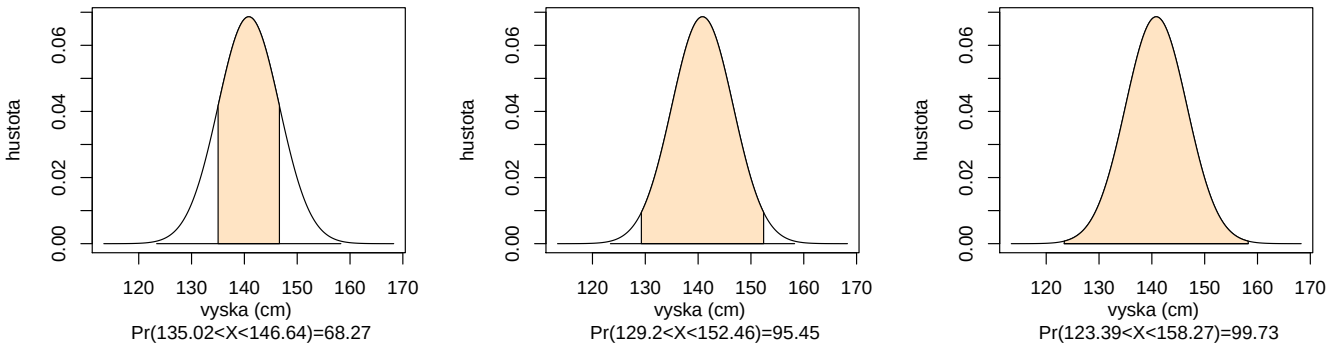
Zadání příkladů - cvičení č.1 - 15-9-23

Příklad č.1 (porovnání dvou typů modelů) *(přednáška)*
Model rozdělení pravděpodobnosti je modelem náhodné proměnné X, např. (1) model rozdělení pravděpodobnosti náhodné proměnné X šířka dolní čelisti, nebo (2) model rozdělení pravděpodobnosti náhodné proměnné X hrubost kožních řas u dospělých zdravých žen. *Statistický model* je modelem náhodné proměnné $Y|X$ (Y kauzálně závisí na X), např. (1) model závislosti náhodné proměnné Y šířka dolní čelisti na proměnné X pohlaví, nebo (2) model závislosti náhodné proměnné Y hrubost kožních řas u dospělých zdravých žen na proměnné X BMI. Všimněme si, že náhodné proměnné označujeme X anebo Y podle toho, jaký model je charakterizuje.

Příklad č.2 (jednoduchý náhodný výběr)
V jednoduchém náhodném výběru o rozsahu n z populace s konečným rozsahem N má každý prvek stejnou pravděpodobnost vybrání. Pokud vybíráme bez vracení (opakování), mluvíme o *jednoduchém náhodném výběru bez vracení* (Dalgaard, 2008). Pokud vybíráme s vracením, mluvíme o *jednoduchém náhodném výběru s vracením.* Mějme množinu $\mathcal{M}$ s $N = 10$ prvky a chceme z ní vybrat $n = 3$ prvky (a) bez vracení, (b) s vracením. Kolik máme možností? Jak vypadá jedna takováto možnost, pokud $\mathcal{M} = \{1, 2, \ldots, 10\}$? Zopakujte to samé pro $N = 100$, $n = 30$ a množinu $\mathcal{M} = \{1, 2, \ldots, 100\}$.

Příklad č.3 (jednoduchý náhodný výběr)
Mějme skupinu lidí označených identifikačními čísly (ID) od 1 do 30. Vyberte (a) náhodně 5 lidí z 30-ti bez návratu, (b) náhodně 5 lidí ze 30-ti s návratem a nakonec (c) náhodně 5 lidí ze 30-ti bez návratu, přičemž lidé s ID od 28-mi do 30-ti mají pravděpodobnost vybrání 4× vyšší než lidé s ID od 1 do 27.

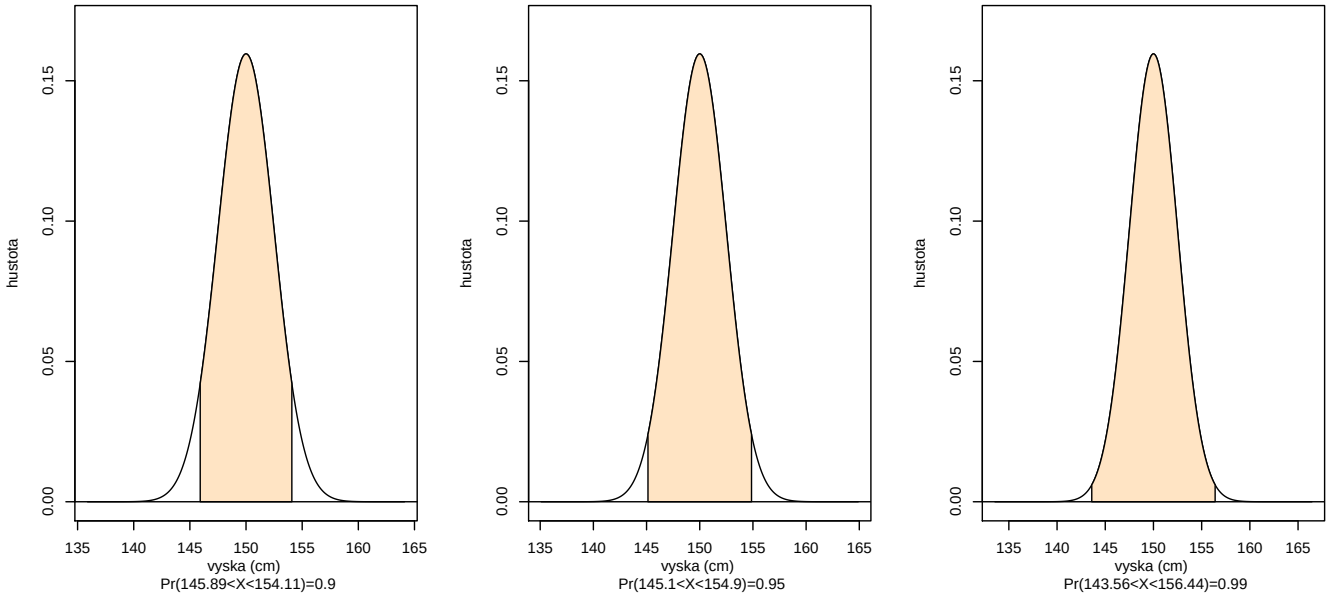
Příklad č.4 (normální rozdělení)
Mějme náhodnou proměnnou X (může to být např. výška postavy desetiletých dívek) a předpokládejme, že tato náhodná proměnná má normální rozdělení s parametry μ (střední hodnota) a σ^2 (rozptyl), což zapisujeme jako $X \sim N(\mu, \sigma^2)$, $\mu = 140.83$, $\sigma^2 = 33.79$. Normální rozdělení představuje model rozdělení pravděpodobnosti pro tuto náhodnou proměnnou. Vypočítejte pravděpodobnost $\Pr(a \leq X \leq b) = \Pr(X < b) - \Pr X < a) = F_X(b) - F_X(a)$, kde $a = \mu - k\sigma$, $b = \mu + k\sigma$, $k = 1, 2, 3$. Nakreslete hustotu rozdělení pravděpodobnosti, vybarvěte oblast mezi body a a b a popište osy x a y tak, jako je uvedeno na obrázku 1.



Obrázek 1: Míry normálního rozdělení; křivka hustoty s vybarveným obsahem pod touto křivkou mezi příslušnými kvantily na ose x; obsah je rovný pravděpodobnosti výskytu subjektů s danou výškou v rozpětí těchto kvantilů.

Dostaneme pravidlo $68.27 - 95.45 - 99.73$ (tzv. *míry normálního rozdělení.*

Příklad č.5 (normální rozdělení)
Mějme $X \sim N(\mu, \sigma^2)$, kde $\mu = 150$, $\sigma^2 = 6.25$. Vypočítejte $a = \mu - x_{1-\alpha/2}\sigma$ a $b = \mu + x_{1-\alpha/2}\sigma$ tak, aby $\Pr(a \leq X \leq b) = 1-\alpha$, byla rovná 0.9, 0.95, 0.99. Číslo $x_{1-\alpha/2}$ je kvantil normovaného normálního rozdělení, t.j. $\Pr(Z = \frac{X-\mu}{\sigma} < x_{1-\alpha}$, $Z \sim N(0,1)$. Nakreslete hustotu rozdělení pravděpodobnosti, vybarvěte oblast mezi body a a b a popište osy x a y tak, jako je uvedeno na obrázku 2.



Obrázek 2: Upravené míry normálního rozdělení; křivka hustoty s vybarveným obsahem pod touto křivkou mezi příslušnými kvantily na ose x; obsah je rovný pravděpodobnosti výskytu subjektů s danou normovanou výškou v rozpětí těchto kvantilů.

Dostaneme pravidlo $90 - 95 - 99$ (tzv.*upravené míry normálního rozdělení*. Použili jsme nerovnost $\Pr(u_{\alpha/2<Z<u_{1-\alpha/2}} = \Phi(x_{1-\alpha/2}) - \Phi(x_{\alpha/2}) = 1 - \alpha$, kde Φ je distribuční funkce normálního normovaného rozdělení a všeobecně ($\alpha \in (0, 1/2)$; v příkladě $\alpha = 0.1, 0.05$ a 0.01.

Příklad č.6 (normální rozdělení)
Předpokládejme model normálního rozdělení $N(132, 13^2)$ pro systolický krevní tlak. Jaká část populace (v %) bude mít hodnoty vyšší než 160 mm Hg?

Příklad č.7 (binomické rozdělení)
Předpokládejme, že počet lidí upřednostňujících léčbu A před léčbou B se řídí modelem binomického rozdělení s parametry N (rozsah náhodného výběru) a p (pravděpodobnost výskytu), ozn. $Bin(N, p)$, kde $N = 20$, $p = 0.5$, t.j. lidé preferují oba dva typy léčby stejnou měrou. (a) Jaká je pravděpodobnost, že 16 a více pacientů upřednostní léčbu A před léčbou B? (b) Jaká je pravděpodobnost, že 16 a více a zároveň 4 a méně pacientů upřednostní léčbu A před léčbou B?

