Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference I

Téma 7: Normální model

Veronika Bendová

bendova.veroonika@gmail.com

Normální rozdělení $N(\mu, \sigma^2)$

- $X_1, \ldots, X_n$ ... nezávislé náhodné veličiny
- Normální rozdělení
  - $X \sim N(\mu, \sigma^2)$
  - $\boldsymbol{\theta} = (\mu, \sigma^2)^T$
  - hustota

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}. \tag{7.1}$$

  - vlastnosti $\mathrm{E}[X] = \mu$; $\mathrm{Var}[X] = \sigma^2$
  - dnorm(x, mu, sigma), pnorm(x, mu, sigma), rnorm(M, mu, sigma), qnorm(alpha, mu, sigma)

- Standardizované normální rozdělení
  - $X \sim N(0, 1)$
  - $\boldsymbol{\theta} = (0, 1)^T$
  - hustota

$$f(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad x \in \mathbb{R}. \tag{7.2}$$

  - vlastnosti $\mathrm{E}[X] = 0$; $\mathrm{Var}[X] = 1$
  - dnorm(x), pnorm(x), rnorm(M), 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 $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
  - Věta 2: Nechť $X_1, \ldots, X_{n_1}$ jsou nezávislé náhodné veličiny pocházející z normálního rozdělení $N(\mu_1, \sigma_1^2)$ a $Y_1, \ldots, Y_{n_2}$ jsou nezávislé náhodné veličiny pocházející z normálního rozdělení $N(\mu_2, \sigma_2^2)$. Potom rozdíl náhodných veličin $\bar{X}_{n_1} - \bar{Y}_{n_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$.

- **Dataset 6: 03-paired-means-clavicle2.txt**
  - Datový soubor obsahuje osteometrické údaje o délkách klíčních kostí (*clavicula*). Data pochází z anglického souboru dokumentovaných skeletů (Parsons, 1916). V souboru se nachází délky klíčních kostí na pravé a levé straně těla v párovém uspořádání. Jednotlivé kosti bez druhostranné kosti nebyly do souboru zařazeny.
  - Přehled proměnných v datasetu:
    - id . . . ID jedince;
    - sex . . . pohlaví jedince (m - muž, f - žena);
    - length.L . . . délka klíční kosti z levé strany (v mm);
    - length.R . . . délka klíční kosti z pravé strany (v mm).

Příklad 7.1. Hustota normálního modelu

Naprogramujte v R funkci dnormal(x, mu, sigma2) počítající hodnoty hustoty normálního rozdělení $N(\mu, \sigma^2)$ v hodnotě x. Správnost funkce otestujte na výpočtu $f(x)$, $x = -1, 0, 2$, pro $X \sim N(\mu, \sigma^2)$, kde $\mu = 0$ a $\sigma^2 = 2$. Výsledky ověřte s výsledky funkce dnorm().

Řešení příkladu 7.1.

```
1  dnormal <- function(...){ # fce s povinnymi vstupnimi argumenty x, mu, sigma2
2    fx <- ... # hustota rozdeleni N(mu, sigma ^ 2); viz vzorec 7.1
3    return(...)
4  }
5
6  dnormal(...) # hustota rozdeleni N(0, 2); funkce dnormal()
7  dnorm(...) # hustota rozdeleni N(0, 2); funkce dnorm()
```

```
      f(-1)      f(0)      f(2)                                                  8
1 0.2196956 0.2820948 0.1037769                                                  9
```

f(-1) = 0.2197; f(0) = 0.2821; f(2) = 0.1038.

Příklad 7.2. Základní číselné charakteristiky spojitého znaku

Načtěte datový soubor 03-paired-means-clavicle2.txt. Nechť náhodná proměnná X popisuje délku klíční kosti z levé strany u mužů. Proměnná X je potom spojitého typu. Pro délku klíční kosti z levé strany u mužů vytvořte tabulku základních číselných charakteristik.

Řešení příkladu 7.2

V tabulce základních číselných charakteristik budou obsaženy následující charakteristiky: aritmetický průměr, směrodatná odchylka, koeficient variace, minimální hodnota, dolní kvartil, medián, horní kvartil, maximální hodnota, interkvartilové rozpětí, koeficient šikmosti a koeficient špičatosti.

