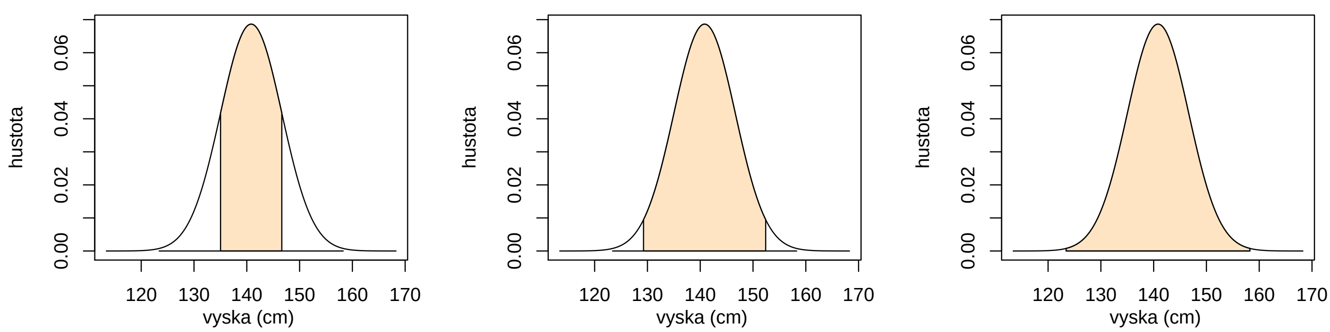
Zadání příkladů - cvičení č.1 - 15-9-23

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.

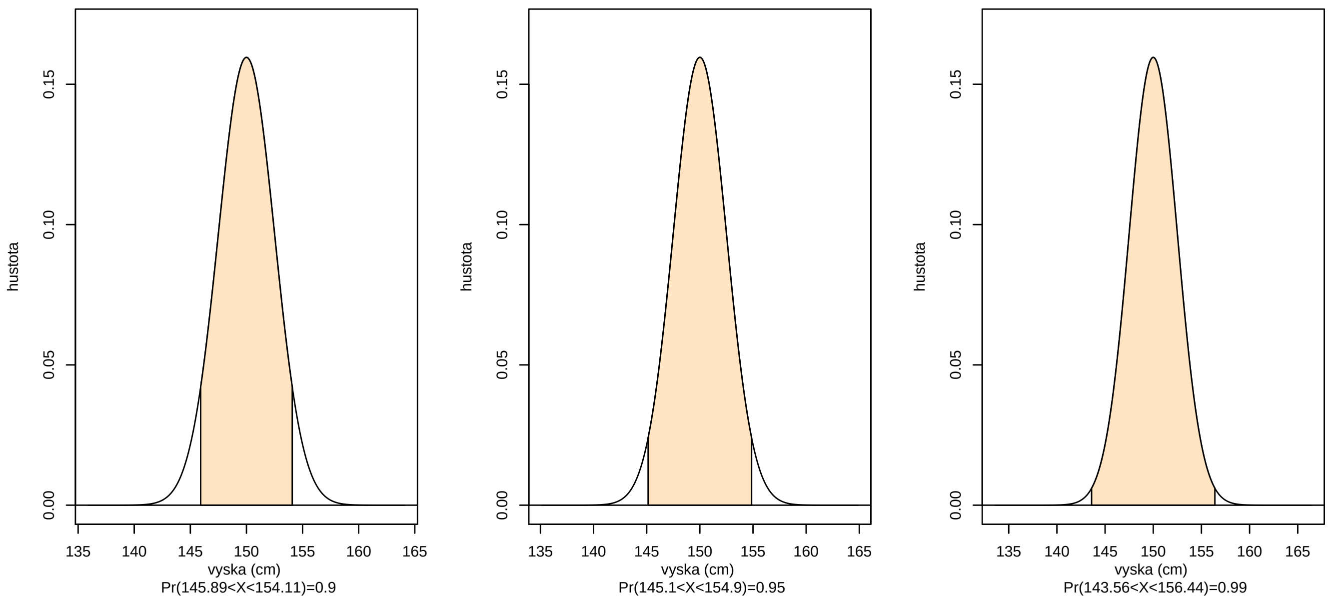


Pr(135.02<X<146.64)=68.27 Pr(129.2<X<152.46)=95.45 Pr(123.39<X<158.27)=99.73

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 $\Pr(a \le X \le b) = 1 - \alpha$, 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.

