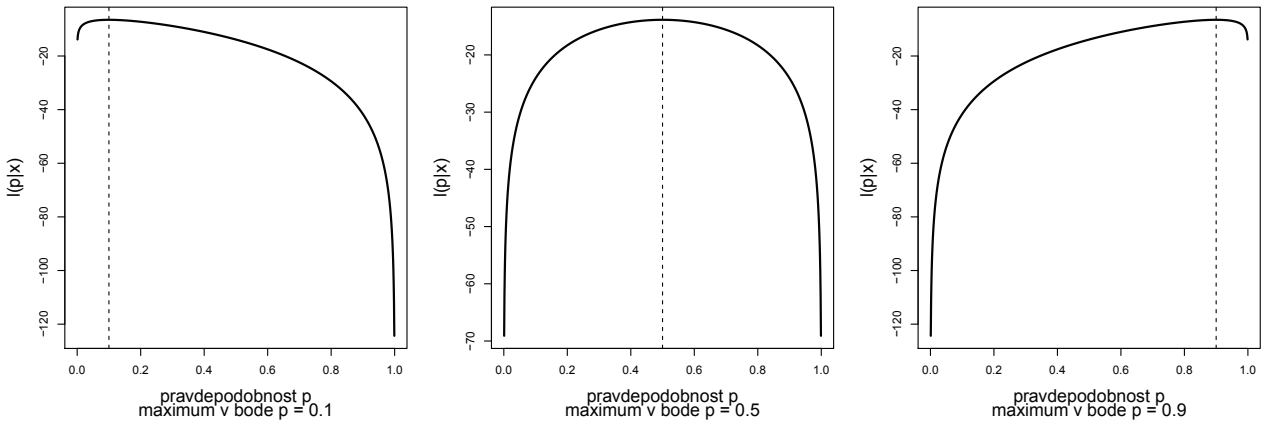
Obr. 17: Profilová funkcia vierohodnosti pre μ (vľavo), σ^2 (uprostred) a funkcia vierohodnosti pre oba parametre (vpravo); $X \sim N(4,1)$); maximálne vierohodné odhady strednej hodnoty a rozptylu sú označené zvislou čiarkovanou čiarou (vľavo a uprotred) a maximálne vierohodný odhad vektora parametrov je označený • (vpravo)

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.

Príklad 104 (maximálne vierohodné odhady) *Za predpokladu normality rozdelenia náhodnej premennej X vypočítajte maximálne vierohodné odhady strednej hodnoty μ (ozn. $\widehat{\mu}$) a rozptylu σ^2 (ozn. $\widehat{\sigma}^2$) pomocou logaritmov funkcií vierohodnosti $l(\mu|\mathbf{x})$, resp. $l(\sigma^2|\mathbf{x})$. Porovnajte tieto odhady s aritmetickým priemerom $\overline{x}$ a rozptylom s^2. Musí platiť $\widehat{\mu} = \overline{x}$ a $\widehat{\sigma}^2 = \frac{(n-1)}{n}s^2$. Realizáciami náhodnej premennej X sú hodnoty $x_i, i = 1,2,\ldots,n$, premenných: (a) dĺžka pravej kľúčnej kosti (*`length.R`*; dáta:* `paired-means-clavicle2.txt`*); (b) morfologická výška tváre (*`face.H`*; dáta:* `one-sample-correlation-skull-mf.txt`*); (c) šírka lebky (*`skull.B`*; dáta:* `one-sample-mean-skull-mf.txt`*).*

Príklad 105 (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.*

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



Obr. 18: Funkcia vierohodnosti pre $X \sim Bin(N, p)$ ($p = 0.1, 0.5, 0.9$ a $N = 20$); odhady $\widehat{p}$ sú označené zvislou čiarkovanou čiarou

