Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference I

Téma 8: Dvourozměrný normální model

Veronika Bendová

bendova.veroonika@gmail.com

Dvourozměrné normální rozdělení $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ ... dvojice nezávislých stejně rozdělených náhodných (iid) veličin
- $(X, Y)^T$ ... dvourozměrný náhodný vektor
- Dvourozměrné normální rozdělení
  - $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$
  - hustota

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

$$= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} e^{-\frac{1}{2}\begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix}}$$

$$= \frac{1}{2\pi(\det(\boldsymbol{\Sigma}))^{1/2}} e^{-\frac{1}{2}U^T\boldsymbol{\Sigma}^{-1}U},$$

  kde $(x, y)^T \in \mathbb{R}$, $(\mu_1, \mu_2)^T \in \mathbb{R}$, $\sigma_j^2 > 0$, $j = 1, 2$ a korelační koeficient $\rho \in \langle -1; 1 \rangle$.
  - marginální rozdělení $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$.
  - mvtnorm::dmvnorm(x, Mu, Sigma), mvtnorm::rmvnorm(n, Mu, Sigma)

- Standardizované dvourozměrné normální rozdělení
  - $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $\boldsymbol{\theta} = (0, 0, 1, 1, \rho)^T$
  - hustota

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

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix}}$$

    kde $(x, y)^T \in \mathbb{R}$ a korelační koeficient $\rho \in \langle -1; 1 \rangle$.
  - margiální rozdělení $X \sim N(0, 1)$, $Y \sim N(0, 1)$.
- **Dataset 7: 13-two-samples-correlations-trunk.txt**
  - Máme k dispozici soubor hodnot délky trupu (rozdíl akrominální a spinální výšky těla) a délky dolní končetiny (spinální výška těla) mladých dospělých jedinců, převážně studentů vysokých škol z Brna a Ostravy (Králík, nepublikovaná data).
  - Přehled proměnných v datasetu:
    - sex . . . pohlaví jedince (m - muž, f - žena);
    - lowex.L . . . délka dolní končetiny (v mm);
    - tru.L . . . délka trupu (v mm).

Příklad 8.1. Hustota dvourozměrného normálního modelu

Naprogramujte v R funkci dnorm2(), jejímiž vstupy budou hodnoty x, y, μ_1, μ_2, σ_1, σ_2 a ρ a výstupem bude hodnota hustoty dvourozměrného normálního rozdělení $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ s parametry $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ a $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ v hodnotách x a y. Správnost funkce otestujte na výpočtu $f(x, y)$ pro (a) $x = 1$, $y = 1$, $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = 0.5$; (b) $x = 2.5$, $y = 1.5$, $\mu_1 = 3$, $\mu_2 = 2$, $\sigma_1^2 = 9$, $\sigma_2^2 = 16$, $\rho = 0.75$. Výsledky ověřte s výsledky funkce dmvnorm() z knihovny mvtnorm.

Řešení příkladu 8.1

```
1  dnorm2 <- function(x, y, mu1, mu2, s1, s2, rho){
2    U <- c(...) # vektor (x - mu1, y - mu2)^T
3    S <- matrix(...) # variancni matice Sigma
4    Z <- ... # hustota f(x, y) rozdeleni N2(Mu, Sigma)
5    return(...)
6  }
7  dnorm2(...) # hustota N2(Mu, Sig); fce dnorm2()
8  mvtnorm::dmvnorm(x = ..., mean = ...,
9                   sigma = matrix(...)) # hustota N2(Mu, Sigma); fce dmvnorm()
```

```
    f(1, 1) f(2.5, 1.5)                                                    10
1 0.0943539  0.01977507                                                    11
```

(a) $f(1, 1) = 0.0944$; (b) $f(2, 3) = 0.0198$.