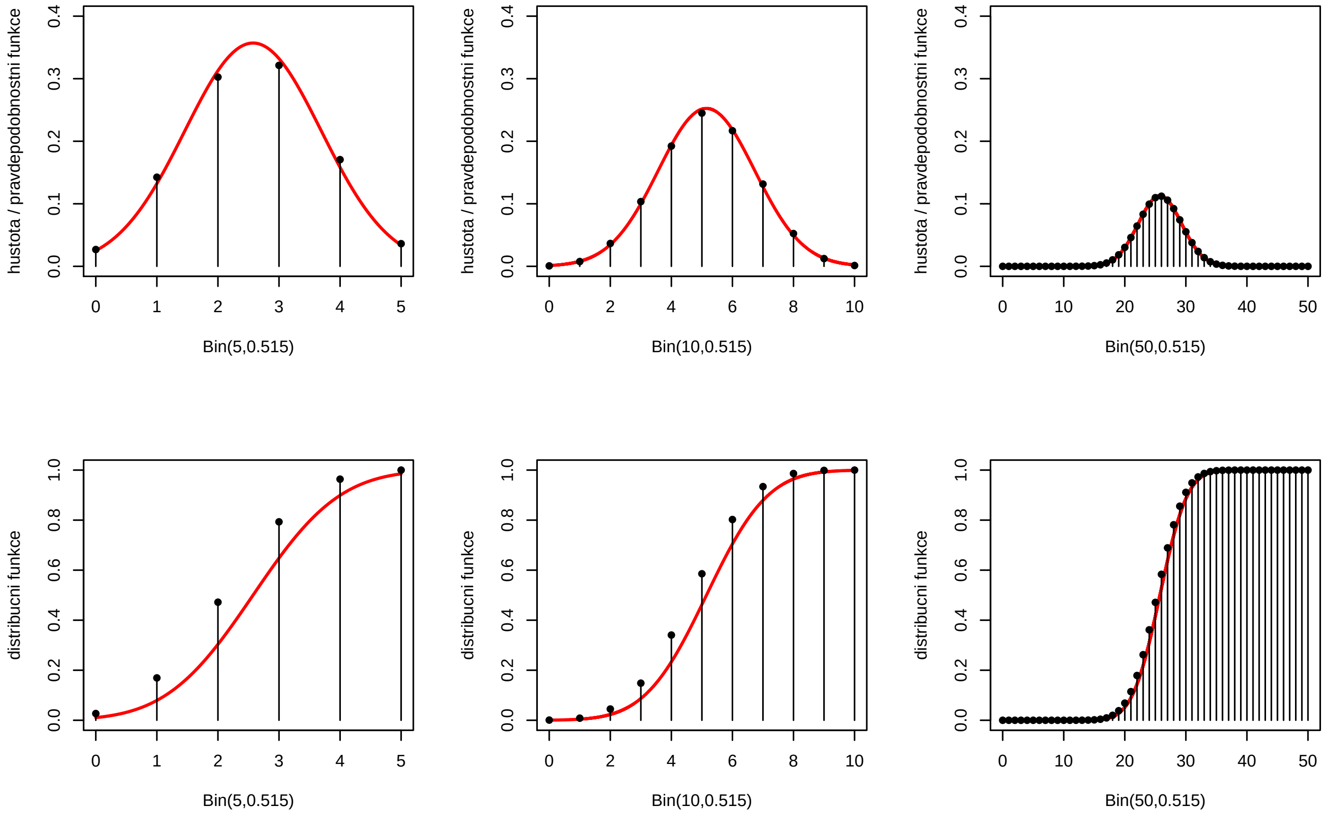
Příklad č.10 (binomické rozdělení) *(přednáška)*
Pokud $X \sim Bin(N, \theta)$, $\theta = p \in \langle 0; 1 \rangle$, potom $\mathcal{Y}_\theta$ je stejný pro všechny θ a koinciduje s výběrovým prostorem $\mathcal{Y} = \{0, 1, \ldots, N\}$.

