Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Aplikovaná statistika I

Téma 5: Spojité náhodné veličiny

Veronika Bendová

bendova.veroonika@gmail.com

Úvod a motivace

- víc než výsledek nás často zajímají jeho číselné interpretace
- **náhodná veličina** X = pravidlo, které zobrazuje základní prostor možných výsledků do množiny reálných čísel
- i-tá realizace náh. veličiny X se značí x_i
  - X ... počet puntíků na vrchní straně kostky: $x_1 = 4$, $x_2 = 1$ ...
  - Y ... dokončené vzdělání; $y_1 = 1$ (ZŠ), $y_2 = 3$ (SŠm) ...
  - Y ... počet starších sourozenců $y_1 = 0$, $y_2 = 2$ ...
  - X ... porodní hmotnost v g; $x_1 = 3470$, $x_2 = 3240$ ...
  - Y ... největší šířka mozkovny v mm; $y_1 = 145$, $y_2 = 139$ ...
- dva typy náhodných veličin
  - diskrétní náhodné veličiny
  - spojité náhodné veličiny

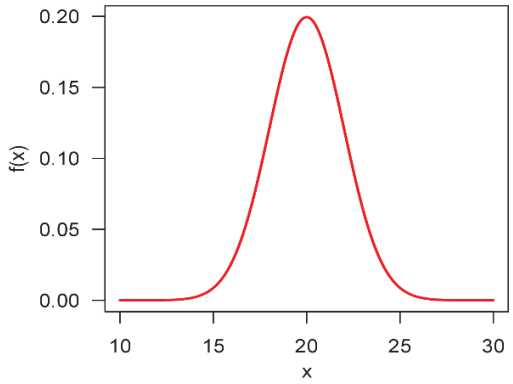
Spojité náhodné veličiny

- ze své podstaty mohou nabývat libovolných i neceločíselných hodnot
  - porodní hmotnost (v g):
    - základní prostor rozdělíme na intervaly: I1: 0–1500; I2: 1500–2500; I3: 2500–3500; I4: 3500–4500; I5:>4500
    - $\Pr(X \in \langle 3500\,;\, 4500 \rangle) = \ldots$
    - $\Pr(X \leq 3500) = \ldots$
    - $\Pr(X > 3500) = \ldots$
- hustota $f(x)$
  - pravděpodobnost realizace X v libovolném intervalu I se dá vyjádřit jako plocha pod křivkou pomocí integrálního tvaru:

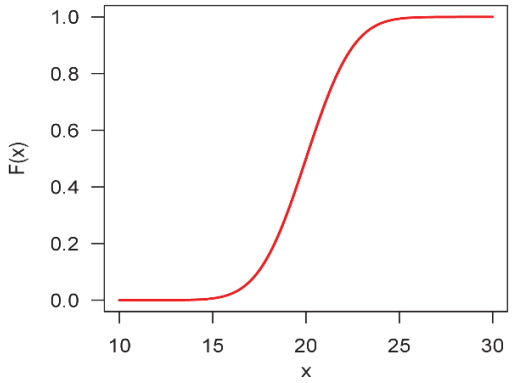
$$\Pr(X \in I) = \int_{x \in I} f(x) dx,$$

kde $f(x)$ je *hustota* pravděpodobnosti spojité náhodné veličiny

- hustota normálního rozdělení (Gaussova křivka)



- nezáporná: $f(x) \geq 0$; normovaná (plocha pod křivkou hustoty $= 1$)
- distribuční funkce $F(x)$
  - $F(x) = \Pr(X \leq x)$



  - $\Pr(X = x) = 0$
  - **Komplementarita**:
    $\Pr(X > x) = \Pr(X \geq x) = 1 - \Pr(X < x) = 1 - \Pr(X \leq x) = 1 - F(x)$

Normální rozdělení

- $X_1, \ldots, X_n$ ... nezávislé náhodné veličiny
- Normální rozdělení
  - $X \sim N(\mu, \sigma^2)$
  - hustota
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}.$$
  - vlastnosti: $\mathrm{E}[X] = \mu$; $\mathrm{Var}[X] = \sigma^2$
  - dnorm(x, mu, sigma), pnorm(x, mu, sigma), qnorm(alpha, mu, sigma)
- Standardizované normální rozdělení
  - $X \sim N(0, 1)$
  - hustota
$$f(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad x \in \mathbb{R}.$$
  - vlastnosti: $\mathrm{E}[X] = 0$; $\mathrm{Var}[X] = 1$
  - dnorm(x), pnorm(x), qnorm(alpha)
- Vlastnosti normálního rozdělení
  - **Věta 1:** Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny z normálního rozdělení $N(\mu, \sigma^2)$. Potom náhodná veličina $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Příklad 5.1. Výpočet parametrů μ a σ normálního rozdělení