Příklad č.8 (binomické rozdělení)
Předpokládejme, že $\Pr(vir) = 0.533 = p_1$ je pravděpodobnost výskytu dermatoglyfického vzoru vír na palci pravé ruky mužů české populace a $\Pr(ostatni) = 0.467 = p_2$ je pravděpodobnost výskytu ostatních vzorů na palci pravé ruky mužů české populace, přičemž X je počet vírů a Y je počet ostatních vzorů, kde $X \sim Bin(N, p_1)$ a $Y \sim Bin(N, p_2)$. Vypočítejte (1) $\Pr(X \leq 120)$, když $N = 300$ a (2) $\Pr(Y \leq 120)$, když

$N = 300$.

Příklad č.9 (parametry) *(přednáška)*
Příklady parametrů θ - střední hodnota μ, rozptyl σ^2, korelační koeficient ρ, pravděpodobnost p výskytu nějaké události, rozdíl dvou středních hodnot $\mu_1 - \mu_2$, podíl dvou rozptylů σ_1^2/σ_2^2, rozdíl dvou korelačních koeficientů $\rho_1 - \rho_2$, rozdíl dvou pravděpodobností $p_1 - p_2$ apod.

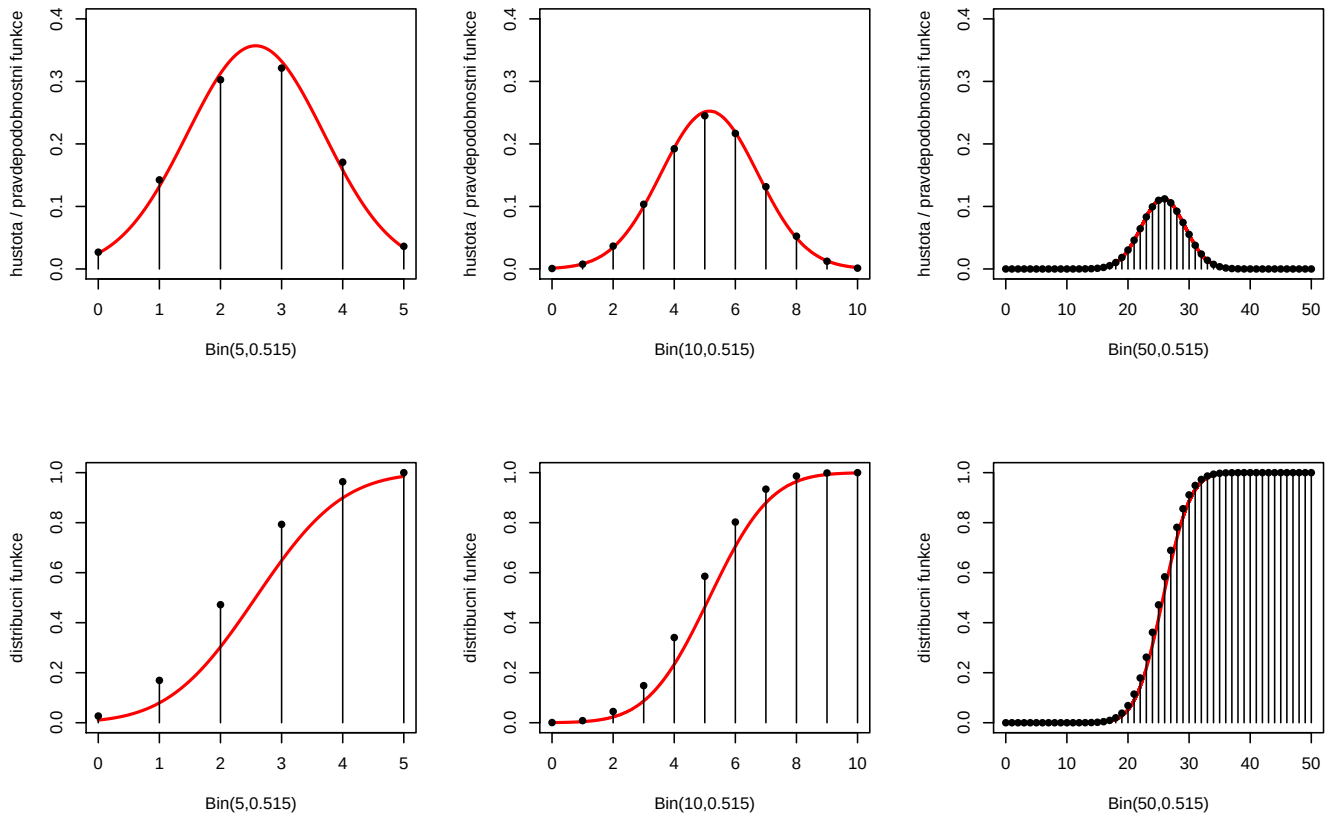
Příklad č.10 (binomické rozdělení) *(přednáška)*
Pokud $X \sim Bin(N, \theta)$, $\theta = p \in \langle 0; 1 \rangle$, potom $\mathcal{Y}_\theta$ je stejný pro všechny θ a koinciduje s výběrovým prostorem $\mathcal{Y} = \{0, 1, \ldots, N\}$.

Příklad č.11 (počet členů v mnohorozměrném LRM) *(z přednášky)*
Mějme mnohorozměrný lineární regresní model $\mathcal{L}$ o 20-ti proměnných, ve kterém jsou obsaženy všechny možné interakce těchto proměnných (dvojné, trojné,...). Kolik členů (jednoduché regresory + všechny interakce) má takový model?

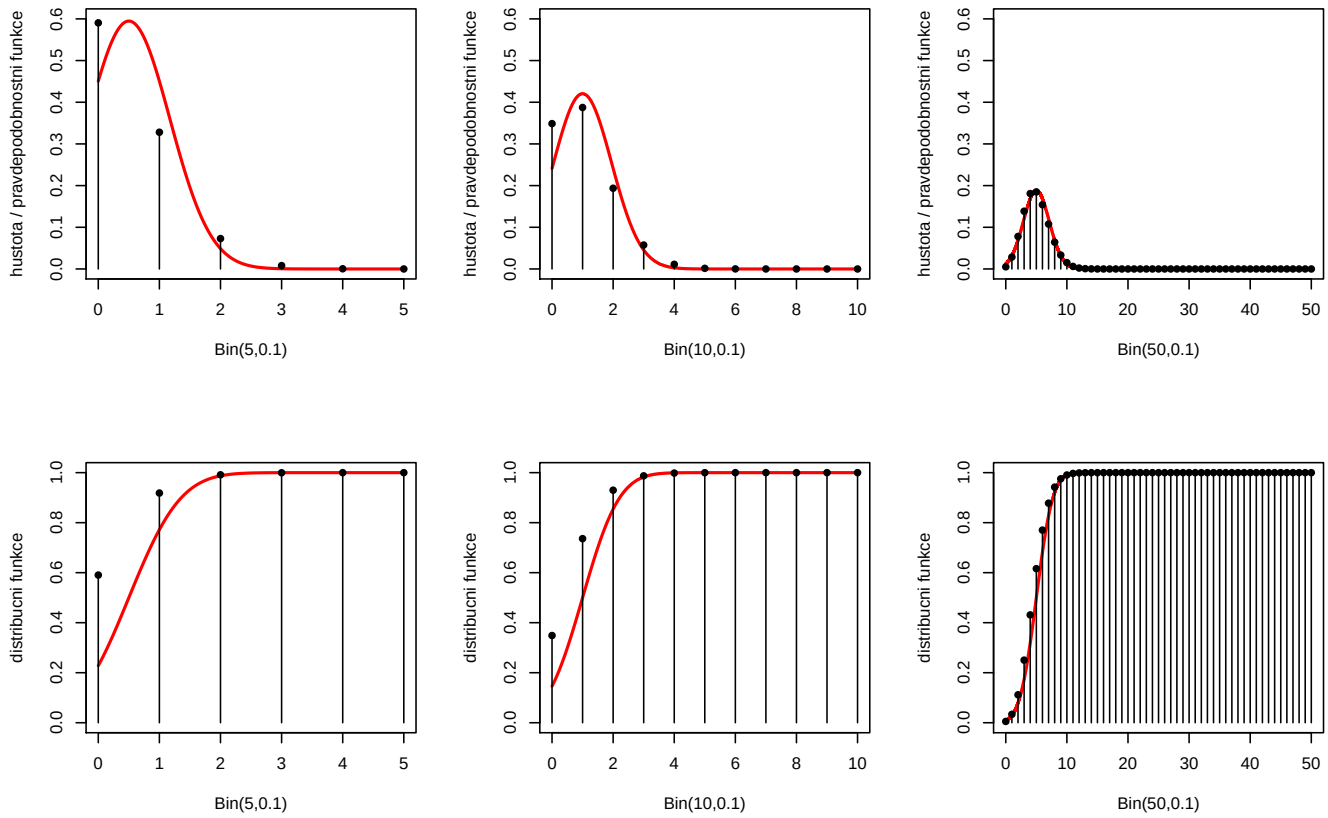
Příklad č.11 (aproximace binomického rozdělení normálním)
Nechť $\Pr(\text{muz}) = p = 0.515$ znamená pravděpodobnost výskytu mužů v populaci a $\Pr(\text{zena}) = q = 0.485$ pravděpodobnost výskytu žen. Nechť X je počet mužů a Y počet žen. Za předpokladu modelu $Bin(N, p)$ vypočítejte (a) $\Pr(X \leq 3)$ pokud $N = 5$, (b) $\Pr(X \leq 5)$, pokud $N = 10$ a (c) $\Pr(X \leq 25)$, pokud $N = 50$. Porovnejte vypočítané pravděpodobnosti s pravděpodobnostmi aproximovanými normálním rozdělením $N(Np, Npq)$.

Nakreslete hustotu rozdělení pravděpodobnosti normálního rozdělení a superponujte ji pravděpodobnostní funkcí binomického rozdělení tak, jak je uvedeno na obrázku 3. Nakreslete distribuční funkci normálního rozdělení a superponujte ji distribuční funkcí binomického rozdělení tak, jak je uvedeno na obrázku 3.