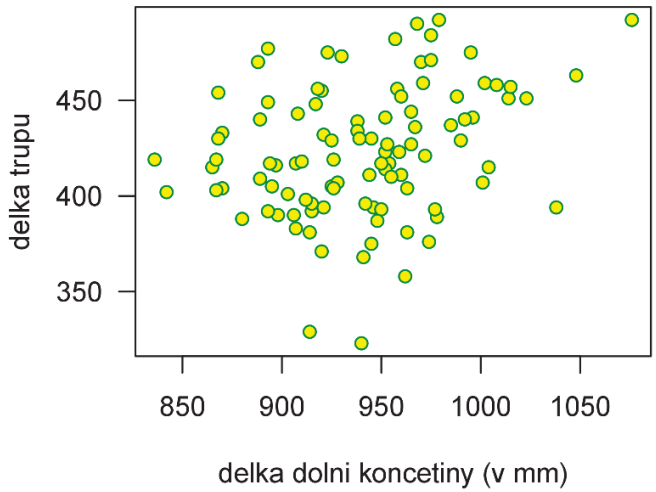
Příklad 8.2. Základní číselné charakteristiky dvojice spojitých znaků
Načtěte datový soubor 13-two-samples-correlations-trunk.txt. Nechť náhodná proměnná X popisuje délku dolní končetiny (v mm) a náhodná proměnná Y popisuje délku trupu u žen. Pomocí tečkového diagramu vizualizujte vztah proměnných X a Y. Za předpokladu, že data pochází z dvourozměrného normálního rozdělení $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ odhadněte hodnoty parametrů μ_1, μ_2, σ_1^2, σ_2^2, σ_{12} a ρ. Výsledky řádně interpretujte.

Řešení příkladu 8.2

```
12  data <- read.delim(..., sep = ..., dec = ...) # nacteni datoveho souboru
13  data.F <- data[data$sex == ..., c('lowex.L', 'tru.L')] # potrebne udaje pro zeny
14  data.F <- na.omit(...) # odstraneni NA
15  lowex.LF <- ... # vyber vektoru delek dolni koncetiny zen
16  tru.LF <- ...  # vyber vektoru delek trupu zen
17  n <- ... # rozsah nahodneho vyberu
18  m1 <- ... # vyberovy prumer delky dolni koncetiny
19  m2 <- ... # vyberovy prumer delky trupu
20  s1 <- ... # sm. odchylka delky dolni koncetiny
21  s2 <- ... # sm. odchylka delky trupu
22  s12 <- cov(..., ...) # vyberova kovariance
23  r12 <- cor(..., ...) # vyberovy koef. korelace
24  tab <- data.frame(..., row.names = ...) # souhrnna tabulka vysledku
```

```
                    n   mean      sd      s12    rho                    25
delka d.koncetiny 100 940.50 45.4712 441.3081 0.2853                    26
delka trupu       100 423.17 34.0229 441.3081 0.2853                    27
```

```
28  par(...) # okraje grafu 4, 5, 1, 1
29  plot(lowex.LF, tru.LF, ...) # teckovy diagram
```



Obrázek: Dvourozměrný tečkový diagram pro délku dolní končetiny a délku trupu

Datový soubor obsahuje údaje o délce dolní končetiny a délce trupu 100 žen. Délka dolní končetiny se pohybuje okolo hodnoty 940.50 mm se směrodatnou odchylkou 45.47 mm. Délka trupu se pohybuje okolo hodnoty 423.17 mm se směrodatnou odchylkou 34.02 mm. Hodnota kovariance $s_{12} = 441.31$. Mezi délkou dolní končetiny a délkou trupu žen existuje nízký stupeň přímé lineární závislosti ($r_{12} = 0.2853$).

Příklad 8.3. Test dvourozměrné normality
Načtěte datový soubor 13-two-samples-correlations-trunk.txt. Nechť náhodná proměnná X popisuje délku dolní končetiny a náhodná proměnná Y popisuje délku trupu žen. Na hladině významnosti $\alpha = 0.05$ testujte hypotézu o dvourozměrné normalitě vektoru $(X, Y)^T$. K otestování použijte Mardiův test.