- výběrový průměr (aritmetický průměr)

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

- výběrový rozptyl

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2,$$

- výběrová směrodatná odchylka

$$s = \sqrt{s^2}$$

- výběrový koeficient variace

$$v = \frac{s}{m}$$

- α-kvantily (dolní kvartil, medián, horní vartil)

$$n\alpha = c \ldots \begin{cases} \text{je celé číslo} \rightarrow x_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2} \\ \text{není celé číslo} \rightarrow \lceil c \rceil \rightarrow x_\alpha = x_{(\lceil c \rceil)} \end{cases}$$

- interkvartilové rozpětí (mezikvartilové rozpětí)

$$IQR = x_{0.75} - x_{0.25}$$

- koeficient šikmosti

$$b_1 = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - m)^3}{s^3},$$

- koeficient špičatosti

$$b_2 = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - m)^4}{s^4} - 3,$$

```r
data <- read.delim(..., sep = ..., dec = ...) # nacteni datoveho souboru
data.M <- na.omit(data[data$sex == 'm', ]) # udaje pro muze + odstraneni NA
l.LM <- data.M$length.L # vyber delek klicnich kosti z leve strany
n <- length(...) # rozsah nahodneho vyberu
m <- mean(...) # vyberovy prumer
s <- sd(...) # vyberova sm. odchylka
v <- ... # vyberovy koeficient variace
min <- min(...) # minimalni namerena hodnota
max <- max(...) # maximalni namerena hodnota
x0.25 <- quantile(l.LM, probs = 0.25, type = 2) # dolni kvartil
x0.50 <- quantile(...) # median
x0.75 <- quantile(...) # horni kvartil
IQR <- ... # interkvartilove rozpeti
sikmost <- e1071::skewness(l.LM, type = 3) # koeficient sikmosti b1
spicatost <- e1071::kurtosis(...) # koeficient spicatosti b2
Xtab <- data.frame(...) # souhrnna tabulka vysledku
```

```
      n     m    s    v min dolni.kv median horni.kv max IQR sikmost spicatost
leva 50 153.6 9.95 0.06 130      147  154.5      158 176  11    0.21     -0.29
```

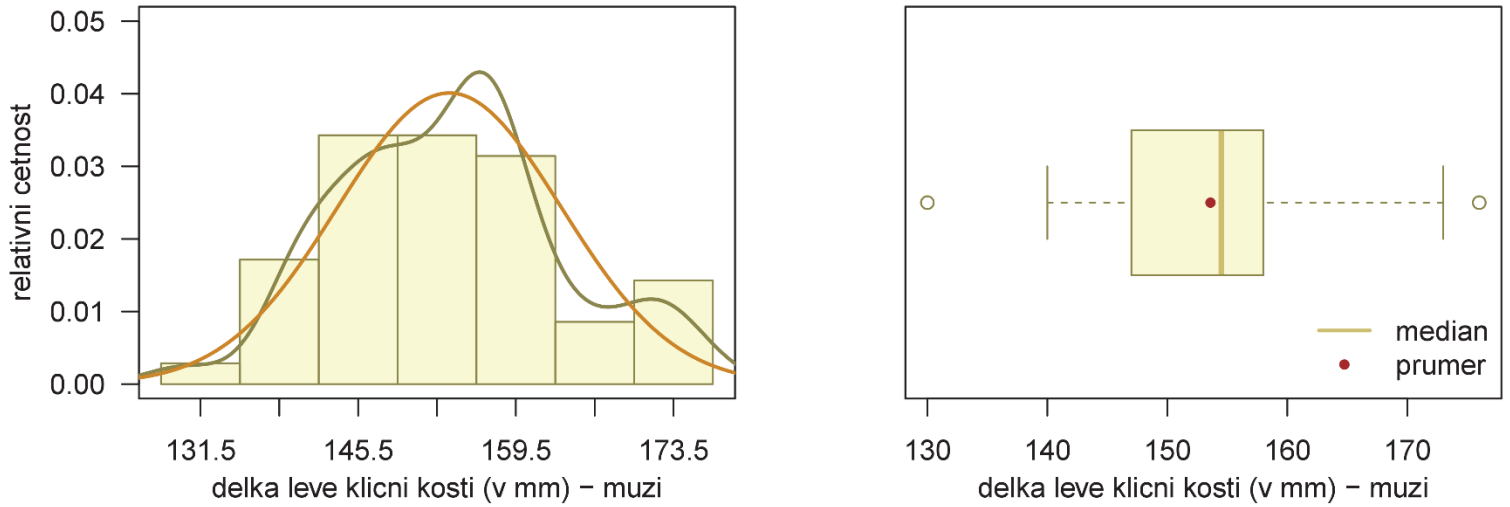
Datový soubor obsahuje údaje o délkách klíční kosti z levé strany 50 mužů. Naměřené hodnoty délky klíční kosti z levé strany u mužů se pohybují v rozmezí 130–176 mm. Průměrná délka klíční kosti z levé strany u mužů je 153.60 mm se směrodatnou odchylkou 9.95 mm. Směrodatná odchylka představuje 6 % aritmetického průměru. 25 % naměřených hodnot je menších nebo rovných 147.00 mm, 50 % hodnot je menších nebo rovných 154.50 mm, 75 % hodnot je menších nebo rovných 158.00 mm. Interkvartilové rozpětí naměřených hodnot je 11 mm. Hodnota koeficientu šikmosti ($b_1 = 0.21$) ukazuje na mírné kladné vychýlení dat (hodnoty vyšikmené doleva s prodlouženým pravým koncem), hodnota koeficientu špičatosti ($b_2 = -0.29$) ukazuje na mírně zploštělý charakter dat.