Nakonec zvolte parametr $p = 0.1$ a vygenerujte analogické grafy hustoty a distribuční funkce pro tento nový parametr. Z obrázků je vidět, že pro p blížící se k 1 nebo k 0 je potřebné mít větší početnosti než pro p blízké hodnotě 0.5. Viz obrázek 4.



Obrázek 3: Aproximace binomického rozdělení normálním pro $p = 0.515$ a $N = 5{,}10$ a 50; spojnicový graf superponovaný hustotou (první řádek) a distribiční funkcí (druhý řádek).



Obrázek 4: Aproximace binomického rozdělení normálním pro $p = 0.515$ a $N = 5{,}10$ a 50; spojnicový graf superponovaný hustotou (první řádek) a distribiční funkcí (druhý řádek).

Příklad č.12 (normální rozdělení)
Model pro náhodný výběr $X_1, X_2, \ldots, X_n$ je z $N(\mu, \sigma^2)$ a říkáme, že $X_1, X_2, \ldots, X_n$ pochází z normálního rozdělení, t.j. $X \sim N(\mu, \sigma^2)$. Parametr modelu $N(\mu, \sigma^2)$ je vektor $\boldsymbol{\theta} = (\mu, \sigma^2)$. Hustota tohoto rozdělení má tvar

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}.$$

Příklad č.13 (standardizované normální rozdělení)
Model pro náhodný výběr $X_1, X_2, |dots, X_n$ pochází ze standardizovaného normálního rozdělení, t.j. $X \sim N(\mu, \sigma^2)$, kde $\mu = 0$, $\sigma^2 = 1$. Parametr modelu $N(\mu, \sigma^2)$ je vektor $\boldsymbol{\theta} = (0, 1)$. Hustota tohoto rozdělení má tvar

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}, x \in \mathbb{R}.$$

Příklad č.14 (dvojrozměrné normální rozdělení)
Náhodný vektor $(X, Y)^T$ má dvojrozměrné normální rozdělení

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

s hustotou

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

kde $(x, y)^T \in \mathbb{R}^2$, $\mu_j \in \mathbb{R}$, $\sigma_j^2 > 0$, $j = 1, 2$, $\rho \in \langle -1, 1 \rangle$ jsou parametry. Potom $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Výraz v exponentu můžeme zapsat jako

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

Marginální rozdělení [1] jsou $X \sim N\left(\mu_1, \sigma_1^2\right)$ a $Y \sim N\left(\mu_2, \sigma_2^2\right)$, ρ je koeficient korelace[2](Viz obrázek 5)

Příklad č.15 (dvojrozměrné normální rozdělení)

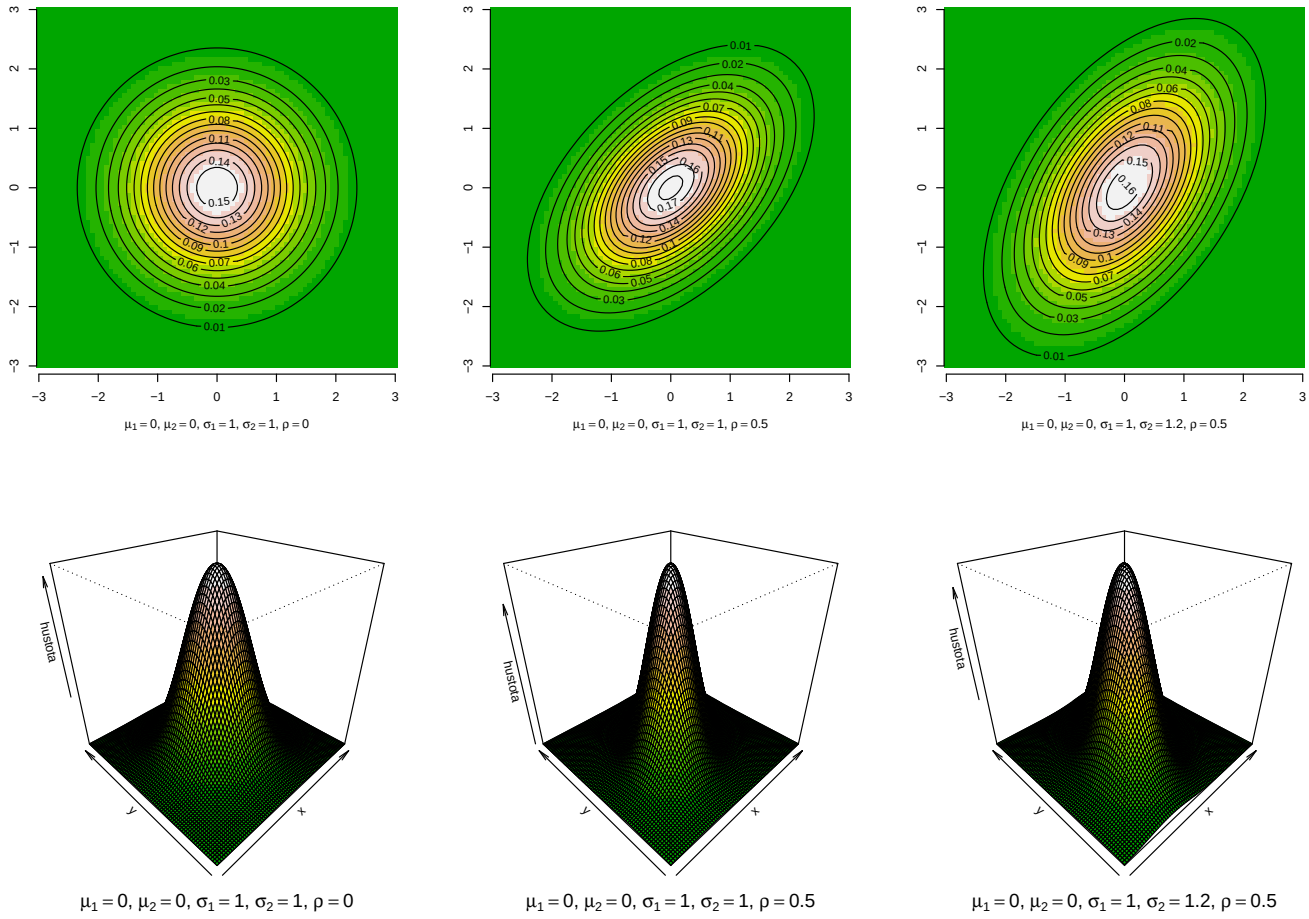
(1) Nakreslete hustotu dvojrozměrného normálního rozdělení $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pomocí funkce image() a superponujte ho s konturovým grafem hustoty toho stejného rozdělení pomocí funkce contour(). (2) Nakreslete hustotu dvojrozměrného normálního rozdělení $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pomocí funkce persp(). Hustotu rozsekejte na 12 intervalů, kde hodnoty v těchto intervalech budou odpovídat barvám terrain.colors(12). Použijte následující parametry:

- $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1, \rho = 0$;
- $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1, \rho = 0.5$;
- $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1.2, \sigma_2 = 1, \rho = 0.5$.

Vzorové řešení je uvedeno na obrázku 5.

[1] Marginální rozdělení je rozdělení náhodné proměnné, zde X nezávisle na Y a naopak Y nezávisle na X.

[2] Z tohoto příkladu je zřejmé, že na dostatečný popis dvojrozměrného normálního rozdělení potřebujeme pět parametrů, t.j. střední hodnotu a rozptyl pro marginální rozdělení náhodných proměnných X a Ya korelační koeficient $\rho = \rho(X, Y)$ popisující sílu lineárního vztahu X a Y.



Obrázek 5: Hustoty dvojrozměrného normálního rozdělení při různých parametrech (první řádek – konturový graf; druhý řádek - perspektivní trojrozměrný graf v podobě plochy); čím je ρ odlišnější od nuly, tím více se kontury liší od kruhů (mění se na elipsy); se zvyšujícím se rozdílem mez σ_1 a σ_2 se zvětšuje rozdíl rozptýlení koncentrických kruhů ve směru jednotlivých os (říkáme, že rozdíl variability proměnných X a Y se zvětšuje.)

Příklad č.17 (standardizované normální rozdělení)
Náhodný vektor $(X, Y)^T$ má dvojrozměrné normální rozdělení

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

s hustotou

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

kde $(x, y)^T \in \mathbb{R}^2$, $\rho \in \langle -1, 1\rangle$ jsou parametry, potom $\boldsymbol{\theta} = (0, 0, 1, 1, \rho)$. Výraz v exponentu můžeme psát jako

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

marginální rozdělení jsou obě $N(0, 1)$ a ρ je koeficient korelace.