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?

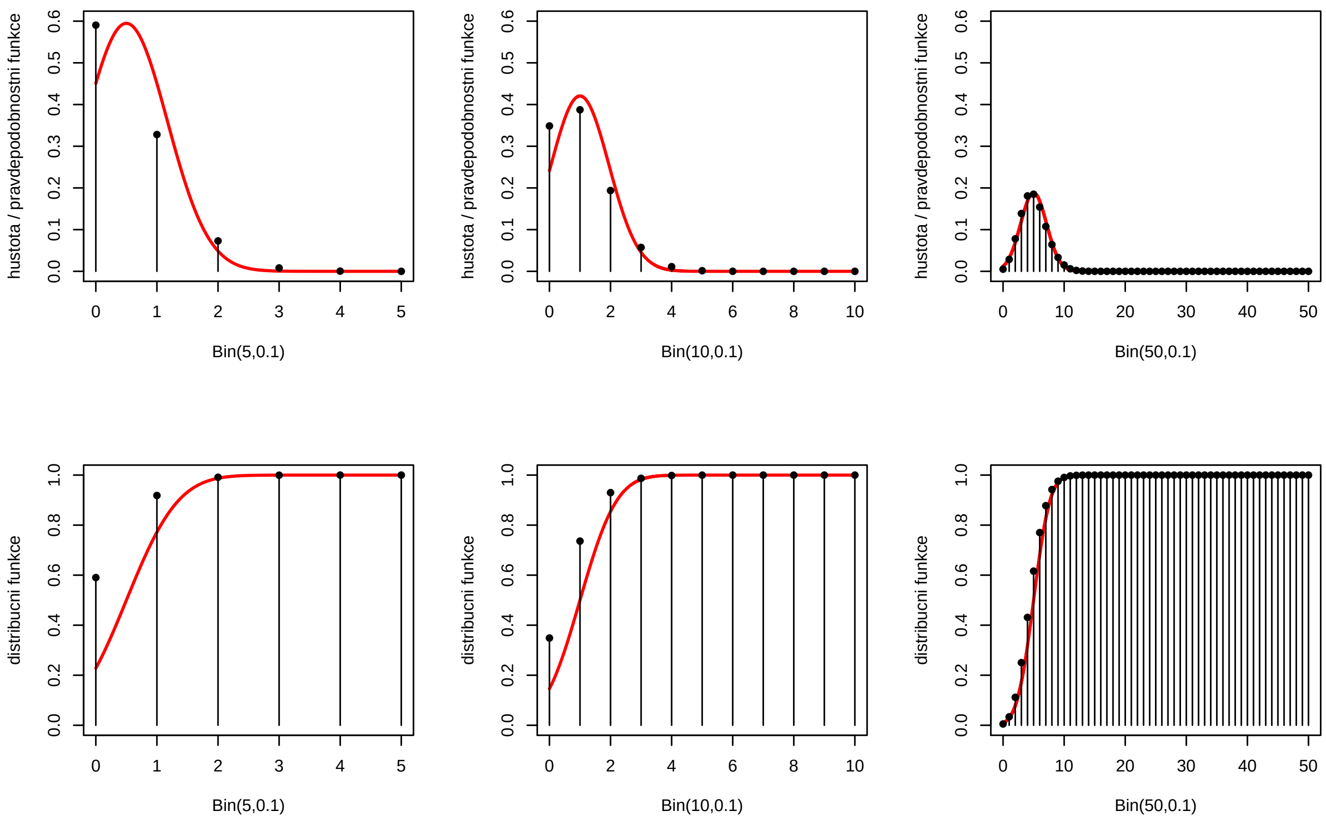
Příklad č.11 (aproximace binomického rozdělení normálním)
Nechť $\Pr(\text{muz}) = p = 0.515$ znamená pravděpodobnost výskytu mužů v populaci a $\Pr(\text{zena}) = q = 0.485$ pravděpodobnost výskytu žen. Nechť X je počet mužů a Y počet žen. Za předpokladu modelu $Bin(N, p)$ vypočítejte (a) $\Pr(X \leq 3)$ 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