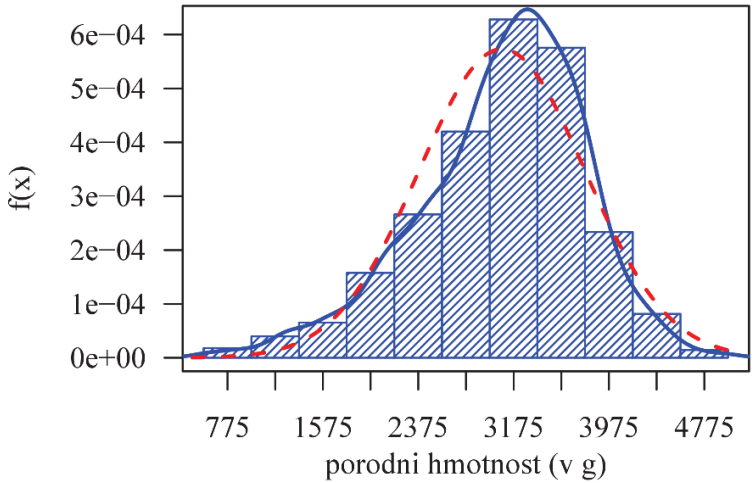
Mějme datový soubor 17-anova-newborns-2.txt obsahujícího údaje o porodní hmotnosti novorozenců v jedné okresní nemocnici za období jednoho roku (Alánová, 2008). Za předpokladu, že náhodná veličina X popisující porodní hmotnost novorozenců pochází z normálního rozdělení $N(\mu, \sigma^2)$ odhadněte parametr střední hodnoty μ a rozptylu σ^2. Finální rozdělení porovnejte s naměřenými údaji.

Řešení příkladu 5.1

```
1  data <- read.delim('17-anova-newborns-2.txt') # nacteni dat
2  head(data, n = 3) # vypis prvnich tri radku
3  data <- na.omit(data) # odstraneni NA hodnot
4  wei<- data$weight.C # vyber por. hmotnosti novorozencu
5  n <- length(wei) # pocet hodnot
6  mu <- mean(wei) # prumer por. hmotnosti novorozencu
7  sigma <- round(sd(wei)) # sm. odchylka por. hmotnosti novorozencu
```

```
  edu.M prch.N sex.C weight.C weight.K                                    8
1     2      0     m     3470        2                                    9
2     2      0     m     3240        2                                    10
3     2      0     f     2980        2                                    11
```

```
12  par(mar = c(4, 5, 1, 1)) # nastaveni okraju grafu 4, 5, 1, 1
13  hist(wei, breaks = b, prob = T, density = 30, col = 'blue', xlab = '',
14       ylab = '', main = '', las = 1, axes = F) # histogram
15  box(bty = 'o') # ramecek okolo grafu
16  axis(1, centr) # osa x
17  axis(2, las = 1) # osa y
18  mtext('porodni hmotnost (v g)', side = 1, line = 2) # popisek osy x
19  mtext('f(x)', side = 2, line = 4) # popisek osy y
20  lines(density(wei), col = 'blue',
21        lwd = 2) # krivka jard. odhadu hustoty (modra silna cara)
22
23  xfit <- seq(min(wei) - 100, max(wei) + 100, length = 512) # posl. hodnot x
24  yfit <- dnorm(xfit, mu, sigma) # hustota norm. rozdeleni N(mu, sigma^2)
25  lines(xfit, yfit, col = 'red', lwd = 2,
26        lty = 2) # krivka hustoty (cervena silna prerusovana cara)
```



Interpretace výsledků: Náhodná veličina X popisující porodní hmotnost novorozenců pochází z normálního rozdělení se střední hodnotou $\mu =$..... a rozptylem $\sigma^2 =$ ....., tj. $X \sim N($.....; .....$)$.

