7 Základní pojmy matematické statistiky

7.1 Vybrané statistiky pro jeden jednorozměrný náhodný výběr a jejich vlastnosti

$X_1, \ldots, X_n$ je náhodný výběr z rozložení se střední hodnotou μ a rozptylem σ^2, $n \geq 2$.

Definice statistik

- $M = \frac{1}{n} \sum_{i=1}^{n} X_i$ ... výběrový průměr,
- $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M)^2$ ... výběrový rozptyl,
- $S = \sqrt{S^2}$ ... výběrová směrodatná odchylka.

Vlastnosti statistik

- M je nestranným bodovým odhadem μ,
- S^2 je nestranným bodovým odhadem σ^2,
- S je asymptoticky nestranným bodovým odhadem σ.

Příklad 7.1. Řešený příklad
Načtěte datový soubor 17-anova-newborns.txt. Pro porodní hmotnost novorozence (weight.C) ženského pohlaví stanovte (a) hodnotu bodového odhadu μ; (b) hodnotu bodového odhadu σ^2; (c) hodnotu bodového odhadu σ.

Řešení příkladu 7.1
Nestranným bodovým odhadem μ je výběrový průměr M. Jeho hodnotu stanovíme pomocí funkce mean(). Nestranným bodovým odhadem σ^2 je výběrový rozptyl S^2. Jeho hodnotu stanovíme pomocí funkce var(). Asymptoticky nestranným bodovým odhadem σ je výběrová směrodatná odchylka S. Její hodnotu stanovíme pomocí funkce sd().

```
1 data <- read.delim('17-anova-newborns.txt')
2 weight.CF <- na.omit(data[data$sex == 'f', 'weight.C'])
3 m <- mean(weight.CF) # 3012,832
4 s2 <- var(weight.CF) # 455722,4
5 s <- sd(weight.CF) # 675,0722
```

Hodnota výběrového průměru $m = 3012,8\,\text{mm}$, hodnota výběrového rozptylu $s^2 = 455722,4\,\text{mm}^2$, hodnota výběrové směrodatné odchylky $s = 675,1\,\text{mm}$. ★

Příklad 7.2. Neřešený příklad
Načtěte datový soubor 18-more-samples-variances-clavicle.txt. Pro největší délku klíční kosti z pravé strany (cla.L) mužů řecké populace z Atén stanovte (a) hodnotu bodového odhadu μ; (b) hodnotu bodového odhadu σ^2; (c) hodnotu bodového odhadu σ.
Výsledky: (a) $m = 153,5\,\text{mm}$; (b) $s^2 = 83,\text{mm}^2$; (c) $s = 9,1\,\text{mm}$. ★

7.2 Vybrané statistiky pro jeden dvourozměrný náhodný výběr a jejich vlastnosti

$(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ je náhodný výběr z dvourozměrného rozložení se středními hodnotami μ_1, μ_2, rozptyly σ_1^2, σ_2^2, kovariancí σ_{12} a koeficientem korelace ρ.

Definice statistik

- $M_1 = \frac{1}{n} \sum_{i=1}^{n} X_i$, $M_2 = \frac{1}{n} \sum_{i=1}^{n} Y_i$ ... výběrové průměry,
- $S_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_1)^2$, $S_2^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - M_2)^2$ ... výběrové rozptyly,
- $S_{12} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_1)(Y_i - M_2)$ ... výběrová kovariance,

- $R_{12} = \begin{cases} \frac{S_{12}}{S_1 S_2} \text{ pro } S_1 S_2 \neq 0, \\ 0 \text{ jinak,} \end{cases}$ ... výběrový koeficient korelace.

Vlastnosti statistik

- S_{12} je nestranným bodovým odhadem σ_{12},
- R_{12} je vychýleným bodovým odhadem ρ (vychýlení je zanedbatelné pro $n \geq 30$).

Příklad 7.3. Řešený příklad
Načtěte datový soubor 05-one-sample-correlation-skull-mf.txt. Pro největší výšku mozkovny (skull.pH) a morfologickou výšku tváře (face.H) žen stanovte (a) hodnotu bodového odhadu σ_{12}; (b) hodnotu bodového odhadu ρ.