Příklad č.18 (standardizované normální rozdělení)
Nechť náhodnou proměnnou $X \sim N(\mu_1, \sigma_1^2)$ je největší výška mozkovny (skull.pH; v mm) a náhodnou

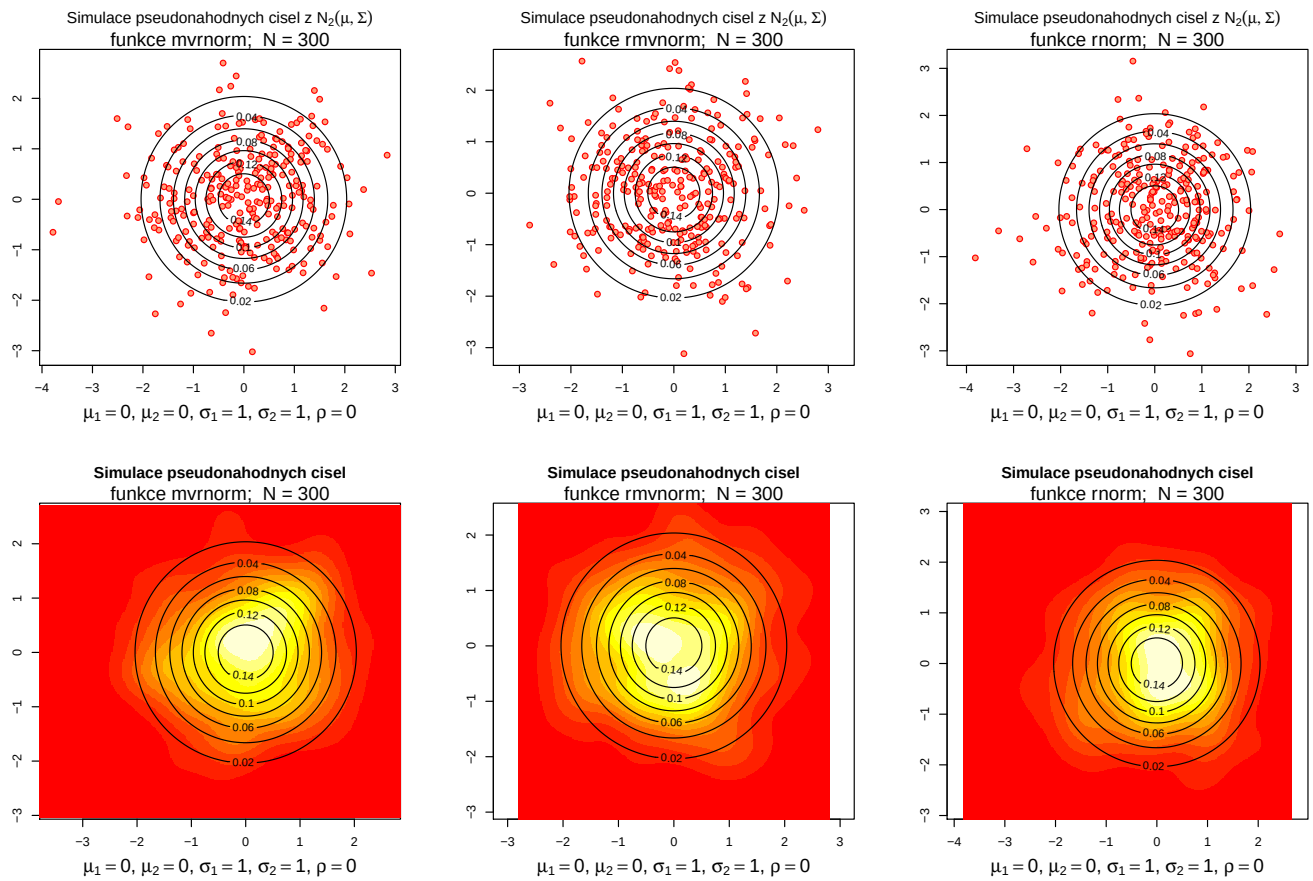
proměnnou $Y \sim N(\mu_2, \sigma_2^2)$ je morfologická výška tváře (face.H; v mm). Nechť X a Y mají dvojrozměrné normální rozdělení s parametry $(\mu_1, \mu_2)^T$ a σ_1^2, σ_2^2 a ρ jsou parametry kovarianční matice $\mathbf{\Sigma}$. Když od náhodné proměnné X odpočítáme její střední hodnotu μ_1 a tento rozdíl podělíme odmocninou z rozptylu (σ_1), dostaneme náhodnou proměnnou Z_X, která má asymptoticky normální rozdělení se střední hodnotou $\mu_1 = 0$ a rozptylem $\sigma_1^2 = 1$, což zapisujeme jako $Z_X \sim N(0, 1)$. Pokud od náhodné proměnné Y odečteme její střední hodnotu μ_2 a tento rozdíl podělíme odmocninou z rozptylu (σ_2), dostaneme náhodnou proměnnou Z_Y, která má asymptoticky normální rozdělení se střední hodnotou $\mu_2 = 0$ a rozptylem $\sigma_2^2 = 1$, což zapisujeme jako $Z_Y \sim N(0, 1)$. Potom $(Z_X, Z_Y)^T$ má standardizované dvourozměrné normální rozdělení $N_2(\boldsymbol{\mu}, \mathbf{\Sigma})$ s parametry $\boldsymbol{\mu} = (0, 0)^T$ a $\sigma_1^2 = 1$, $\sigma_2^2 = 1$ a ρ jsou parametry kovarianční matice $\mathbf{\Sigma}$.

Příklad č.19 (dvourozměrné normální rozdělení)
Simulaci pseudonáhodných čísel z $N_2(\boldsymbol{\mu}, \mathbf{\Sigma})$ můžeme v R vytvořit následujícími způsoby:

1. použitím funkce mvrnorm() z knihovny MASS;
2. použitím funkce rmvnorm() z knihovny mvtnorm
3. použitím funkce rnorm() a následujícího algoritmu:
   Nechť $X_1 \sim N(0, 1)$ a $X_2 \sim N(0, 1)$; potom $(Y_1, Y_2)^T \sim N_2(\boldsymbol{\mu}, \mathbf{\Sigma})$, kde $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ je vektor středních hodnot a σ_1^2 a σ_2^2 a ρ jsou parametry kovarianční matice $\mathbf{\Sigma}$, přičemž síla lineárního vztahu Y_1 a Y_2 je daná velikostí a znaménkem 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 pseudonáhodná čísla Y_1 a Y_2 z $N_2(\boldsymbol{\mu}, \mathbf{\Sigma})$. Vypočítejte dvourozměrný jádrový odhad hustoty $(Y_1, Y_2)^T$ pomocí funkce kde2d(). Nakreslete jej také pomocí funkce image() a superponujte jej kontúrovým grafem hustoty dvourozměrného normálního rozdělení $N_2(\boldsymbol{\mu}, \mathbf{\Sigma})$ pomocí funkce contour(). Hustotu rozsekejte na 12 intervalů, kde hodnoty v těchto intervalech budou odpovídat barvám terrain.colors(12). Při simulaci použijte následující parametry:

   (a) $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$, $\rho = 0$; (1) $n = 50$, (2) $n = 500$
   (b) $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$, $\rho = 0.5$; (1) $n = 50$, (2) $n = 500$
   (c) $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1.2$, $\rho = 0.5$; (1) $n = 50$, (2) $n = 500$

   Vzorové řešení viz obrázek 6.

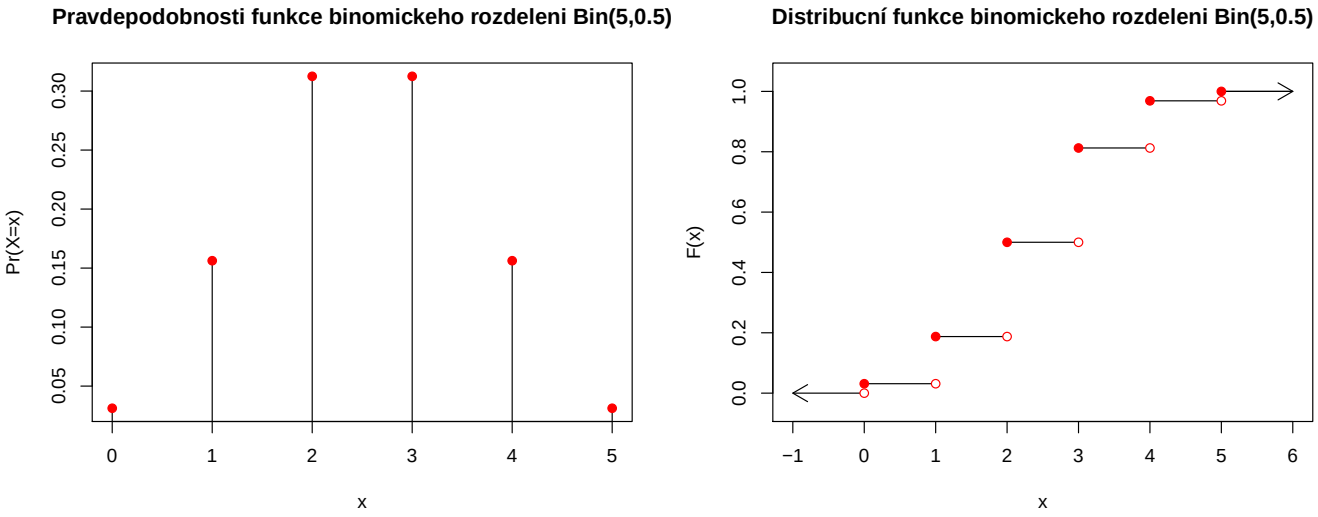


Obrázek 6: Hustoty dvourozměrného normálního rozdělení

Příklad č.23 (binomické rozdělení, binomický experiment)
Experiment sestávající z fixního počtu Bernoulliho experimentů (ozn. N) se nazývá binomický experiment. Pravděpodobnost úspěchu označme p, pravděpodobnost neúspěchu $q = 1 - p$. Náhodná proměnná X je počet pozorovaných úspěchů po dobu experimentu. Pravděpodobnost $X = x$ za podmínky, že X pochází z binomického rozdělení $Bin(N, p)$, píšeme jako

$$\Pr(X = x) = \binom{N}{x} p^x (1-p)^{N-x}, x = 0, 1, \ldots, N \tag{1}$$

(Ugarte a kol. 2008). Střední hodnota $E[X] = Np$ a rozptyl $Var[X] = Np(1-p)$. Naprogramujte a zobrazte v R pravděpodobnostní funkci a (kumulativní) distribuční funkci pro $Bin(5, 0.5)$. Řešení viz obrázek 7.



Obrázek 7: Pravděpodobnostní a distribuční funkce binomického rozdělení $Bin(5, 0.5)$

Příklad č.26 (Poissonovo rozdělení; počet havárií za týden)
Pokud každý z 50 milionů lidí řídí v Itálii řídí auto následující týden nezávisle, potom pravděpodobnost smrti při autonehodě bude 0.000002, kde počet úmrtí má binomické rozdělení Bin$(50mil, 0.000002)$ anebo limitní Poissonovo rozdělení s parametrem $\lambda = 50mil \times 0.000002 = 100$.