Příklad 5.2. Výpočet pravděpodobností na základě normálního rozdělení
Za předpokladu, že porodní hmotnost novorozenců pochází z normálního rozdělení $N(3078.027, 696^2)$, vypočítejte pravděpodobnost, že porodní hmotnost novorozence bude (a) menší než 3800 g; (b) v rozmezí 2500–4200 g; (c) větší než 4000 g, (d) rovná 2100 g.

Řešení příkladu 5.2
(a) pravděpodobnost, že porodní hmotnost novorozence bude menší než 3800 g

```
27  mu <- ... # stredni hdonota mu
28  sigma <- 696 # sm. odchylka sigma (!NE ROZPTYL sigma^2!)
29  pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.8502061
```
30

(b) pravděpodobnost, že porodní hmotnost novorozence bude v rozmezí 2500–4200

```
31  pnorm(...) - pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.7433938
```
32

(c) pravděpodobnost, že porodní hmotnost novorozence bude větší než 4000 g

```
1 - pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.09263967
```

(d) pravděpodobnost, že porodní hmotnost novorozence bude rovná 2100 g

Interpretace výsledků: Pravděpodobnost, že porodní hmotnost novorozenců bude menší než 3800 g, je .....%. Pravděpodobnost, že porodní hmotnost novorozenců bude v rozmezí 2500–4200 g, je .....%. Pravděpodobnost, že porodní hmotnost novorozenců bude větší než 4000 g, je .....%. Pravděpodobnost, že porodní hmotnost novorozenců bude rovná 2100 g, je .....%, protože data pochází z normálního rozdělení, což je ..... typ rozdělení, proto $\Pr(X = 2100) =$ ..... .

Příklad 5.3. Výpočet pravděpodobností na základě normálního rozdělení
Za předpokladu, že porodní hmotnost novorozenců pochází z normálního rozdělení $N(3078.027, 696^2)$, vypočítejte pravděpodobnost, že **průměrná** porodní hmotnost **pěti** novorozenců bude (a) menší než 3800 g; (b) v rozmezí 2500–4200 g; (c) větší než 4000 g; (d) rovná 2100 g.

Řešení příkladu 5.3
(a) pravděpodobnost, že **průměrná** porodní hmotnost **pěti** novorozenců bude menší než 3800 g

```
35  n <- 5
36  mu <- ... # stredni hodnota mu
37  sigma <- ... # sm. odchylka sigma
38  sigma5 <- sqrt(sigma ^ 2 / n) # sm. odchylka pro prumer (sqrt(sigma^2/n))
39  pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.9898164                                                              40
```

(b) pravděpodobnost, že **průměrná** porodní hmotnost **pěti** novorozenců bude v rozmezí 2500–4200

```
41  pnorm(...) - pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.9681918                                                              42
```

(c) pravděpodobnost, že **průměrná** porodní hmotnost **pěti** novorozenců bude větší než 4000 g

```
1 - pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.001527937
```

(d) pravděpodobnost, že **průměrná** porodní hmotnost **pěti** novorozenců bude rovná 2100 g

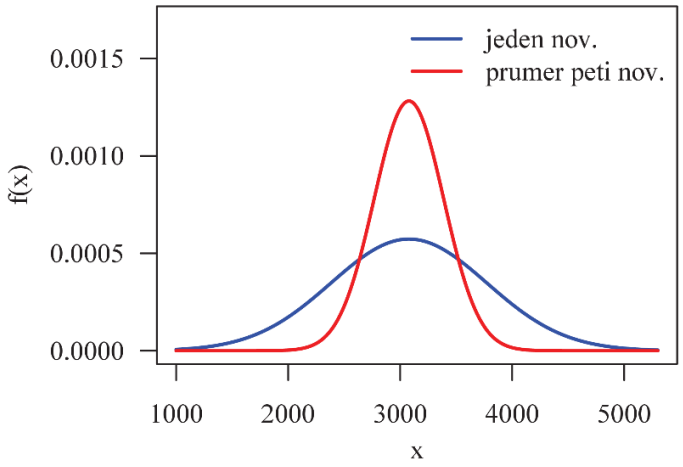
Interpretace výsledků: Pravděpodobnost, že průměrná porodní hmotnost pěti novorozenců bude menší než 3800 g je .....%. Pravděpodobnost, že průměrná porodní hmotnost pěti novorozenců bude v rozmezí 2500–4200 g je .....%. Pravděpodobnost, že průměrná porodní hmotnost pěti novorozenců bude větší než 4000 g je .....%. Pravděpodobnost, že průměrná porodní hmotnost pěti novorozenců bude rovná 2100 g je .....%, protože data pochází z normálního rozdělení, což je ..... typ rozdělení, proto $\Pr(X = 2100) =$ ..... .