Řešení příkladu 7.3
Nestranným bodovým odhadem σ_{12} je výběrová kovariance S_{12}. Její hodnotu stanovíme pomocí funkce cov(). Vychýleným bodovým odhadem korelačního koeficientu ρ je výběrový koeficient korelace R_{12}. Jeho hodnotu stanovíme pomocí funkce cor().

```
6  data <- read.delim('05-one-sample-correlation-skull-mf.txt')
7  data.F <- na.omit(data[data$sex == 'f', ])
8  skull.pHF <- data.F$skull.pH
9  face.HF <- data.F$face.H
10 cov(skull.pHF, face.HF) # 1,808192
11 cor(skull.pHF, face.HF) # 0,06417166
```

Hodnota výběrové kovariance $s_{12} = 1,808$, hodnota výběrového korelačního koeficientu $r_{12} = 0,064$. Mezi největší výškou mozkovny a morfologickou výškou tváře existuje zanedbatelný stupeň přímé lineární závislosti. ★

Příklad 7.4. Neřešený příklad
Načtěte datový soubor 19-more-samples-correlations-skull.txt. Pro výšku nosu (nose.H) a šířku nosu (nose.B) peruánské populace stanovte (a) hodnotu bodového odhadu σ_{12}; (b) hodnotu bodového odhadu ρ.
Výsledky: $s_{12} = 0,699$; $r_{12} = 0,137$. Mezi výškou a šířkou nosu mužů peruánské populace existuje slabý stupeň přímé lineární závislosti. ★

7.3 Vybrané statistiky pro dva nezávislé náhodné výběry a jejich vlastnosti

$X_{11}, \ldots, X_{1n_1}$ a $X_{21}, \ldots, X_{2n_2}$ jsou dva nezávislé náhodné výběry, první z rozložení se střední hodnotou μ_1 a rozptylem σ^2, druhý z rozložení se střední hodnotou μ_2 a rozptylem σ^2, $n_1 \geq 2$, $n_2 \geq 2$.

Definice statistik

- $M_1 - M_2 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$ ... rozdíl výběrových průměrů,
- $\frac{S_1^2}{S_2^2} = \frac{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - M_1)^2}{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2i} - M_2)^2}$ ... podíl výběrových rozptylů,
- $S_*^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$ ... vážený průměr výběrových rozptylů.

Vlastnosti statistik

- $M_1 - M_2$ je nestranným bodovým odhadem $\mu_1 - \mu_2$,
- S_*^2 je nestranným bodovým odhadem σ^2.

Příklad 7.5. Řešený příklad
Načtěte datový soubor 15-anova-means-skull.txt. Pro výšku horní části tváře (upface.H) malajské a čínské populace

stanovte (a) hodnotu bodového odhadu $\mu_1 - \mu_2$; (b) hodnotu bodového odhadu σ^2.

Řešení příkladu 7.5
Nestranným bodovým odhadem $\mu_1 - \mu_2$ je rozdíl výběrových průměrů $M_1 - M_2$. Hodnoty výběrových průměrů vypočítáme pomocí funkce mean(). Nestranným bodovým odhadem σ^2 je vážený průměr výběrových rozptylů S_*^2. Hodnoty výběrových rozptylů vypočítáme pomocí funkce var(), rozsahy náhodných výběrů stanovíme pomocí funkce length().

```
12 data <- read.delim('15-anova-means-skull.txt')
13 data <- na.omit(data)
14 upface.HM <- data[data$pop == 'mal', 'upface.H']
15 upface.HC <- data[data$pop == 'cin', 'upface.H']
16 m1 <- mean(upface.HM) # 70,13043
17 m2 <- mean(upface.HC) # 72
18 m1 - m2 # -1,869565
19 s1 <- var(upface.HM) # 24,52685
20 s2 <- var(upface.HC) # 20,82353
21 n1 <- length(upface.HM) # 69
22 n2 <- length(upface.HC) # 18
23 sh <- ((n1 - 1) * s1 + (n2 - 1) * s2) / (n1 + n2 - 2) # 23,78619
```