$$-\frac{1}{2}\left(\begin{array}{c} x-\mu_1 \\ y-\mu_2 \end{array}\right)^T \left(\begin{array}{cc} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{array}\right)^{-1} \left(\begin{array}{c} x-\mu_1 \\ y-\mu_2 \end{array}\right).$$

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

Příklad č.15 (dvojrozměrné normální rozdělení)

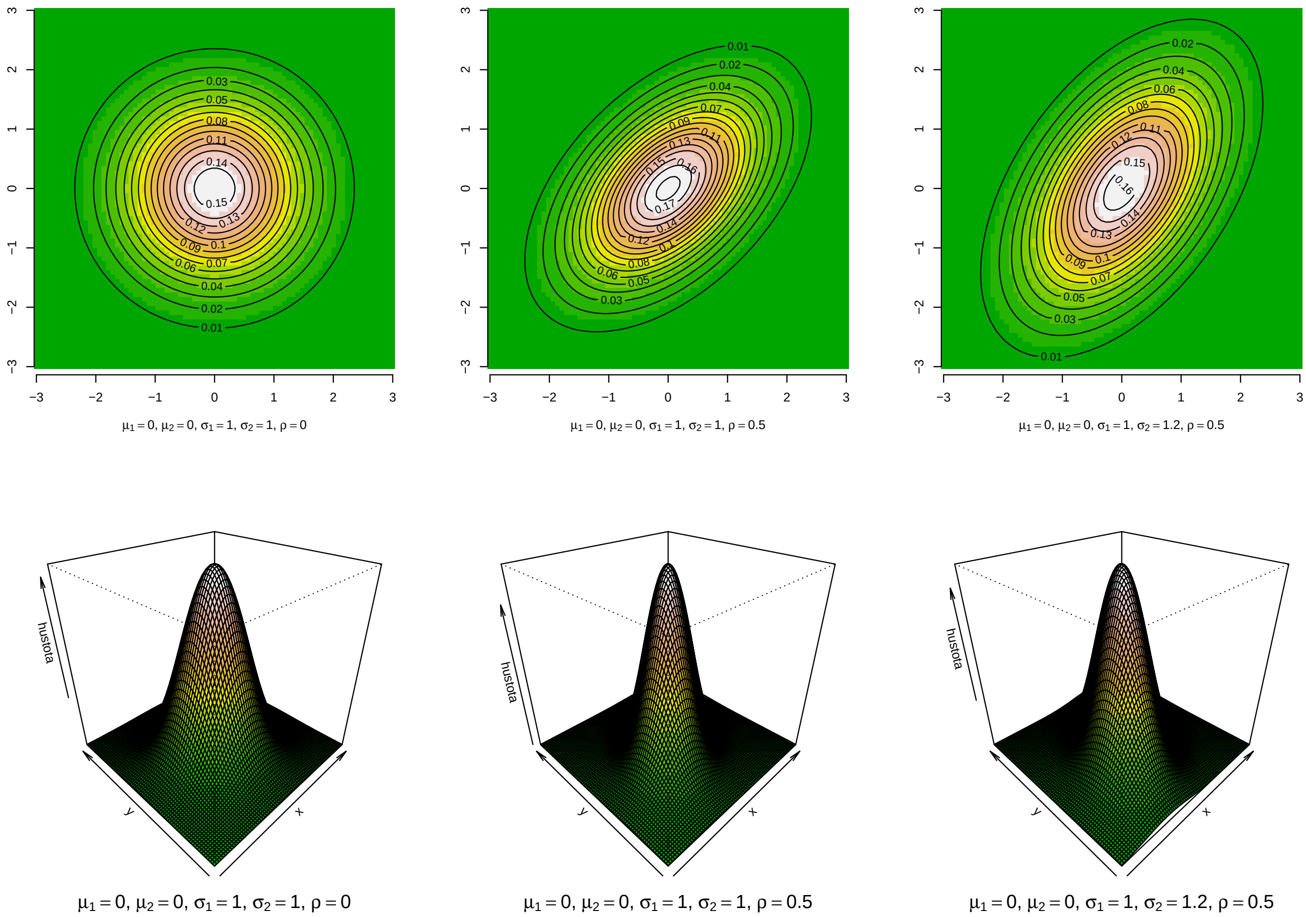
(1) Nakreslete hustotu dvojrozměrného normálního rozdělení $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pomocí funkce `image()` a superponujte ho s konturovým grafem hustoty toho stejného rozdělení pomocí funkce `contour()`. (2) Nakreslete hustotu dvojrozměrného normálního rozdělení $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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