Příklad č.27 (Poissonovo rozdělení; pruské armádní jednotky)
Nechť početnosti úmrtí X jako následek kopnutí koněm v Pruských armádních jednotkách (Bortkiewicz, 1898) mají Poissonovo rozdělení s parametrem λ, tj. $X \sim Poiss(\lambda)$. Pravděpodobnost, že někdo bude smrtelně zraněný v daném dni, je extrémně malá. Mějme 10 vojenských jednotek za 20-letou periodu s rozsahem $M = 200$ $(200 = 10 \times 20)$, kde, při početnostech úmrtí $n = 1, 2, 3, 4, 5+$ v dané jednotce a v daném roce, zaznamenáváme také početnosti vojenských jednotek m_n při daném n, kde $M = \sum m_n$ (viz tabulka). Vypočítejte očekávané početnosti, za předpokladu $X \sim Poiss(\lambda)$, kde

$$\lambda = \frac{\sum_n n m_n}{\sum_n m_n}. \tag{2}$$

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

Příklad č.28 (podíl chlapců a dívek v rodinách)
Nechť X představuje početnost chlapců mezi dětmi v rodinách. Zde můžeme předpokládat, že $X \nsim Bin(N, p)$, tj. rodina může mít vychýlený poměr pohlaví dětí ve směru k chlapcům nebo k dívkám. V realitě tedy můžeme mít velmi mnoho rodin jen s chlapci nebo jen s děvčaty a nemáme dostatek rodin s poměrem pohlaví blízkým 51 : 49 (poměr chlapců ku dívkám). Z toho nám vyplývá, že rozptyl početnosti chlapců bude ve skutečnosti větší než rozptyl předpokládaný binomickým rozdělením $Bin(n, P)$.

Příklad č.29 (overdispersion v binomickém modelu)
V klasické studii poměru pohlaví u lidí z roku 1889 na základě záznamů z nemocnic v Sasku (více informací viz Lindsey a Altham, (1998)) zaznamenal Geissler (1889) rozdělení počtu chlapců v rodinách. Mezi $M = 6115$ rodinami s $N = 12$ dětmi pozoroval následující početnosti chlapců (n jsou početnosti chlapců a m_n početnosti rodin s n chlapci).

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

Vypočítejte m_n za předpokladu, že početnosti chlapců X v rodinách mají binomické rozdělení s parametry

$$\pi = \frac{\sum_{n=0}^{N} n m_n}{NM} = 0.5192 \tag{3}$$

a $N = 12$, ozn. $X \sim Bin(N, \pi)$.

Příklad č.30 (overdispersion v Poissonově modelu)
Mějme početnosti úrazů n mezi dělníky v továrně, kde početnosti dělníků m_n při daném n (viz tabulka) (Greenwood a Yule (1920)).

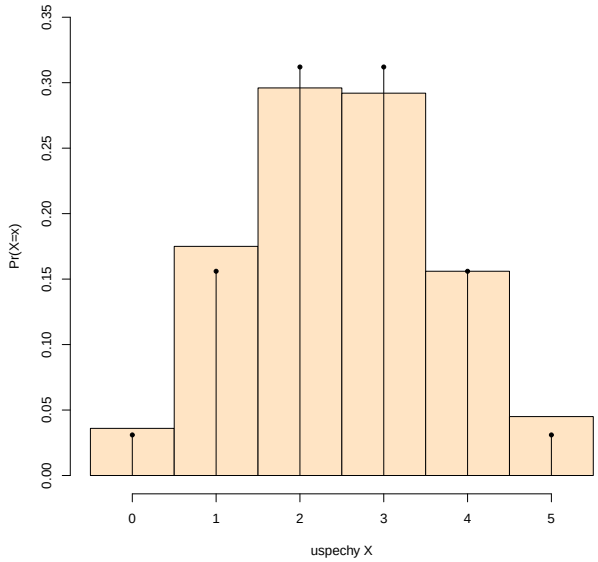
n	0	1	2	3	4	≥5
m_n	447	132	42	21	3	2

Vypočítejte očekávané početnosti dělníků za předpokladu, že početnosti úrazů na dělníka X mají Poissonovo rozdělení s parametrem

$$\lambda = \frac{\sum_n n m_n}{\sum_n m_n} = 0.47. \tag{4}$$

Ozn. $X \sim Poiss(\lambda)$.

Příklad č.31 (binomické rozdělení, simulační studie)
Vygenerujte pseudonáhodná čísla X (početnosti úspěchů) opakovaná M-krát ($M = 1000$) z $Bin(N, p)$, kde $N = 5$ a $p = 0.5$. Vytvořte tabulku vygenerovaných (simulovaných) i teoretických relativních početností (pro $n = 0, 1, \ldots, 5$). Superponujte histogram vygenerovaných pseudonáhodných čísel s teoretickou pravděpodobnostn funkcí (viz obrázek 8).



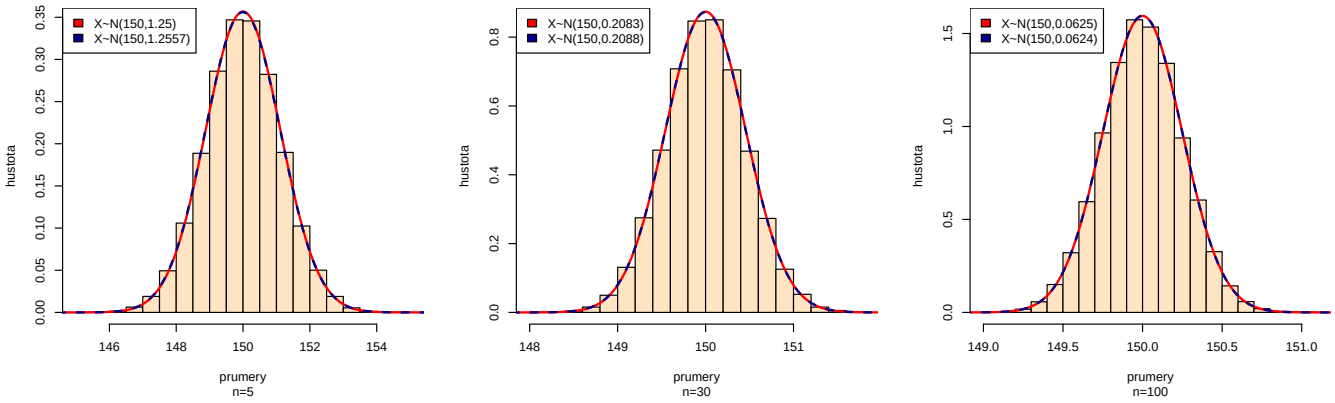
Obrázek 8: Histogram vygenerovaných pseudonáhodných čísel superponovaný teoretickou pravděpodobnostní funkcí $Bin(N, p)$.

Příklad č.32 (binomické vs normální rozdělení)
Nechť $X_N \sim Bin(N, p)$, potom můžeme aproximovat binomické rozdělení normálním následovně: $X_N \sim N(Np, Np(1-p))$, kde také platí

$$Z_N = \frac{X_N - Np}{\sqrt{Np(1-p)}} \sim N(0, 1).$$

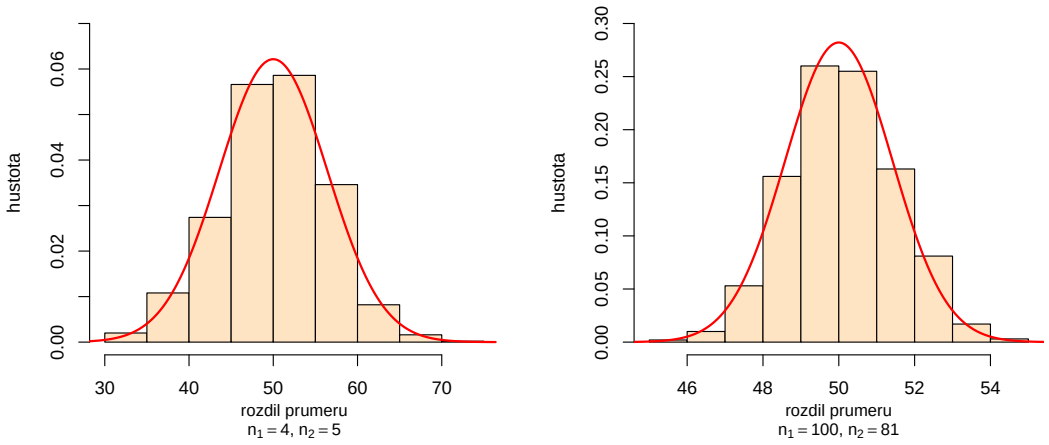
Ukažte, že CLV platí pro $N = 100$ a $p = 0.5$ na tři desetinná místa.

Příklad č.33 (normální rozdělení, simulační studie)
Na základě simulační studie prověřte, že pokud $X \sim N(150, 6.25)$, potom $\bar{[X]}_n \sim N(150, 6.25/n)$. Použijte $n = 30$. Pro každou simulaci X vypočítejte aritmetické průměry $\bar{x}_m$, $m = 1, 2, \ldots, M$, kde $M = 500\,000$. Superponujte je histogramem v relativní škále s teoretickou křivkou hustoty pro $\bar{X}_n$. Vypočítejte $\Pr(\bar{X}_n > 151)$ ze simulovaných dat a porovnejte tento výsledek s teoretickou (očekávanou) pravděpodobností. Řešení viz obrázek 9.



Obrázek 9: Histogram vygenerovaných průměrů superponovaný teoretickou křivkou hustoty $\bar{X}_n$.

Příklad č.34 (normální rozdělení, simulační studie)
Nechť $X \sim N(\mu_1, \sigma_1^2)$ a $Y \sim N(\mu_2, \sigma_2^2)$. Potom $\bar{X}_{n_1} - \bar{Y}_{n_2} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$. Generujte pseudonáhodná čásla X a Y rozdělení $N(\mu_j, \sigma_j^2)$, $j = 1, 2$, kde $\mu_1 = 100$, $\sigma_1 = 10$, $\mu_2 = 50$, $\sigma_2 = 9$ při (a) $n_1 = 4$, $n_2 = 5$, (b) $n_1 = 100$, $n_2 = 81$. Pro každou simulaci X a Y vypočítejte rozdíl $\bar{x}_m - \bar{y}_m$, $m = 1, 2, \ldots M$, kde $M = 1000$. Superponujte histogram těchto rozdílů v relativní škále s teoretickou křivkou hustoty rozdílu $\bar{X}_{n_1} - \bar{Y}_{n_2}$. Pro případ (a) i (b) vypočítejte $\Pr(\bar{X}_{n_1} - \bar{Y}_{n_2} < 52)$ na základě empirického (vygenerovaného) a teoretického rozdělení $\bar{X}_{n_1} - \bar{Y}_{n_2}$.