Příklad 5.4. Graf hustoty a distribuční funkce normálního rozdělení
Nakreslete graf hustoty a distribuční funkce náhodné veličiny $X \sim N(3078.027, 696^2)$ popisující porodní hmotnost jednoho novorozence a porovnejte je s křivkami hustoty a distribuční funkce náhodné veličiny $\bar{X} \sim N(3078.027, \frac{696^2}{5})$ popisující průměrnou porodní hmotnost pěti novorozenců.

Řešení příkladu 5.4

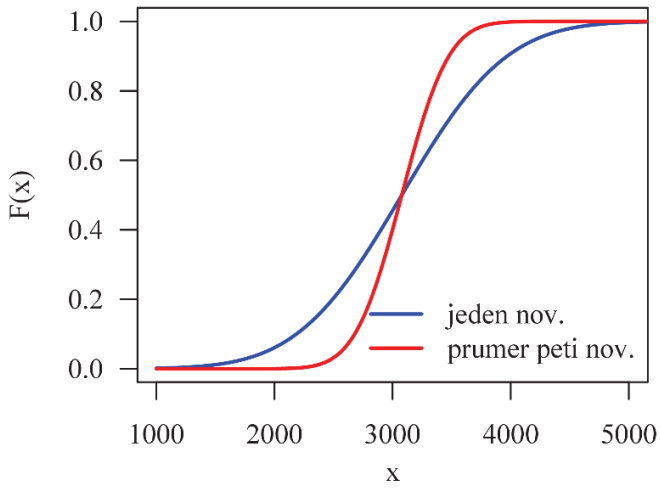
Graf hustoty $f(x)$

```
45  x  <- seq(...) # posl. od 1000 do 5300 o delce 512
46  fx <- dnorm(x, mean = mu, sd = sigma) # hustota rozdeleni N(mu, sigma^2)
47  fx5 <- dnorm(x, mean = mu, sd = sigma5) # hustota rozdeleni N(mu, sigma^2/n)
48
49  par(...) # nastaveni okraju grafu 3, 5, 1, 1
50  plot(x, fx, type = 'l', lwd = 2, xlim = c(1000, 5300), ylim = c(0, 0.0017),
51       col = ..., xlab = '', ylab = '',
52       las = ...) # modra silna cara hustoty N(mu, sigma ^ 2)
53  mtext(...) # popisek osy x
54  mtext(...) # popisek osy y
55  lines(x, fx5, ...) # cervena silna cara hustoty N(mu, sigma^2/n)
56  legend('topright', col = c(..., ...), lwd = c(2, 2),
57         legend = c(..., ...), bty = 'n') # legenda
```



Graf distribuční funkce funkce $F(x)$

```
58  Fx <- pnorm(x, mean = mu, sd = sigma) # distr. fce N(mu, sigma^2)
59  Fx5 <- pnorm(x, mean = mu, sd = sigma5) # distr. fce N(mu, sigma^2/n)
60
61  par(...) # nastaveni okraju 3, 5, 1, 1
62  plot(x, Fx, xlim = c(1000, 5000), xlab = '',
63       ...) # modra silna cara distr. fce N(mu, sigma ^ 2)
64  lines(x, Fx5, ...) # cervena silna cara distr. fce N(mu, sigma^2/n)
65  mtext(...) # popisek osy x
66  legend('bottomright', col = c(...), lwd = c(...),
67         legend = c(...), bty = 'n') # legenda
```



Dvourozměrné normální rozdělení

- $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ ... dvojice nezávislých stejně rozdělených náhodných veličin
- $(X, Y)^T$ ... dvourozměrný náhodný vektor
  - $(X, Y)^T \sim \mathbf{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
    - $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ ... vektor středních hodnot
    - $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ ... varianční matice
  - $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$, kde $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1^2, \sigma_2^2 > 0$, $\rho \in \langle -1; 1 \rangle$
  - hustota

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

  - vlastnosti $\mathrm{E}[(X, Y)^T] = \boldsymbol{\mu}$; $\mathrm{Var}[(X, Y)] = \boldsymbol{\Sigma}$
  - marginální rozdělení $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$.
- Grafická vizualizace dat
  - dvourozměrný tečkový diagram superponovaný konturovým diagramem
  - 3D-graf

Dataset: 03-paired-means-clavicle2.txt
Datový soubor 03-paired-means-clavicle2.txt obsahuje osteometrické údaje o délkách klíčních kostí na pravé a levé straně těla v párovém uspořádání. Data pochází z anglického souboru dokumentovaných skeletů (Parsons, 1916).