Příklad 7.3. Vizualizace dat z normálního modelu
Načtěte datový soubor 03-paired-means-clavicle2.txt. Nechť náhodná proměnná X popisuje délku klíční kosti z levé strany u mužů. Pomocí histogramu a krabicového diagramu vhodně vizualizujte rozdělení délky klíční kosti z levé strany u mužů. Histogram superponujte (a) křivkou jádrového odhadu hustoty; (b) křivkou teoretické hustoty normálního rozdělení $N(\mu, \sigma^2)$. Hodnoty parametrů μ a σ^2 odhadněte na základě dat.

Řešení příkladu 7.3

```
28  xfit <- seq(...) # posl. od min - 10 do max + 10 o delce 512
29  yfit <- dnorm(...) # hustota rozdeleni N(m, s^2) nad posl. xfit
30  r <- round(3.3 * log10(n) + 1) # Sturgesovo pravidlo; pocet tr. intervalu = 7
31  # 176 - 130 = 46 -> 49 / 7 = 7 ... opt. sirka = 7 -> seq(128, 177, by = 7)
32
33  par(...) # okraje grafu 4, 4, 1, 1
34  hist(l.LM, prob = T, breaks = seq(128, 177, by = 7), axes = F,
35       ylim = c(0, 0.05), col = ..., border = ..., density = ..., main = '',
36       xlab = '', ylab = ...) # histogram delek levych kl. kosti
37  box(...) # ramecek okolo grafu
38  axis(..., seq(131.5, 173.5, by = 7)) # osa x; stredy tridicich intervalu
39  axis(...) # osa y
40  mtext(...) # popisek osy x
41  lines(density(...), col = ..., lwd = ...) # krivka jadr. odhadu hustoty
42  lines(xfit, yfit, col = 'orange3', lwd = 2) # krivka hustoty N(m, s^2)
```

```
43  boxplot(l.LM, type = 2, xlab = '', las = 1, horizontal = T,
44          col = ..., border = ..., medcol = ...) # krabicovy diagram
45  mtext(...) # popisek osy x
46  points(m, 1, pch = ..., col = ...) # aritm. prumer jako bod
47  legend('bottomright', lty = c(1, NA), pch = c(NA, 20),  lwd = c(2, NA),
48         col = ..., legend = ..., bty = ...) # legenda
```



Obrázek: Vizualizace délky levé klíční kosti u mužů pomocí histogramu (vlevo) a krabicového diagramu (vpravo)

Příklad 7.4. Výpočet pravděpodobností na základě normálního modelu
Za předpokladu, že náhodná veličina X udávající délku klíční kosti z levé strany u mužů pochází z normálního rozdělení $N(153.6, 9.95^2)$ vypočítejte pravděpodobnost, že délka klíční kosti z levé strany je (a) menší než 140 mm; (b) větší než 160 mm; (c) v rozmezí 150–160 mm; (d) rovná 155 mm.