Z grafov na obrázku 18 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 106 (maximálne vierohodné odhady; binomické rozdelenie) *Za predpokladu, že náhodná premenná X má binomické rozdelenie, vypočítajte maximálne vierohodný odhad $\widehat{p}$ pomocou logaritmu funkcie vierohodnosti $l(p|\mathbf{x})$. Porovnajte tento odhad s výrazom $\sum_{i=1}^{N} x_i/N$. Realizáciami náhodnej premennej X sú nasledujúce binárne premenné: (a) pohlavie (`sex`; dáta: `one-sample-probability-sexratio.txt`), kde ozn. pohlavia dievča „f“ preznačíme na 1 a ozn. pohlavia chlapec „m“ preznačíme na 0; (b) pohlavie (`sex`; dáta: `two-samples-probabilities-sexratio.txt`), kde ozn. pohlavia muž „m“ preznačíme na 1 a ozn. pohlavia žena „f“ preznačíme na 0. V prípade (a) počítame pravdepodobnosť výskytu dievčat a v prípade (b) pravdepodobnosť výskytu chlapcov.*

Príklad 107 (maximálne vierohodné odhady; multinomické rozdelenie) *Za predpokladu, že náhodná premenná X má multinomické rozdelenie vypočítajte maximálne vierohodné odhady $\widehat{p}_1$ a $\widehat{p}_2$ pomocou logaritmu funkcie vierohodnosti $l(\mathbf{p}|\mathbf{x})$. Porovnajte odhad p_1 s odhadom p z príkladu 106, kde pravdepodobnosť $\widehat{p}_1$ bola označená ako $\widehat{p}$. Realizáciami X sú binárne premenné: (a) pohlavie (`sex`; dáta: `one-sample-probability-sexratio.txt`), kde ozn. pohlavia dievča „f“ preznačíme na 1 a ozn. pohlavia chlapec „m“ preznačíme na 0; (b) pohlavie (`sex`; dáta: `two-samples-probabilities-sexratio.txt`), kde ozn. pohlavia muž „m“ preznačíme na 1 a ozn. pohlavia žena „f“ preznačíme na 0. Pravdepodobnosť $\widehat{p}_1$ je (a) pravdepodobnosť výskytu dievčat a (b) pravdepodobnosť výskytu chlapcov.*

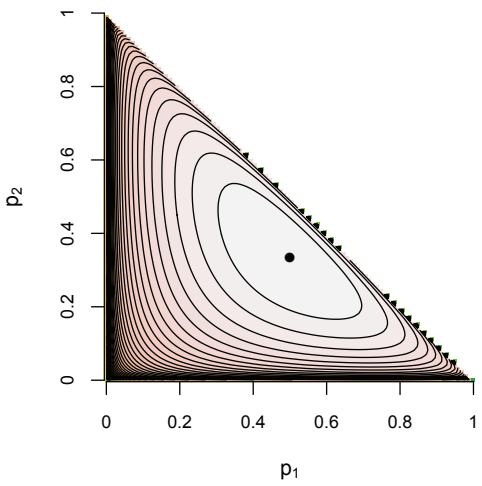
Príklad 107 hovorí o tom, že parameter p_1 dvojrozmerného multinomického rozdelenia je parametrom p binomického rozdelenia.

Príklad 108 (maximálne vierohodné odhady; multinomické rozdelenie) *Majme dáta `more-samples-probabilities-pubis.txt`. Nakreslite logaritmus štandardizovanej funkcie vierohodnosti $\mathcal{L}(\boldsymbol{\theta}|\mathbf{x})$, kde $\boldsymbol{\theta} = (p_1, p_2)^T$, Európskej populácie ($n_1 = 30$, $n_2 = 20$ a $n_3 = 10$) pomocou funkcie `contour()`. Dokreslite do obrázku jej maximum v bode $\widehat{\boldsymbol{\theta}} = (\widehat{p}_1, \widehat{p}_2)^T$.*

Riešenie v R (pozri obrázok 19)

```
poz.poc <- c(30,20,10)
N <- sum(poz.poc)
ocak.prav <- poz.poc/N
# riesenie cez maximalnu vierohodnost
"ll" <- function(theta,data) {
```

```
# logaritmus funkcie vierohodnosti
N <- sum(data)
p1 <- theta[1]
p2 <- theta[2]
p3 <- 1-p1-p2
if ((p1>0)&&(p1<1)&&(p2>0)&&(p2<1)&&(p3>0)&&(p3<1)) {
ocak.prav <- c(p1,p2,p3)
log.vier <- dmultinom(poz.poc,prob=ocak.prav,size=N,log=TRUE)
return(log.vier)
}
else {return(NA)}
}
# startovacie parametre optimalizacie c(0.1,0.2)
# maximalizacia, ak control$fnscale je rovne -1
OPTtheta <- optim(c(0.1,0.2),ll,control=list(fnscale=-1),
                  hessian=TRUE,data=poz.poc)
theta.hat <- OPTtheta$par # 0.4999662 0.3333621
```