Popis proměnných v datasetu:

- id . . . ID jedince;
- sex . . . pohlaví jedince (m - muž, f - žena);
- length.L . . . délka levé klíční kosti (v mm);
- length.R . . . délka pravé klíční kosti (v mm).

Příklad 5.5. Výpočet parametrů μ a Σ dvourozměrného normálního rozdělení
Načtěte datový soubor 03-paired-means-clavicle2.txt. Nechť náhodná veličina X popisuje délku levé klíční kosti a náhodná veličina Y popisuje délku pravé klíční kosti u mužů. Pomocí tečkového diagramu vizualizujte vztah délky levé a pravé klíční kosti. Za předpokladu, že data pochází z dvourozměrného normálního rozdělení $(X, Y)^T \sim \mathbf{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ odhadněte hodnoty parametrů μ_1, μ_2, σ_1^2, σ_2^2 a ρ a stanovte tvar vektoru středních hodnot a varianční matice.

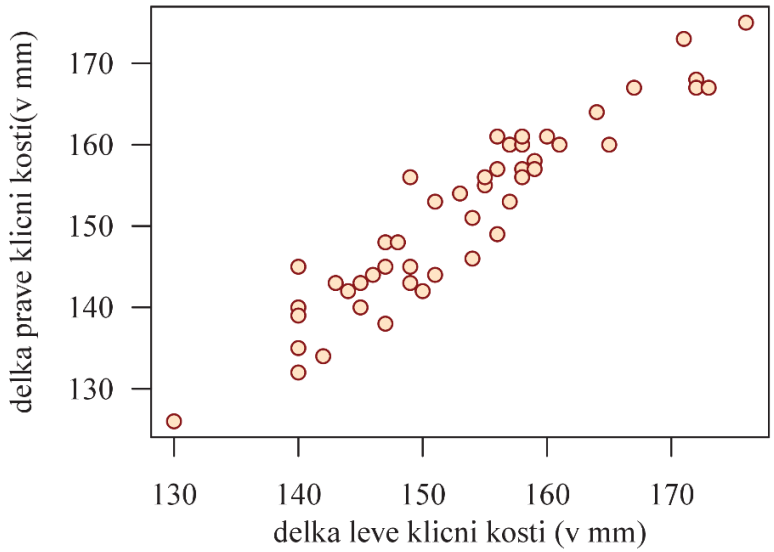
Řešení příkladu 5.5

```
68 data <- read.delim(...) # nacteni datoveho souboru
69 head(...) # vypis prvnich ctyr radku
```

```
  id sex length.R length.L
1 66   m      126      130
2 69   m      158      159
3 71   m      153      151
4 72   m      145      147
```

```r
data.M <- data[...] # vyber radku tykajicich se pouze muzu
data.M <- na.omit(...) # odstraneni NA hodnot
length.ML <- data.M$length.L # vyber delek levych kl. kosti muzu
length.MR <- data.M$length.R # vyber delek pravych kl. kosti muzu

par(...) # nastaveni okraju grafu 4, 4, 1, 1
plot(length.ML, length.MR, pch = ..., bg = ..., col = ...,
     xlab = ..., ylab = ..., las = ...) # teckovy graf
mtext(...) # popisek osy x
```



```r
mu1 <- mean(...) # odhad mu1
mu2 <- mean(...) # odhad mu2
sigma1na2 <- var(...) # odhad sigma1^2
sigma2na2 <- var(...) # odhad sigma2^2
sigma1 <- sd(...) # odhad sigma1
sigma2 <- sd(...) # odhad sigma2
rho <- cor(..., ...) # odhad rho
sigma12 <- cov(..., ...) # odhad sigma12
tab <- data.frame(...) # sumarizacni tabulka vysledku
round(tab, 4)
```

```
                  mu sigma.na.2   sigma     rho  sigma12
leva strana  153.60    98.9388  9.9468 0.9371 102.5061
prava strana 151.74   120.9310 10.9969 0.9371 102.5061
```