Řešení příkladu 7.4 (a)

```
49  p1 <- pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.08576968                                                    50
```

(b)

```
51  p2 <- 1 - pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.2599748                                                     52
```

(c)

```
53 p3 <- pnorm(...) - pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.3813214
```
54

(d)

```
55 p4 <- ... # pravdepodobnost (d)
```

```
[1] 0
```
56

Pravděpodobnost, že délka klíční kosti z levé strany u mužů je menší než 140 mm je 8.58 %. Pravděpodobnost, že délka klíční kosti z levé strany u mužů je větší než 160 mm je 26.00 %. Pravděpodobnost, že délka klíční kosti z levé strany u mužů je v rozmezí 150–160 mm je 38.13 %. Protože délka klíční kosti z levé strany u mužů pochází z normálního rozdělení, což je rozdělení spojitého typu, je tato délka rovná 155 mm s pravděpodobností 0 %.

Příklad 7.5. Výpočet pravděpodobností na základě normálního modelu
Za předpokladu, že náhodná veličina X udávající délku klíční kosti z levé strany u mužů pochází z normálního rozdělení $N(153.6, 9.95^2)$ vypočítejte pravděpodobnost, že **průměrná délka pěti** klíčních kostí z levé strany u mužů je (a) menší než 140 mm; (b) větší než 160 mm; (c) v rozmezí 150–160 mm; (d) rovná 155 mm.

Řešení příkladu 7.5
(a)

```
57  s5 <- ... # rozptyl N(mu, sigma ^ 2 / n)
58  p1 <- pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.001116635
```
59

(b)

```
60  p2 <- 1 - pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.07511239
```
61

(c)

```
62 p3 <- pnorm(...) - pnorm(...) # vypocet pravdepodobnosti
```

```
[1] 0.7157135                                                    63
```

(d)

```
64 p4 <- ... # pravdepodobnost (d)
```

```
[1] 0                                                            65
```

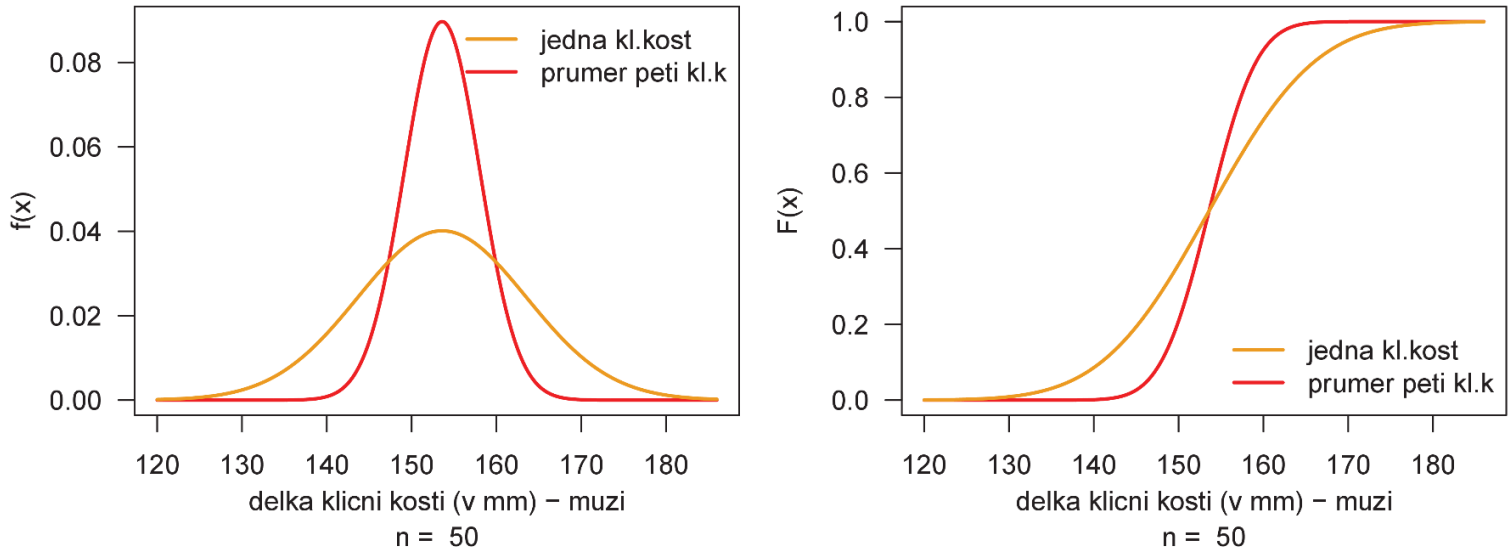
Pravděpodobnost, že průměrná délka pěti klíčních kostí z levé strany u mužů je menší než 140 mm je 0.11 %. Pravděpodobnost, že průměrná délka pěti klíčních kostí z levé strany u mužů je větší než 160 mm je 7.51 %. Pravděpodobnost, že průměrná délka pěti klíčních kostí z levé strany u mužů je v rozmezí 150–160 mm je 71.57 %. Protože průměrná délka pěti klíčních kostí z levé strany u mužů pochází z normálního rozdělení, což je rozdělení spojitého typu, je tata délka rovná 155 mm s pravděpodobností 0.00 %.