Řešení příkladu 8.3
Na hladině významnosti $\alpha = 0.05$ testujeme H_0: *Data pochází z dvourozměrného normálního rozdělení.* oproti H_1: *Data nepochází z dvourozměrného normálního rozdělení.*

Mardiův test – sestává ze dvou testů:

(i) test šikmosti
- H_{0a}: Data nejsou kladně ani záporně vyšikmená.
- H_{1a}: Data jsou kladně nebo záporně vyšikmená.

(ii) test špičatosti
- H_{0b}: Data nejsou kladně ani záporně zešpičatělá.
- H_{1b}: Data jsou kladně nebo záporně zešpičatělá.

Poznámka: Náhodný výběr pochází z dvourozměrného normálního rozdělení, pokud nevykazuje kladné ani záporné zešikmení ani kladné nebo záporné zešpičatění.

```
30 MVN::mvn(data.F, mvnTest = 'mardia')$multivariateNormality
```

```
             Test           Statistic            p value Result                    31
1 Mardia Skewness   6.31326657225727 0.176942962210473    YES                     32
2 Mardia Kurtosis -0.207066071208097 0.835958259081491    YES                     33
3             MVN                <NA>               <NA>    YES                  34
```

Protože p-hodnota testu šikmosti $p = 0.1769$ je větší než α, H_{0a} nezamítáme na hladině významnosti $\alpha = 0.05$. Protože p-hodnota testu špičatosti $p = 0.8360$ je větší než α, H_{0b} nezamítáme na hladině významnosti $\alpha = 0.05$. Data nejsou kladně ani záporně zešikmená ani zešpičatělá. Data pochází z dvourozměrného normálního rozdělení.

Příklad 8.4. Vizualizace dat z dvourozměrného normálního modelu

Načtěte datový soubor 13-two-samples-correlations-trunk.txt. Nechť náhodná proměnná X popisuje délku dolní končetiny a náhodná proměnná Y popisuje délku trupu žen. Na základě řešení příkladů 8.2 a 8.3 předpokládáme, že data pochází z dvourozměrného normálního rozdělení $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ s odhadem středních hodnot $\widehat{\mu}_1 = 940.50$, $\widehat{\mu}_2 = 423.17$, s odhadem rozptylů $\widehat{\sigma}_1^2 = 45.47^2$ a $\widehat{\sigma}_2^2 = 34.02^2$ a s odhadem korelačního koeficientu $\widehat{\rho} = 0.2853$.

(a) Sestrojte tečkový diagram pro délku dolní končetiny a délku trupu. Tečkový diagram superponujte (i) konturami hustoty dvourozměrného normálního rozdělení (funkce dnorm2() + contour()); (ii) jádrovým odhadem dvourozměrné hustoty (funkce kde2d() z knihovny MASS + funkce contour()).

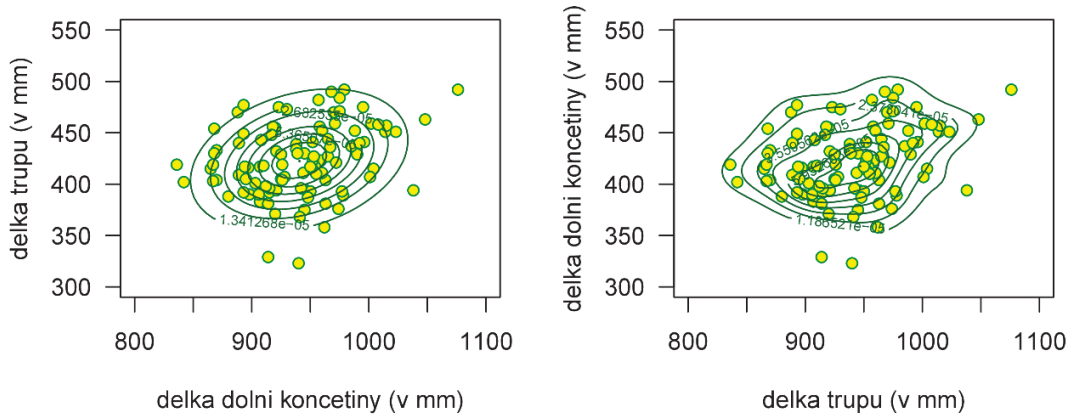
(b) Sestrojte (i) vrstevnicový diagram hustoty dvourozměrného normálního rozdělení délky dolní končetiny a délky trupu superponovaný svými konturami (dnorm2() + image() + contour()); (ii) vrstevnicový diagram jádrového odhadu hustoty délky dolní končetiny a délky trupu superponovaný svými konturami (MASS::kde2d() + image() + contour()).