Interpretace výsledků: Náhodný vektor $(X, Y)^T$ popisující délku klíční kosti z levé a pravé strany u mužů pochází z dvourozměrného normálního rozdělení s vektorem středních hodnot $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$, kde $\mu_1 =$ ..... mm a $\mu_2 =$ ..... mm a s varianční maticí $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, kde $\sigma_1 =$ ..... mm, $\sigma_2 =$..... mm a $\rho =$ ..... . Délka klíční kosti z levé strany u mužů pochází marginálně z normálního rozdělení se střední hodnotou $\mu_1 =$..... mm a směrodatnou odchylkou $\sigma_1 =$ ..... mm. Délka klíční kosti z pravé strany u mužů pochází marginálně z normálního rozdělení se střední hodnotou $\mu_2 =$ ..... mm a směrodatnou odchylkou $\sigma_2 =$ ..... mm.

Příklad 5.6. Parametrické a neparametrické odhady dat z $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Načtěte datový soubor 03-paired-means-clavicle2.txt. Za předpokladu, že náhodný vektor $(X, Y)^T$ popisující délku klíční kosti z levé a pravé strany u mužů pochází z dvourozměrného normálního rozdělení, tj. $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ s odhadem středních hodnot $\widehat{\mu}_1 = 153.6$, $\widehat{\mu}_2 = 151.74$, rozptylů $\widehat{\sigma}_1^2 = 9.95^2$ a $\widehat{\sigma}_2^2 = 11^2$ a odhadem korelačního koeficientu $\widehat{\rho} = 0.9371$.

(a) sestrojte tečkový diagram délky klíční kosti z levé a pravé strany superponovaný teoretickými konturami teoretického dvourozměrného normálního rozdělení;

(b) sestrojte tečkový diagram délky klíční kosti z levé a pravé strany superponovaný konturami jádrového odhadu hustoty;

(c) sestrojte 3D-diagram hustoty teoretického dvourozměrného normálního rozdělení délky klíční kosti z levé a pravé strany;

(d) sestrojte 3D-diagram jádrového odhadu hustoty délky klíční kosti z levé a pravé strany.

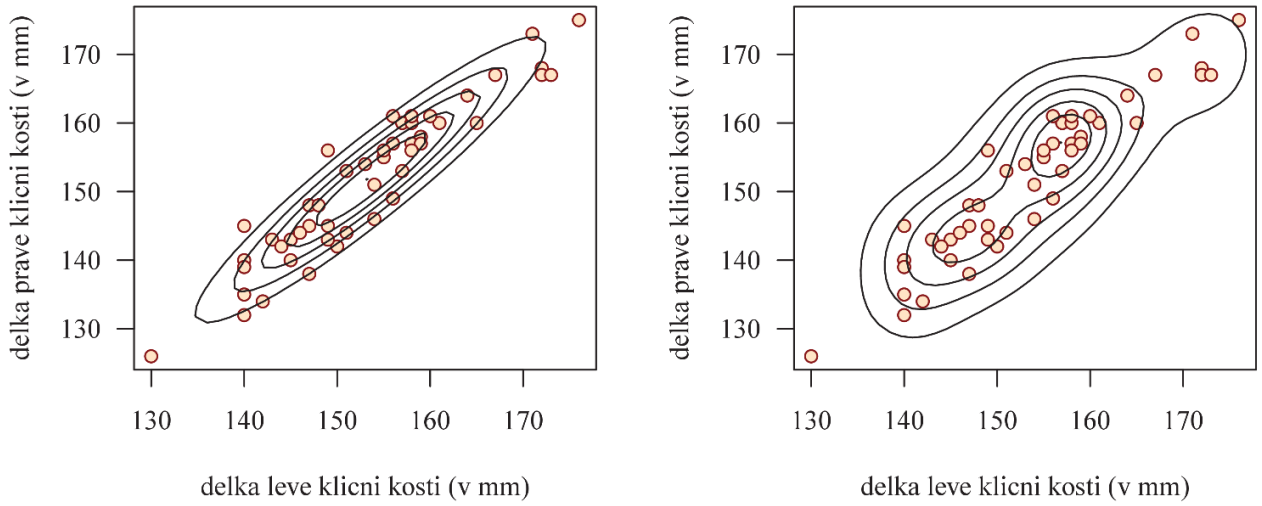
Řešení příkladu 5.6

```
source(...) # nacteni souboru s funkcemi ('Sbirka-AS-I-2018-funkce-2.txt')
Mu <- c(...) # vektor prumeru mL a mR
Sig <- matrix(c(sigma1 ^ 2, rho * sigma1 * sigma2, rho * sigma1 * sigma2,
                sigma2 ^ 2), ncol = 2, nrow = 2) # variancni matice
n <- 50
x <- seq(...) # posloupnost od 125 do 185 o delce n
y <- seq(...) # posloupnost od 120 do 185 o delce n
M <- d2norm(x, y, mean = Mu, sigma = Sig) # teor. hustota N_2(Mu, Sigma)

Z <- MASS::kde2d(length.ML, length.MR, n = n,
                 lim = c(125, 185, 120, 185)) # jadrovy odhad hustoty
```