Obrázek 10: Histogram vygenerovaných rozdílů průměrů superponovaný teoretickou křivkou hustoty rozdělení rozdílu výběrových aritmetických průměrů

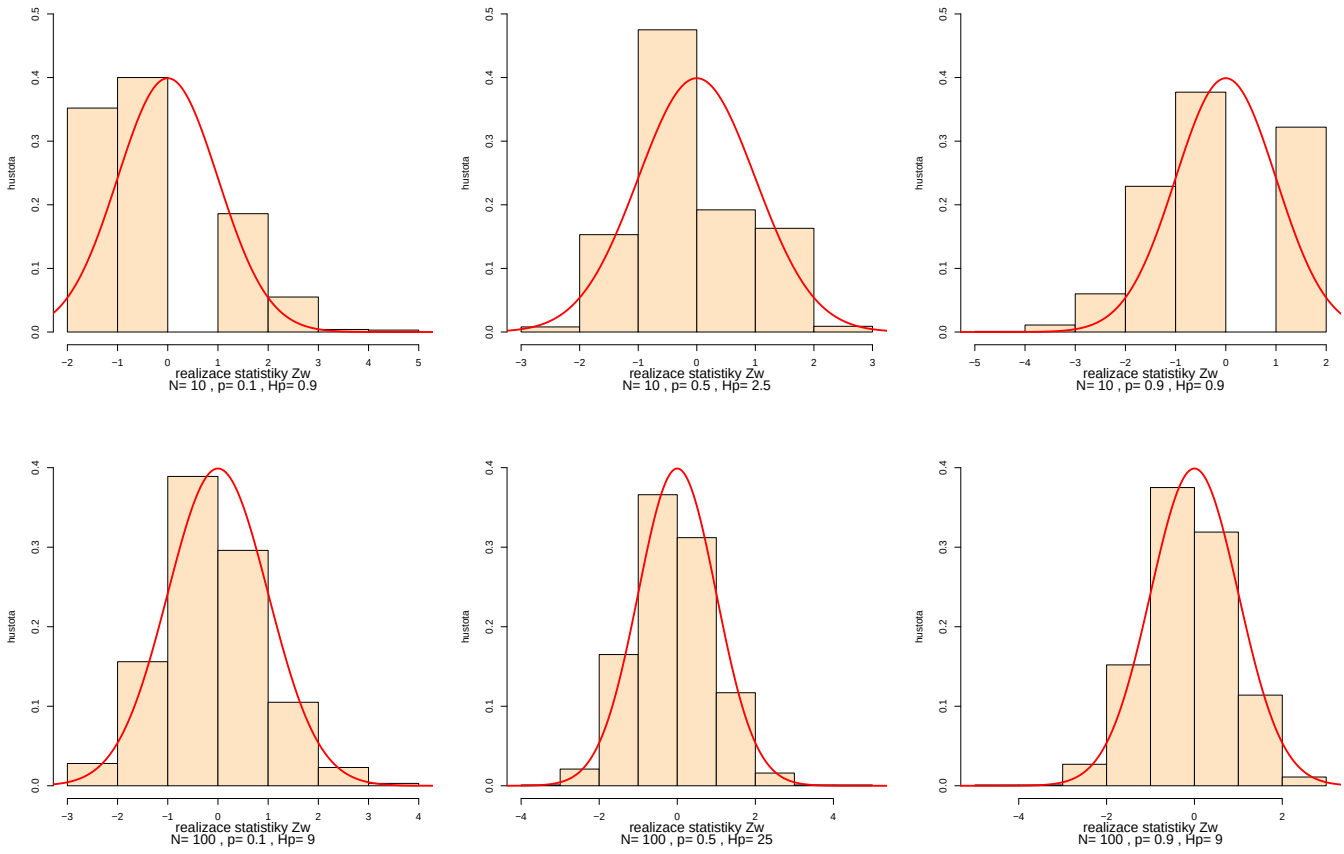
Příklad č.35 (statistika)
Mějme náhodný výběr $(X_1, X_2, \ldots, X_n)^T$, kde $X_i \in \mathbb{R}, i = 1, 2, \ldots, n$, potom příklady statistik jsou:

- $T_1 = \sum_{i=1}^{n} X_i \in \mathbb{R}$,
- $T_2 = \sum_{i=1}^{n} X_i^2 \in \mathbb{R}^+ \cup \{0\}$,
- $T_3 = (\sum_{i=1}^{n} X_1, \sum_{i=1}^{n} X_1^2) \in \mathbb{R}^2$.

Příklad č.36 (testovací statistika, simulační studie)
Na základě simulační studie prověřte, že pokud náhodná proměnná X má asymptoticky binomické rozdělení $Bin(N, p)$, potom testovací statistika

$$Z_W = \frac{X/N - p}{\sqrt{p(1-p)/N}}$$

má asymptoticky normální rozdělení $N(0, 1)$. Použijte $p = 0, 0.1, 0.5, 0.9$ a 1, a $N = 5, 10, 30, 50$ a 100. Okomentujte výsledky ve spojitosti s Haldovou podmínkou $Np(1-p) > 9$. Pro každou simulaci X vypočítejte $z_{W,m}$, $m = 1, 2, \ldots, M$, kde $M = 1000$. Superponujte histogram vygenerovaných testovacích statistik v relativní škále s teoretickou křivkou hustoty Z_W.

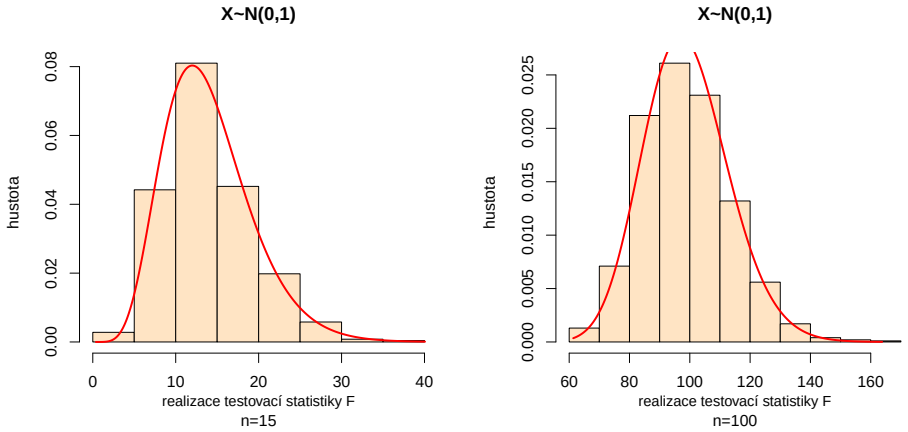


Obrázek 11: Histogram vygenerovaných testovacích statistik v relativní škále superponovaný s teoretickými křivkami hustoty.

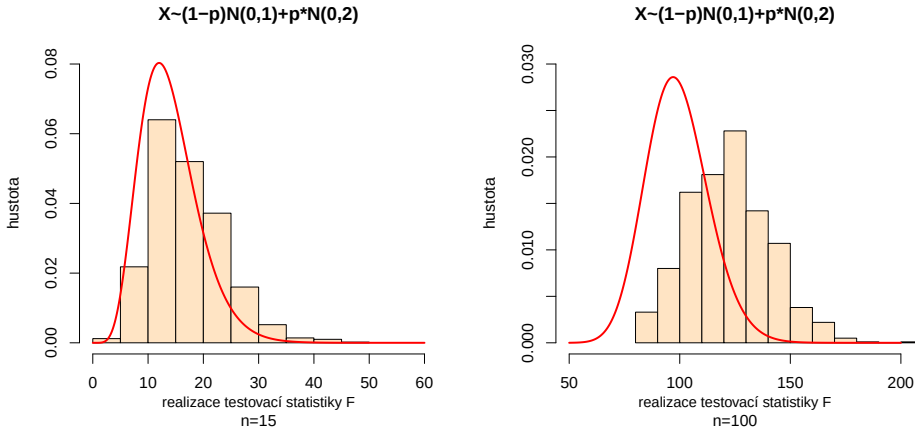
Příklad č.36 mluví o použití jednovýběrové testovací statistiky pro parametr binomického rozdělení (pravděpodob pro různé pravděpodobnosti a různé početnosti. Pokud není Haldova odmínka splněná,není možné testovací

statistiku použít.

Příklad č.37 (testovací statistika, simulační studie)
Na základě simulační studie prověřte, že pokud (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 $\mu = 0$, $\sigma^2 = 1$, $p = 0.05$, $\sigma_1^2 = 2$, potom testovací statistika $F = \frac{(n-1)S^2}{\sigma^2}$ má asymptoticky χ^2_{n-1} rozdělení o $n-1$ stupních volnosti. Použijte rozsahy náhodných výběrů $n = 15$ a $n = 100$. Pro každou simulaci X vypočítejte $F_{poz,m}$, $m = 1, 2, \ldots, M$, kde $M = 1000$. Superponujte histogram vygenerovaných testovacích statistik v relativní škále s teoretickou křivkou hustoty F.



Obrázek 12: Histogram vygenerovaných testovacích statistik v relativní škále superponovaný s teoretickými křivkami hustoty $N(0, 1)$.



Obrázek 13: Histogram vygenerovaných testovacích statistik v relativní škále superponovaný s teoretickými křivkami hustoty $(1-p)N(0, 1) + pN(0, 2)$.