(c) Sestrojte (i) 3D-diagram hustoty dvourozměrného normálního rozdělení délky dolní končetiny a délky trupu (dnorm2() + persp()); (ii) 3D-diagram jádrového odhadu hustoty dvourozměrného normálního rozdělení délky dolní končetiny a délky trupu (MASS::kde2d() + persp()).

Řešení příkladu 8.4

(a) Tečkové diagramy

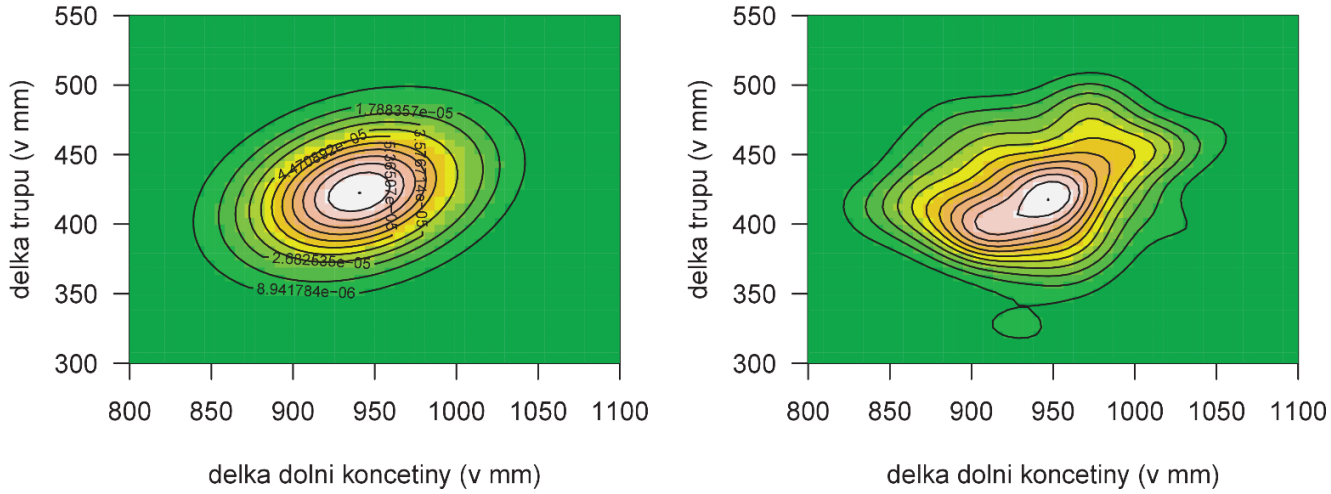
```
n <- 50
x <- seq(...) # posl. bodu osy x; od 800 do 1100 o delce n
y <- seq(...) # posl. bodu osy y; od 300 do 550 o delce n
M <- matrix(NA, ...) # prazdna matice dimenze (n x n)
for(i in 1:length(x)){
  for(j in 1:length(y)){
    M[i, j] <- dnorm2(x[i], y[j], m1, m2, s1, s2, r12)
  }
} # hodnoty hustoty f(x, y) rozdeleni N2(Mu, Sig)
Z <- MASS::kde2d(lowex.LF, tru.LF, lim = c(800, 1100, 300, 550),
                 n = n) # jadrovy odhad hustoty f(x, y)
k <- 8
par(...) # okraje grafu 4, 5, 1, 1
plot(lowex.LF, tru.LF, xlim = c(800, 1100), ylim = c(300, 550),
     ...) # teckovy diagram
contour(x, y, M, drawlabels = T, levels = seq(0, max(M), length = k + 1),
        add = T, col = ...) # kontury hustoty f(x, y) rozdeleni N2(Mu, Sig)
plot(...) # teckovy diagram
contour(Z$x, Z$y, Z$z, ...) # kontury jadroveho odhadu hustoty f(x, y)
```