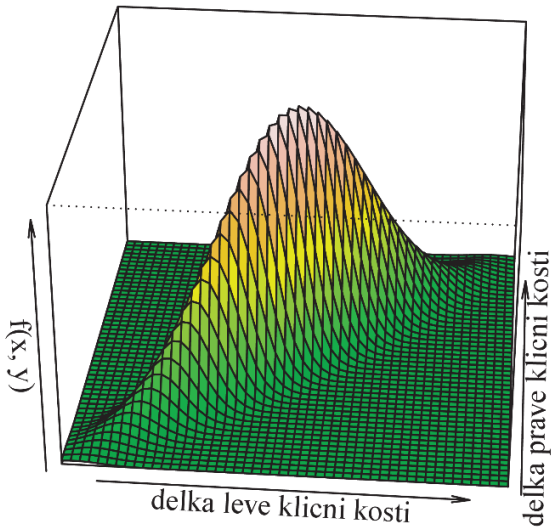
Tečkový diagram s konturami

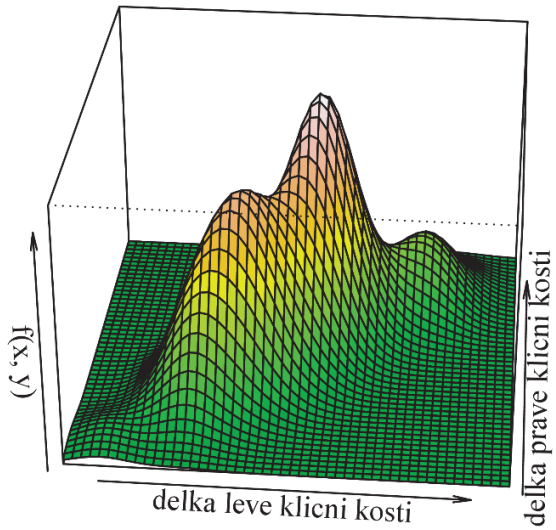
```
108  par(...) # nastaveni okraju grafu 4, 5, 1, 1
109  plot(length.ML, length.MR, ...) # teckovy graf
110  contour(x, y, M, add = T, levels = seq(0, max(M), length = 7),
111          asp = T, drawlabels = F) # konturovy graf (teor. hustota)
112
113  plot(length.ML, length.MR, ...) # teckovy graf
114  contour(Z$x, Z$y, Z$z, add = ..., levels = seq(0, max(Z$z), length = 7),
115          asp = ..., drawlabels = ...) # konturovy graf (jadr. odhad hustoty)
```



3D-graf

```
116  par(...) # nastaveni okraju grafu 3, 4, 1, 4
117  GA::persp3D(x, y, M, phi = 30, theta = 5, ticktype = 'simple',
118              xlab = ..., ylab = ..., zlab = 'f(x, y)', border = 'black',
119              col.palette = terrain.colors) # 3D graf (teor. hustota)
120
121  GA::persp3D(Z$x, Z$y, Z$z, phi = ..., theta = ..., ticktype = ...,
122              xlab = ..., ylab = ..., zlab = ..., border = ...,
123              col.palette = ...) # 3D graf (jadr. odhad hustoty)
```





Interpetace výsledků: Na základě grafické vizualizace předpokládáme, že data pochází / nepochází z dvourozměrného normálního rozdělení.

Poznámka: Hodnocení normality na základě grafické vizualizace je pouze subjektivním hodnocením. V sedmém cvičení budeme normalitu, případně dvourozměrnou normalitu, náhodného výběru posuzovat objektivně, a to na základě testů normality.