Příklad 7.6. Graf funkce hustoty a distribuční funkce normálního modelu
V příkladech 7.2 a 7.3 jsme odhadli hodnoty parametrů μ a σ^2 normálního rozdělení znaku $X =$ *délka klíční kosti z levé strany u mužů* jako $\widehat{\mu}_X = 153.60$ a $\widehat{\sigma}^2_X = 9.95^2$. Nakreslete graf hustoty a distribuční funkce (a) rozdělení $N\left(\mu, \sigma^2\right)$; (b) rozdělení $N\left(\mu, \frac{\sigma^2}{n}\right)$.

Řešení příkladu 7.6

```
66  xfit <- seq(...) # posl. od min - 10 do max + 10 o delce 512
67  fx <- dnorm(...) # hustota rozdeleni N(mu, sigma ^ 2) nad posl. xfit
68  fx5 <- ... # hustota rozdeleni N(mu, sigma ^ 2 / n) nad posl. xfit
69  Fx <- pnorm(...) # distr. fce rozdeleni N(mu, sigma ^ 2) nad posl. xfit
70  Fx5 <- ... # distr. fce rozdeleni N(mu, sigma ^ 2 / n) nad posl. xfit
71
72  par(...) # okraje grafu 5, 4, 1, 1
73  plot(xfit, fx5, type = 'l', xlab = '', ylab = ..., col = ...,
74       lwd = ..., las = ...) # krivka hustoty N(mu, sigma ^ 2 / n); cervena, silna
75  lines(xfit, fx, ...) # krivka hustoty N(mu, sigma ^ 2); oranzova, silna
76  mtext(...) # popisek osy x
77  mtext(paste('n = ', n), ...) # druhy popisek osy x; n = ...
78  legend(...) # legenda
```

```
79  plot(xfit, Fx5, ...) # krivka distr. fce N(mu, sigma ^ 2 / n); cervena, silna
80  lines(...) # krivka distr. fce N(mu, sigma ^ 2); oranzova, silna
81  mtext(...) # popisek osy x
82  mtext(...) # druhy popisek osy x
83  legend(...) # legenda
```