Obrázek: Dvourozměrný tečkový diagram pro délku dolní končetiny a délku trupu žen superponovaný (a) konturami hustoty dvourozměrného normálního rozdělení (vlevo); (b) jádrovým odhadem hustoty dvourozměrného normálního rozdělení (vpravo)

(b) Vrstevnicové diagramy

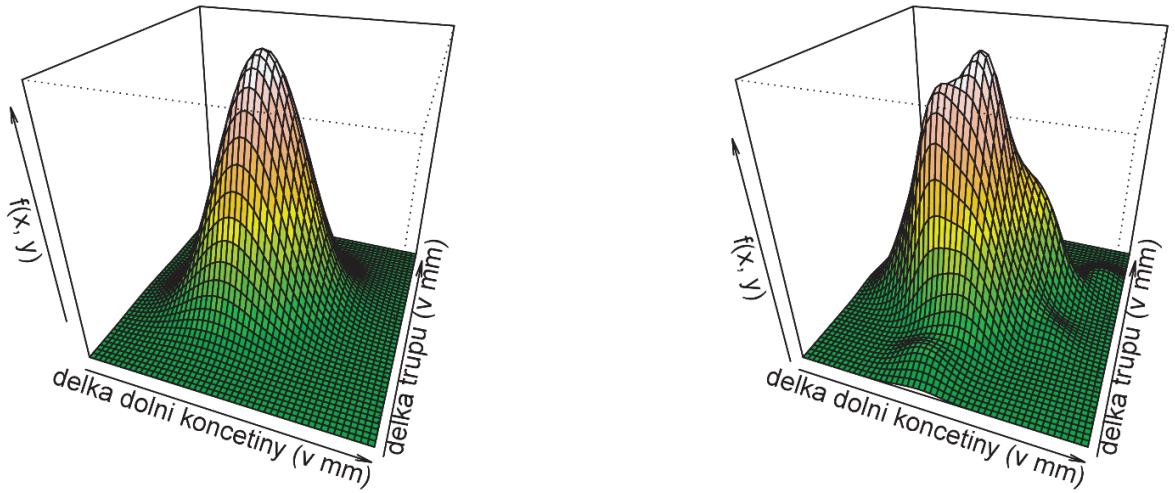
```
54  k <- 12
55  par(...) # okraje grafu 4, 5, 1, 1
56  image(x, y, M, breaks = seq(0, max(M), length = k + 1), xlim = c(800, 1100),
57        ylim = c(300, 550), asp = F, col = terrain.colors(k), las = 1, xlab = ...,
58        ylab = ...) # vrstevnicovy diagram hustoty f(x, y) rozd. N2(Mu, Sig)
59  contour(x, y, M, add = ..., drawlabels = ...,
60          levels = ...) # kontury hustoty f(x, y) rozdeleni N2(Mu, Sig)
61
62  image(Z$x, Z$y, Z$z, ...) # vrstevnicovy diagram jadr. odhadu hustoty f(x, y)
63  contour(...) # kontury jadr. odhadu hustoty f(x, y)
```