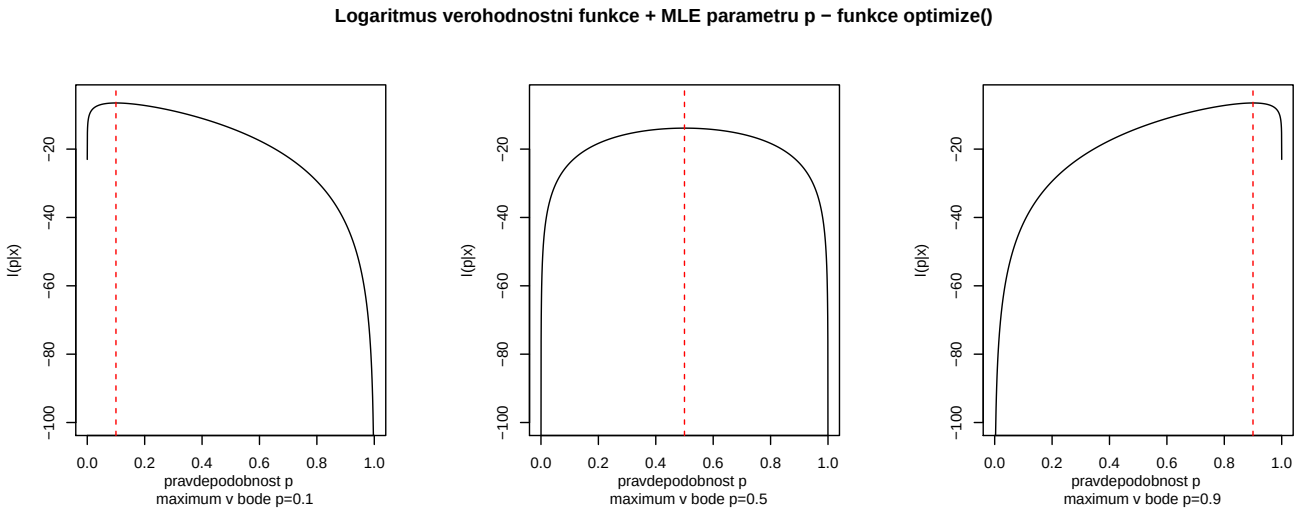
Příklad č.38 (postačující statistika binomického rozdělení)
Nechť X_i, $i = 1, 2, \ldots, N$ jsou iid Bernoulliho pokusy a $X = \sum_{i=1}^{N} X_i$. Potom $X \sim Bin(N, p)$. Ukažte, že $T(\mathbf{X}) = \sum_{i=1}^{N} X_i$ je postačující statistika pro p.

Příklad č.39 (postačující statistika normálního rozdělení)
Nechť $X_i \sim N(\mu, \sigma^2)$, kde $i = 1, 2, \ldots, N$ jsou iid proměnné a σ^2 poznáme. Ukažte, že $T(\mathbf{X}) = \sum_{i=1}^{N} X_i/N =$

$\bar{X}$ je postačující statistika pro μ.

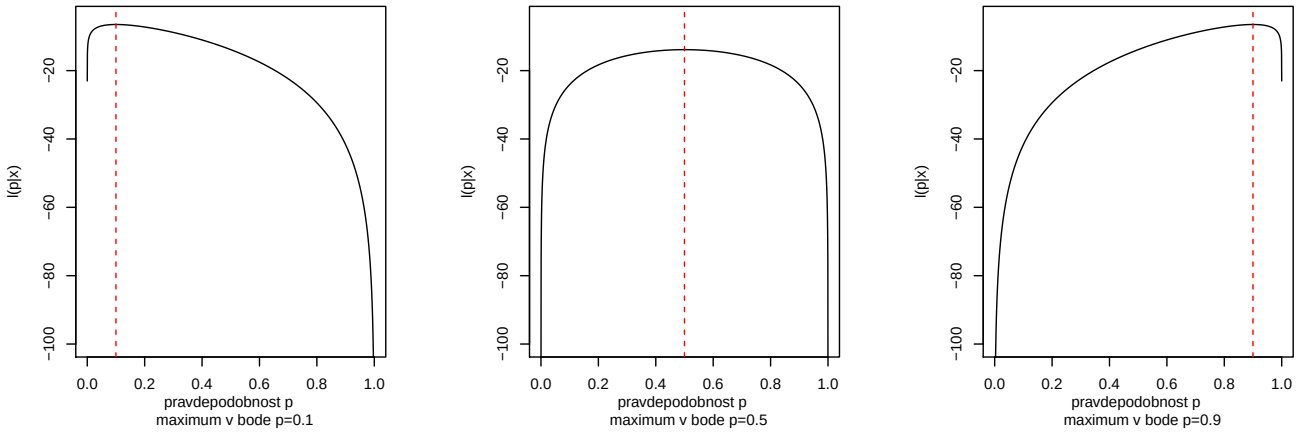
Příklad č.40 (binomické rozdělení; maximálně věrohodný odhad p)
Nechť $X \sim Bin(N, p)$ a realizace X jsou $x = n$. Předpokládejme, že jsme pozorovali (a) $x = 2$, (b) $x = 10$ a (c) $x = 18$ úspěchů v $N = 20$ pokusech.

(a) Pomocí R vypočítejte maximálně věrohodný odhad p. Výsledek zobrazte do grafu spolu s logaritmickou funkcí věrohodnosti (viz graf 14).



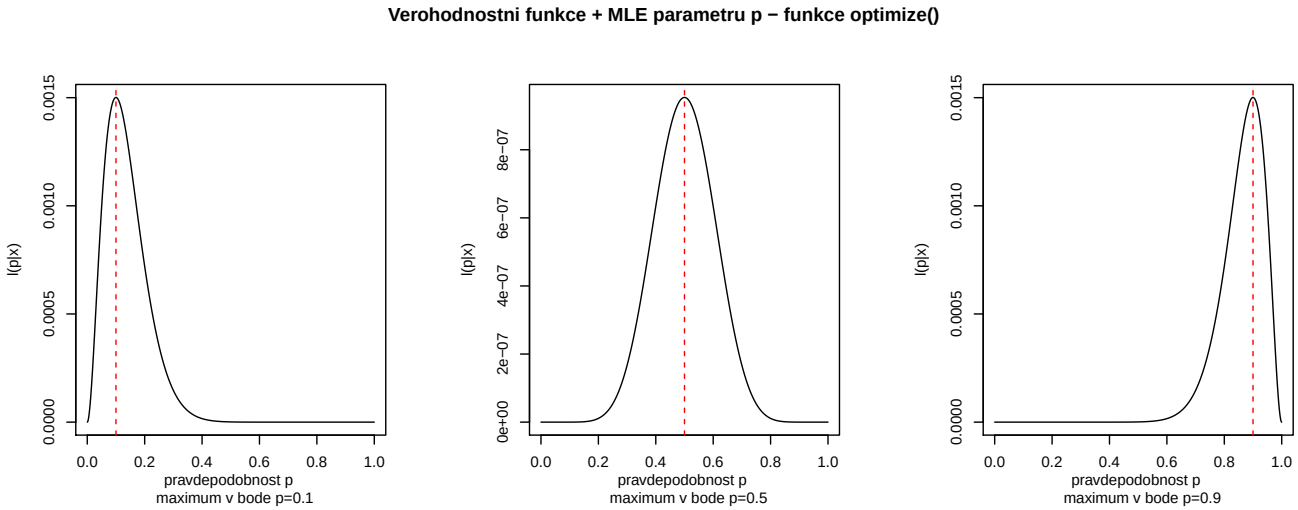
Obrázek 14: Logaritmická funkce věrohodnosti pro $X \sim Bin(N, p)(p = 0.1; 0.5; 0.9$ a $N = 20)$ - funkce optimize()

(b) Naprogramujte Newton-Raphsonovu iterační metodu. Touto metodou nahraďte funkci optimize(), nalezněte maximálně věrohodný odhad parametru p, výsledek zaneste do grafu spolu s logaritmickou funkcí věrohodnosti (grafy budou stejné, jako grafy vygenerované v části (a)).



Obrázek 15: Logaritmická funkce věrohodnosti pro $X \sim Bin(N, p)(p = 0.1; 0.5; 0.9$ a $N = 20)$ - Newton-Raphsonova metoda

(c) Pomocí R vypočítejte maximálně věrohodný odhad p. Výsledek zobrazte do grafu spolu s funkcí věrohodnosti (viz graf 16).



Obrázek 16: Funkce věrohodnosti pro $X \sim Bin(N, p)(p = 0.1; 0.5; 0.9$ a $N = 20)$ - funkce optimize()

Příklad č.41 ($\mathcal{I}(\widehat{p})$ a rozptyl pro p; $X \sim Bin(N, p)$)
Z funkce věrohodnosti odvoďte pozorovanou Fisherovu míru informace $\mathcal{I}(\widehat{p})$ a rozptyl $\widehat{Var[\widehat{p}]}$.

Příklad č.42 (maximálně věrohodné odhady; Poissonovo rozdělení)
Každý rok za posledních pět let byly v nějakém městě registrovány $3, 2, 5, 0$ a 4 zemětřesení za rok. Za předpokladu, že počet zemětřesení za rok (náh. veličina X) má Poissonovo rozdělení s parametrem λ, tj. $X \sim Poiss(\lambda)$:

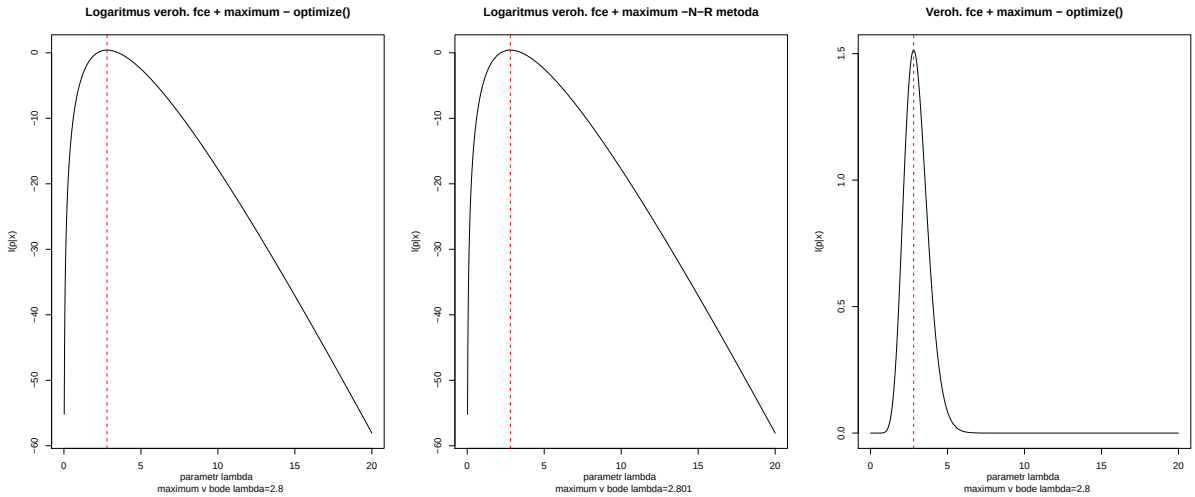
(a) Odvoďte obecný tvar maximálně věrohodného odhadu parametru λ a vypočítejte hodnotu tohoto parametru pro počet zemětřesení (λ představuje očekávanou početnost zemětřesení za rok).