Obrázek: Funkce hustoty (vlevo) a distribuční funkce (vpravo) normálního modelu

Příklad 7.7. Simulační studie: Věta 1: Rozdělení výběrového průměru $\bar{X}_n$

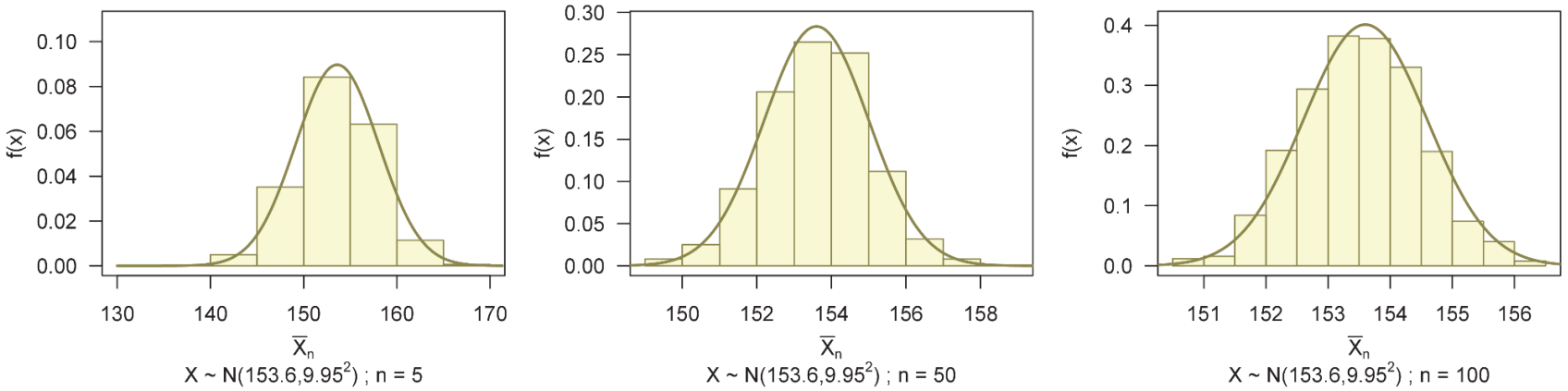
Na základě simulační studie ověřte, že pokud $X \sim N(\mu, \sigma^2)$, potom $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$. Zvolte $\mu = 153.60$, $\sigma^2 = 9.95^2$, (a) $n = 5$, (b) $n = 50$, (c) $n = 100$. Vygenerujte M pseudonáhodných výběrů $X_1, \ldots, X_n$, $M = 1000$. Pro každý výběr vypočítejte realizaci aritmetického průměru $\bar{x}_{n,m}$, $m = 1, \ldots, M$. Následně vygenerujte histogram pro hodnoty $\bar{x}_{n,m}$ a superponujte jej teoretickou křivkou hustoty pro $\bar{X}_n$. Pro všechny tři případy (a), (b) i (c) vypočítejte $\Pr(\bar{X}_n > 152)$ na základě empirického a teoretického rozdělení $\bar{X}_n$. Pravděpodobnosti porovnejte.

Řešení příkladu 7.7

```
84  simulace.mean <- function(n, mu = ..., sigma = ..., M = ..., vypis = ...){
85    m <- replicate(M, mean(rnorm(n, mean = ..., sd = ...))) # M = 1000 vyb. prum.
86    xfit <- seq(...) # posl. od min(m) - 5 do max(m) + 5 o délce 512
87    yfit <- dnorm(...) # hustota rozdeleni N(mu, sigma ^ 2 / n) nad xfit
88    par(...) # okraje grafu 5, 4, 1, 1
89    hist(m, prob = ..., ylim = c(0, max(hist(m, plot = F)$density) + 0.025),
90         ...) # histogram 1000 vyberovych prumeru
91    box(...) # ramecek okolo grafu
92    lines(...) # krivka hustoty N(mu, sigma ^ 2 / n)
93    mtext(expression(paste(bar(X)[n])), ...) # popisek osy x
94    mtext(bquote(paste('X ~ N(', .(mu), ',', .(round(sigma, 2))^2, ') ; n = ',
95                       .(n))), ...) # druhy popisek osy x
96    tab <- data.frame(teor = 1 - pnorm(152, mu, sqrt(sigma ^ 2 / n)),
97                      exact = sum(m > 152) / M) # tabulka teor. a exaktni psti
98    if (vypis == T) return(tab) # pokud vypis == T, vrat jako vystup promennou tab
99  }
```

```r
sim1 <- simulace.mean(n = 5, mu = m, sigma = s) # simulace pro (a) n = 5
sim2 <- simulace.mean(...) # simulace pro (b) n = 50
sim3 <- simulace.mean(...) # simulace pro (c) n = 100
tab <- data.frame(t(rbind(...))) # souhrnna tabulka vysledku tri simulaci
names(tab) <- ... # pojmenovani sloupcu tabulky tab
```

```
             n = 5 n = 50 n = 100
teoreticka 0.6405 0.8723  0.9461
exaktni    0.6420 0.8760  0.9440
```



Obrázek: Histogramy průměrů 1000 náhodných výběrů o rozsahu (a) $n = 5$ (vlevo); (b) $n = 50$ (uprostřed); (c) $n = 100$ (vpravo)