Obrázek: (a) Hustota dvourozměrného normálního rozdělení pro délku dolní končetiny a délku trupu žen superponovaná konturami (vlevo); (b) jádrový odhad hustoty dvourozměrného normálního rozdělení superponovaný konturami (vpravo)

(c) 3D-diagramy

```
64  nrz <- nrow(M) # pocet radku matice Z
65  ncz <- ncol(M) # pocet sloupcu matice z
66  color <- terrain.colors(k) # paleta k barev
67  stredy <- (M[-1, -1] + M[-1, -ncz] + M[-nrz, -1] + M[-nrz, -ncz]) / 4
68    # matice stredu site
69  stredy.col <- cut(stredy, k) # rozdel rozpeti hodnot do 12 ekvidistantnich
70    # intervalu a kazde hodnote prirad interval, do ktereho nalezi
71  par(...) # okraje grafu 1, 1, 1, 1
72  persp(x, y, M, col = color[stredy.col], phi = 30, theta = 20, xlab = ...,
73        ylab = ..., zlab = ...) # 3D-diagram hudtoty f(x, y) rozd. N2(Mu,  Sig)
74
75  nrz <- nrow(Z$z)  # pocet radku matice z promenne Z
76  (...)
77  persp(Z$x, Z$y, Z$z, ...) # 3D diagram jadr. odhadu hustoty f(x, y)
```



Obrázek: (a) 3D-diagram hustoty normálního rozdělení pro délku dolní končetiny a délku trupu (vlevo); (b) 3D-diagram jádrového odhadu hustoty dvourozměrného normálního rozdělení (vpravo)

Příklad 8.5. Simulace dat z dvourozměrného normálního rozdělení
Nasimulujte data $(X, Y)^T$ z dvourozměrného normálního rozdělení s parametry $\mu_1 = 940.50$, $\mu_2 = 423.17$, $\sigma_1^2 = 45.47^2$, $\sigma_2^2 = 34.02^2$ a $\rho = 0.2853$; $n = 100$. Simulaci pseudonáhodných čísel z $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ proveďte použitím funkce rmvnorm() z knihovny mvtnorm. Pro nasimulovaná data vykreslete (a) histogram náhodné veličiny X a histogram náhodné veličiny Y, přičemž každý histogram superponujte křivkou marginálního normálního rozdělení $N(\mu_i, \sigma_i^2)$, $i = 1, 2$ a křivkou jádrového odhadu hustoty; (b) tečkový diagram náhodných veličin X a Y, přičemž graf superponujte konturami jádrového odhadu dvourozměrné hustoty.

Řešení příkladu 8.5

```
78  simnorm2 <- function(n, Mu, Sig, method = 'rmvnorm', col1 = ..., col2 = ...,
79                       col3 = ..., k = 12){
80    mu1 <- Mu[1] # vyber mu1 z vektoru Mu
81    mu2 <- ... # vyber mu2 z vektoru Mu
82    sigma1 <- sqrt(Sig[1, 1]) # vyber sm. odchylky s1 z matice Sig
83    sigma2 <- sqrt(Sig[2, 2]) # vyber sm. odchylky s2 z matice Sig
84    if(method == 'rmvnorm') {
85      data <- mvtnorm::rmvnorm(n = ..., mean = ..., sigma = ...) # simulace dat
86      }
```