Obr. 19: Logaritmus štandardizovanej funkcie vierohodnosti multinomického rozdelenia v parametroch p_1 a p_2 (Európska populácia) s maximom označeným •

Príklad 109 (overdispersion v Poissonovom modeli, pokrač.) *Majme početnosti úrazov n medzi m_n robotníkmi v továrni, pozri tabuľku 11 (Greenwood a Yule, 1920). Vypočítajte očakávané m_n za predpokladu, že početnosti úrazov na robotníka X majú negatívne binomické rozdelenie s parametrami α a π.* angl

Riešenie (pozri tabuľku 15)
Aby sme mohli fitovať negatívne binomické rozdelenie, potrebujeme funkciu vierohodnosti

$$L(\alpha,\pi|\mathbf{x}) = \prod_{n=0}^{4} (\Pr(X=n))^{m_n} \left(1 - \sum_{n=0}^{4} \Pr(X=n)\right)^{m_{\geq 5}}.$$

a jej logaritmus

$$l(\alpha,\pi|\mathbf{x}) = \sum_{n=0}^{4} m_n \ln \Pr(X=n) + m_{\geq 5} \ln \left(1 - \sum_{n=0}^{4} \Pr(X=n)\right).$$

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 15).

Tabuľka 15: Očakávané početnosti robotníkov m_n (zaokrúhlené na nula desatinných miest) s n úrazmi v továrni (negatívne binomické rozdelenie)

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

2.5 Kritériá klasifikácie štatistických modelov

Kritérá delenia štatistických modelov (ako aj k nim prislúchajúcich testov) sú nasledovné:

- **množstvo výberov** – jeden, dva alebo viac ako dva výbery [jedno-, dvoj- a viacvýberový test/model];
- **závislosť výberov** – nezávislé a závislé výbery (opakované merania na subjekte v čase, merania na párových orgánoch) [test/model dvoch a viacerých nezávislých výberov, párový test];
- **množstvo premenných** – jedna, dve alebo viac premenných [jedno-, dvoj- a viac-premenný test/model];
- **typ premenných** – kvalitatívne alebo kvantitatívne premenné [binomický, Poissonov, multinomický, súčinový multinomický model, model normálneho rozdelenia, rôzne modely kauzality];
- **rozmer endpointov/závislých premenných** – jedna, dve alebo viac premenných [jedno-, dvoj- a viac-rozmerný test/model – lineárny regresný model (LRM) vs. mnohorozmerný LRM (MLRM), analýza rozptylu ANOVA vs. mnohorozmerná ANOVA (MANOVA), analýza kovariancie ANCOVA vs. mnohorozmerná ANCOVA (MANCOVA)]; angl
- **typ náhodnosti efektov** – fixné, náhodné alebo zmiešané [(M)ANOVA, (M)ANCOVA model a test/testy v nich];
- **typ kauzálneho vzťahu** – lineárny (priamka) alebo nelineárny (polynóm ľubovoľného stupňa – kvadratický, kubický a iný, alebo ľubovoľná funkcia);
- **typ vzťahu parametrov modelu** – lineárny [LRM – priamka, polynóm ľubovoľného stupňa a pod.] a nelineárny [nelineárny regresný model (NLRM)]; toto kritérium sa často zamieňa s predchádzajúcim, ale tieto dve kritéria nie sú totožné;
- **prítomnosť odľahlých pozorovaní**;
- **typ hypotézy** – jednostranná, obojstranná, stochasticky usporiadaná.