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

s hustotou

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

kde $(x, y)^T \in \mathbb{R}^2$, $\rho \in \langle -1, 1\rangle$ 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:

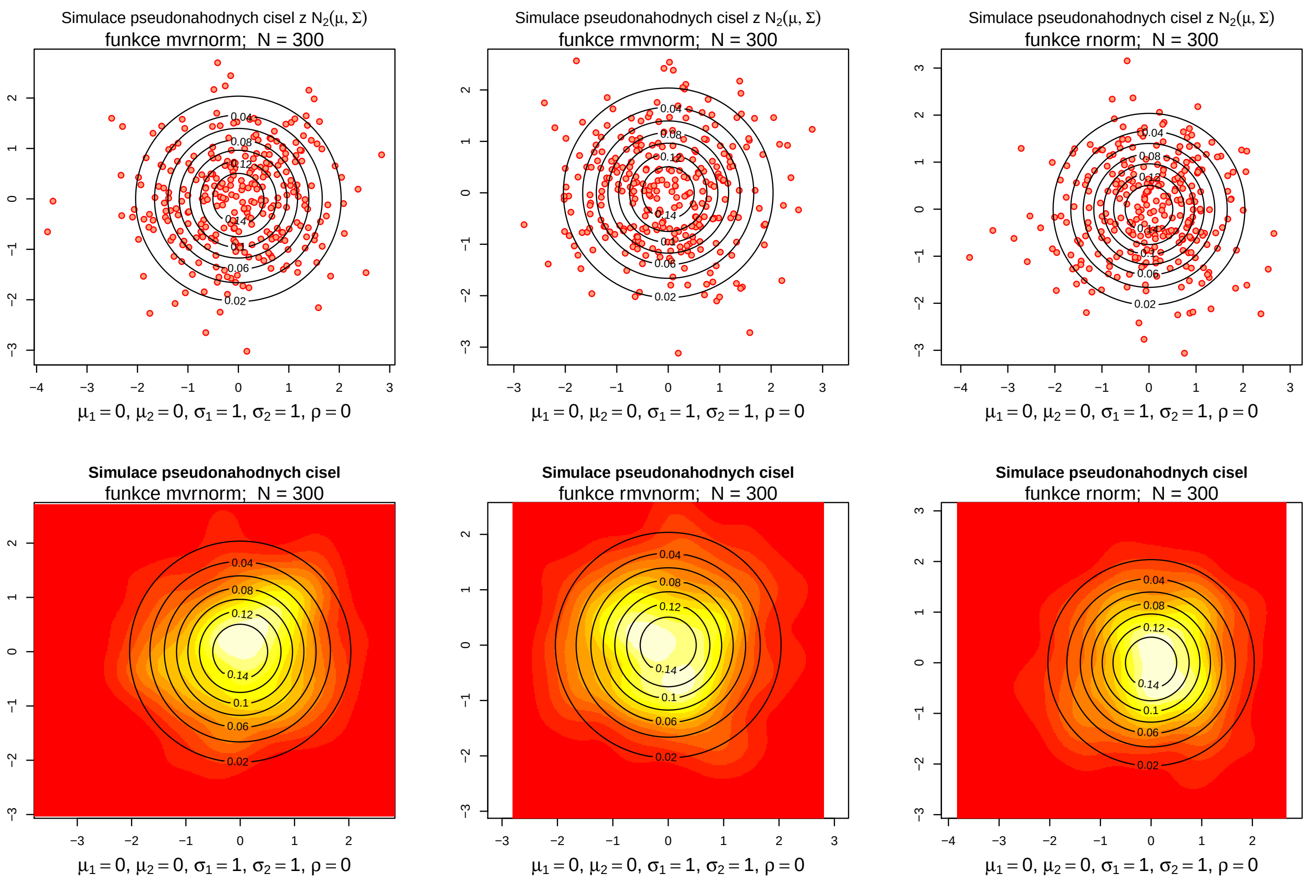
1. použitím funkce mvrnorm() z knihovny MASS;
2. použitím funkce rmvnorm() z knihovny mvtnorm
3. použitím funkce rnorm() a následujícího algoritmu:
   Nechť $X_1 \sim N(0,1)$ a $X_2 \sim N(0,1)$; potom $(Y_1, Y_2)^T \sim N_2(\boldsymbol{\mu}, \mathbf{\Sigma})$, kde $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ 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.