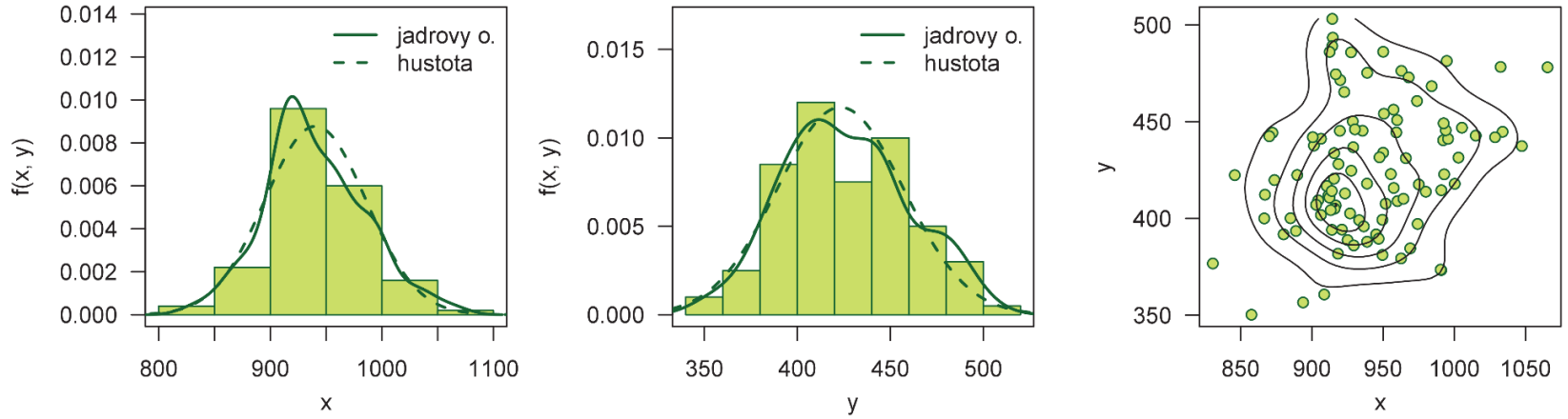
```
  x <- data[, 1] # vyber vektoru x z nasimulovanych dat
  y <- data[, 2] # vyber vektoru y z nasimulovanych dat
  xx <- seq(min(x) - 100, max(x) + 100,
            length = 512) # posl. bodu osy x pro krivku f(x)
  xy <- dnorm(xx, mu1, sigma1) # hustota rozdeleni N(mu1, sigma1^2)
  yx <- seq(...) # posl. bodu osy x pro krivku f(y)
  yy <- ... # hustota rozdeleni N(mu2, sigma2^2)
  Z <- MASS::kde2d(x, y, n = 50) # jadrovy odhad hustoty f(x, y)

  par(...) # okraje grafu 4, 5, 1, 1
  hist(x, prob = T, ylim = c(0, max(xy) + 0.005), ...) # histogram simul. dat x
  box(...) # ramecek okolo grafu
  lines(density(x), ...) # krivka jadroveho odhadu hustoty f(x)
  lines(xx, xy, ...) # krivka hustoty N(mu1, sigma1^2)
  mtext(...) # popisek osy x
  mtext(...) # popisek osy y
  legend(...) # legenda

  hist(...) # histogram nasimulovanych dat y
  (...)

  plot(x, y, ...) # teckovy diagram
  contour(Z$x, Z$y, Z$z, levels = seq(..., ..., length = k + 1),
          ...) # kontury jadr. odhadu hustoty f(x, y)
  mtext(...) # popisek osy x
}

Mu <- c(...) # vektor vyberovych prumeru
Sig <- matrix(...) # kovariancni matice
n <- ... # rozsah nahodneho vyber n
simnorm2(n, Mu, Sig, method = 'rmvnorm', col1 = ..., col2 = ..., col3 = ...,
         k = 6) # provedeni simulace
```



Poznámka: Simulaci dat z dvourozměrného normálního rozdělení lze dále provést

1. pomocí funkce mvrnorm() z knihovny MASS
2. použitím funkce rnorm() a následujícího algoritmu: Nechť $U_1 \sim N(0,1)$ a $U_2 \sim N(0,1)$; potom $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 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 $\boldsymbol{\Sigma}$, přičemž síla lineárního vztahu X a Y je daná velikostí a znaménkem ρ; $X = \sigma_1 U_1 + \mu_1$ a $Y = \sigma_2(\rho U_1 + \sqrt{1-\rho^2} U_2) + \mu_2$.

V případě zájmu si můžete simulaci těmito dvěma způsoby také vyzkoušet.