(b) Odvoďte obecný tvar maximálně věrohodného odhadu rozptylu odhadu parametru λ a vypočítejte hodnotu tohoto odhadu rozptylu pro počet zemětřesení.

(c) Vykreslete maximálně věrohodný odhad parametru λ spolu s logaritmickou věrohodnostní funkcí Poissonova rozdělení (parametr λ odhadněte pomocí funkce optimize()).

(d) Vykreslete maximálně věrohodný odhad parametru λ spolu s logaritmickou věrohodnostní funkcí Poissonova rozdělení (parametr λ odhadněte pomocí vámi naprogramované Newtonovy-Raphsonovy iterační metody).

(e) Vykreslete maximálně věrohodný odhad parametru λ spolu s věrohodnostní funkcí Poissonova rozdělení.

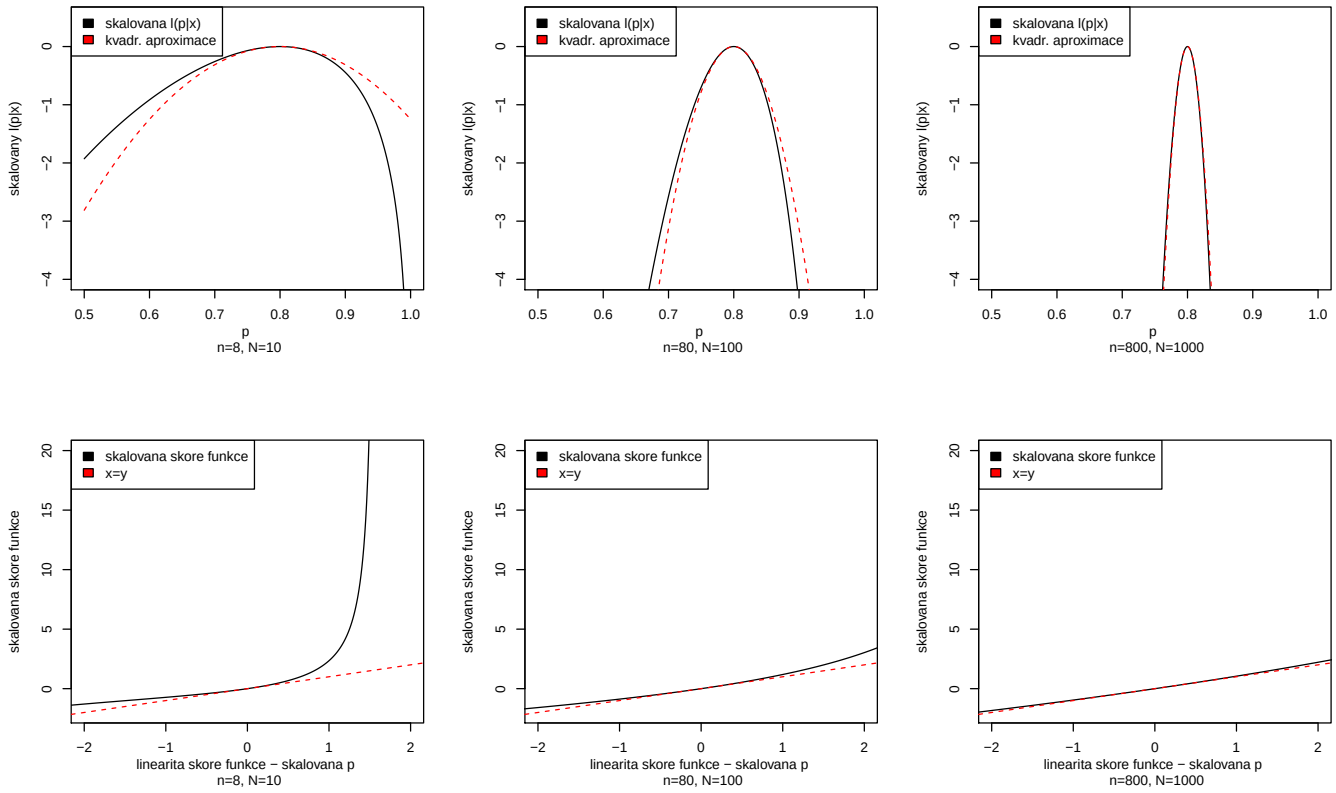


Obrázek 17: $X \sim Po(\lambda)$, zemětřesení: (a) Logaritmická věrohodnostní funkce (použití fce optimize()); (b) Logaritmická věrohodnostní funkce - Newton-Raphsonova metoda; (c) Věrohodnostní funkce - (použití funkce optimize())

Příklad č.43 (kvadratická aproximace logaritmu funkce věrohodnosti)

1. Nakreslete škálovaný logaritmus funkce věrohodnosti binomického rozdělení. Na x-ové ose bude p a na y-ové ose $\ln \mathcal{L}(p) = l(p|\mathbf{x}) - \max(l(p|\mathbf{x}))$. Porovnejte $\ln \mathcal{L}(p)$ s kvadratickou aproximací vypočítanou pomocí Taylorova rozvoje $\ln \mathcal{L}(p) = \ln\left(\frac{L(p|\mathbf{x})}{L(\widehat{p}|\mathbf{x})}\right) \approx -\frac{1}{2}\mathcal{I}(\widehat{p})(p-\widehat{p})^2$.

2. Nechť skóre funkce $S(p) = \frac{\partial}{\partial p} \ln L(p|\mathbf{x})$. Vezmeme-li derivaci kvadratické aproximace uvedené výše, dostaneme $S(p) = -\mathcal{I}(\widehat{p})(p-\widehat{p})$ anebo $-\mathcal{I}^{-1/2}(\widehat{p})S(p) \approx \mathcal{I}^{1/2}(\widehat{p})(p-\widehat{p})$. Potom zobrazením pravé strany na x-ové ose a levé strany na y-ové ose dostaneme asymptoticky lineární funkci s jednotkovým sklonem. Asymptoticky také platí $\mathcal{I}^{1/2}(\widehat{p})(p-\widehat{p}) \sim N(0,1)$, Je postačující mít rozsah x-ové osy $\langle -2; 2\rangle$, protože funkce je asymptoticky (lokálně) lineární na tomto intervalu. Rozumně škálujte y-vou osu. Zobrazte pro (a) $n = 8$, $N = 10$, (b) $n = 80$, $N = 100$ a (c) $n = 800$, $N = 1000$ ($p \in (0.5; 0.99)$). Okomentujte rozdíly mezi (a), (b) a (c). Grafické řešení je na obrázku 18.



Obrázek 18: Porovnání škálovaného logaritmu funkce věrohodnosi s jeho kvadratickou aproximací a prvním řádku a porovnání škálované skóre funkce a přímky $x = y$ v druhém řádku

Příklad č.44 (maximálně věrohodné odhady; binomické rozdělení
Za předpokladu, že náhodná proměnná X má binomické rozdělení, vypočítejte maximálně věrohodný odhad $\widehat{p}$ pomocí logaritmu funkce věrohodnosti $l(p|\mathbf{x})$. Porovnejte tento odhad s výrazem $\sum_{i=1}^{N} x_i/N$. Realizacemi náhodné proměnné X jsou následující binární proměnné:

(a) pohlaví (sex; data: one-sample-probability-sexratio.txt, kde označení pohlaví 'dívka' ('f') přeznačíme na 1 a označení pohlaví 'chlapec' ('m') přeznačíme na 0;

(b) pohlaví (sex; data: two-samples-probabilities-sexratio.txt), kde označení pohlaví 'muž' ('m') přeznačíme na 1 a označení pohlaví 'žena' ('f') přeznačíme na 0.

V případě (a) počítáme pravděpodobnost výskytu děvčat a v případě (b) pravděpodobnost výskytu chlapců.

Příklad č.45 (maximálně věrohodné odhady; multinomické rozdělení)
Nechť náhodná proměnná X má multinomické rozdělení. Potom vypočítejte maximálně věrohodné odhady $\widehat{p}_1$ a $\widehat{p}_2$ pomocí logaritmu funkce věrohodnosti $l(\mathbf{p}|\mathbf{x})$. Porovnejte odhad p_1 s odhadem p z příkladu 44, kde pravděpodobnost $\widehat{p}_1$ byla označená jako $\widehat{p}$. Realizacemi X jsou binární proměnné:

(a) pohlaví (sex; data: one-sample-probability-sexratio.txt, kde označení pohlaví 'dívka' ('f') přeznačíme na 1 a označení pohlaví 'chlapec' ('m') přeznačíme na 0;

(b) pohlaví (sex; data: two-samples-probabilities-sexratio.txt), kde označení pohlaví 'muž' ('m') přeznačíme na 1 a označení pohlaví 'žena' ('f') přeznačíme na 0.

Pravděpodobnost $\widehat{p}_1$ je (a) pravděpodobnost výskytu děvčat a (b) pravděpodobnost výskytu chlapců.

Příklad č. 45 ukazuje, že parametr p_1 dvourozměrného multinomického rozdělení je parametrem p binomického rozdělení.