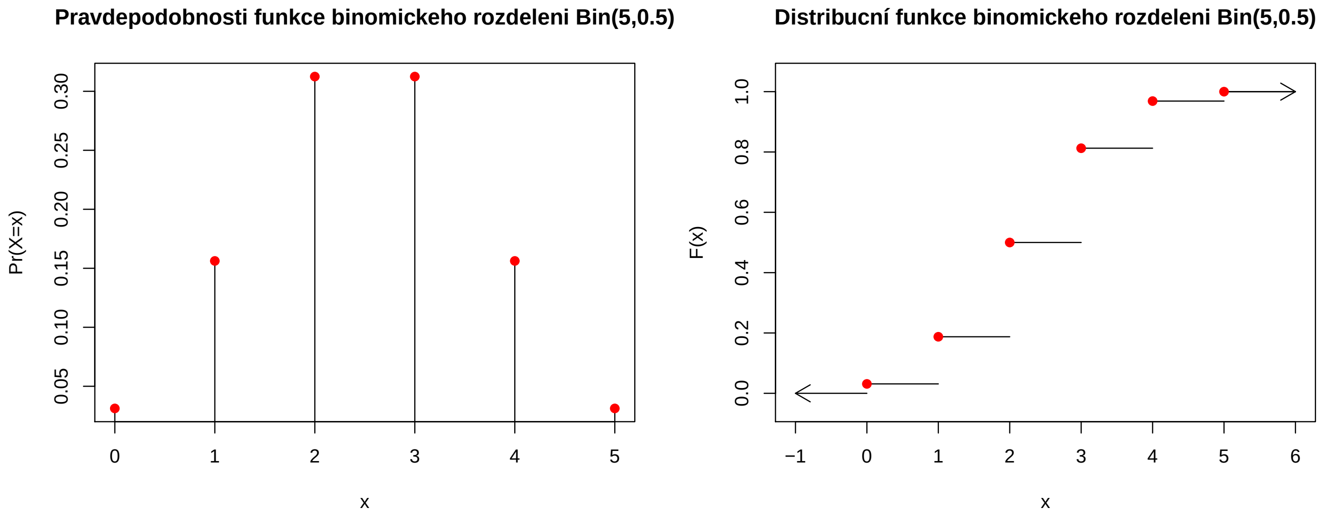


Obrázek 6: Hustoty dvourozměrného normálního rozdělení

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 $Bin(5, 0.5)$