2.6 Praktické dôsledky odchýlok od normality

Základným predpokladom mnohých modelov je normalita rozdelenie spojitých premenných, čo však nemusí byť v praxi splnené. Napr. hodnoty často používanej telesnej hmotnosti nemajú v západnej civilizácii normálne rozdelenie, ale negatívne zošikmené (pozri kapitolu 3.2 Charakteristiky variability). Odchýlky od normality môžu mať mnoho príčin, od **neošetrených metodických nezrovnalostí** až ku **skutočným biologickým procesom v populácii**, z ktorých (náhodný) výber (vzorka) pochádza. Predpoklad normality by nikdy nemal byť samozrejmý (ani u premenných, ktoré obvykle normálne rozdelenie majú), dáta by mali byť pred štatistickými analýzami vhodne zobrazené (pozri kapitolu 3.6 Štatistická grafika) a ich normalita by mala byť v každej vzorke testovaná (pozri kapitolu 5.2 Kolmogorov-Smirnovov test dobrej zhody). Normalita rozdelenia by mala byť starostlivo posudzovaná a v prípade zachytenia odchýlkok od normality v nejakom znaku v mnohých nezávislých výberoch by príčiny odchýlky mali byť bližšie skúmané. Malo by sa zistiť, či ide o metodickú chybu (odľahlé hodnoty žiadajúce odstránenie, nenáhodný výber, atď.) alebo skutočný trend, ktorý má

svoje biologické príčiny (*rozdielna regulácia hornej a dolnej medze intenzity nejakého metabolického procesu* alebo *smerová/direkcionálna selekcia* u zošikmeného rozdelenia, *sledovaným znakom obmedzené vzorkovanie* (nenáhodný výber) alebo *stabilizačná selekcia u leptokurtického rozdelenia* (pozri kapitolu 3.2 Charakteristiky variability), atď.). Zošikmené rozdelenie majú často *inkrementálne (prírastkové) dáta* merané v priebehu ontogenézy, keďže je ich dolná hranica prirodzene obmedzená (nulový prírastok), zatiaľ čo horná môže u každého jedinca dosiahnúť rôzne veľké hodnoty (Garn a Rohmann, 1963). Treba mať na pamäti, že našim hlavným a najdôležitejším cieľom je zistenie podstaty procesov, ktoré prebiehajú v populácii, z ktorej pochádza naša vzorka a ktorej je obrazom.

V bežnej praxi je však časté, že sú používané relatívne malé vzorky, v ktorých sa odchýlky od normality ľahko prejavia. Môžu však nastať aj *kontroverzné situácie*, že rozdelenie dát (pri relatívne malej vzorke) sa nepodobá na normálne, ale test dobrej zhody s normálnym rozdelením hypotézu o zhode zamietne. Rovnako sa často stáva, že meranie alebo experiment nie sú dopredu naplánované z hľadiska náhodného výberu, použitia nejakého modelu, z hľadiska minimalizácie variability a odhadu minimálneho rozsahu súboru. Dôležitosť takéhoto postupu je bežne aplikovanými výskumníkmi ignorovaná, čo môže viesť až do situácie nemožnosti použitia dát samotných alebo nemožnosti použitia akejkoľvek štatistickej metódy.

Súčasne sa utvrdzuje predstava o ideálnej podstate normality v prírode a tiež všeobecnej normalite rozdelenia často testovaných znakov, takže akékoľvek ďalšie, novozískané odchýlky od normality sa považujú za chybu. Skutočnosť ale často ukazuje, že so zvyšujúcim sa počtom prípadov vo vzorke (virtuálne až k celkovej veľkosti populácie) celý rad (empirických) veličín v skutočnosti nesmeruje ku stále dokonalejšej Gausovej krivke, ale má rozdelenie od normálneho viac či menej sa vzďaľujúce alebo nejakým spôsobom vychýlené. V týchto prípadoch sa odporúča

1. použitie nejakej **transformácie dát** (v prípade zošikmenia; logaritmická, odmocninová, Box-Coxova a pod.), ktorá závisí na samotných dátach a nedá sa použiť univerzálne;

2. použitie **urezávania alebo winsorizácie dát** na jednom alebo oboch koncoch rozdelenia (v prípade prítomnosti odľahlých pozorovaní alebo zošikmenia; pozri kapitolu 3.1 Charakteristiky polohy, 3.2 Charakteristiky variability a 3.3 Detekcia odľahlých pozorovaní);

3. **nahradenie asymptotického rozdelenia testovacej štatistiky bootstrapovým alebo permutačným** (v prípade nedostatočného alebo relatívne malého rozsahu vzorky alebo pochybností o asymptotickom rozdelení testovacej štatistiky aj pri väčších rozsahoch).

Po aplikovaní prvých dvoch metód je možné použiť asymptotické testy v nezmenenej podobe a v treťom prípade sa použijú len samotné testovacie štatistiky.