Hodnota rozdílu výběrových průměrů $m_1 - m_2 = -1,9\,\text{mm}$, hodnota váženého průměru výběrových rozptylů $s_*^2 = 23,8\,\text{mm}^2$. ★

Příklad 7.6. Neřešený příklad
Načtěte datový soubor 01-one-sample-mean-skull-mf.txt. Pro největší délku mozkovny mužů a žen (skull.L) stanovte (a) hodnotu bodového odhadu $\mu_1 - \mu_2$; (b) hodnotu bodového odhadu σ^2.
Výsledky: $m_1 - m_2 = 7,5\,\text{mm}$; $s_*^2 = 39,9\,\text{mm}^2$. ★

7.4 Vybrané statistiky pro aspoň tři nezávislé náhodné výběry a jejich vlastnosti

$X_{11}, \ldots, X_{1n_1}$ až $X_{r1}, \ldots, X_{rn_r}$ je $r \geq 3$ nezávislých náhodných výběrů, první z rozložení se střední hodnotou μ_1 a rozptylem σ^2 až r-tý z rozložení se střední hodnotou μ_r a rozptylem σ^2, $n_1 \geq 2, \ldots, n_r \geq 2$. Označme $n = \sum_{j=1}^r n_j$. Nechť $c_1, \ldots, c_r$ jsou reálné konstanty, aspoň jedna nenulová.

Definice statistik

- $\sum_{j=1}^r c_j M_j$ ... lineární kombinace výběrových průměrů,
- $S_*^2 = \frac{\sum_{j=1}^r (n_j - 1) S_j^2}{n-r}$ ... vážený průměr výběrových rozptylů.

Vlastnosti statistik

- $\sum_{j=1}^r c_j M_j$ je nestranným bodovým odhadem $\sum_{j=1}^r c_j \mu_j$,
- S_*^2 je nestranným bodovým odhadem σ^2.

Příklad 7.7. Řešený příklad
Načtěte datový soubor 18-more-samples-variances-clavicle.txt. Pro největší délku klíční kosti z pravé strany (cla.L) řecké populace z Atén, indické populace z Amritsaru a indické populace z Varanasi stanovte (a) hodnotu bodového odhadu $\sum_{j=1}^r c_j \mu_j$, kde $c_1 = -2$, $c_2 = 2$, $c_3 = 3$; (b) hodnotu bodového odhadu σ^2.

Řešení příkladu 7.7
Nestranným bodovým odhadem $\sum_{j=1}^r c_j \mu_j$ je lineární kombinace výběrových průměrů $\sum_{j=1}^r c_j M_j$. Hodnoty výběrových průměrů vypočítáme pomocí funkce mean(). Nestranným bodovým odhadem σ^2 je vážený průměr výběrových rozptylů S_*^2. Hodnoty výběrových rozptylů vypočítáme pomocí funkce var(), rozsahy náhodných výběrů stanovíme pomocí funkce length().

```
data <- read.delim('18-more-samples-variances-clavicle.txt')
data <- na.omit(data)
cla.LG <- data[data$pop == 'gre', 'cla.L']
cla.LA <- data[data$pop == 'ind1', 'cla.L']
cla.LV <- data[data$pop == 'ind2', 'cla.L']
m1 <- mean(cla.LG) # 153,5213
m2 <- mean(cla.LA) # 145,5667
m3 <- mean(cla.LV) # 141,4938
c1 <- -2
c2 <- 2
c3 <- 3
c1 * m1 + c2 * m2 + c3 * m3 # 408,5723
s1 <- var(cla.LG) # 83,15546
s2 <- var(cla.LA) # 76,27283
s3 <- var(cla.LV) # 67,57184
n1 <- length(cla.LG) # 94
n2 <- length(cla.LA) # 120
n3 <- length(cla.LV) # 81
sh <- ((n1 - 1) * s1 + (n2 - 1) * s2 + (n3 - 1) * s3) / (n1 + n2 + n3 - 3) # 76,08107
```

Hodnota lineární kombinace $-2m_1 + 2m_2 + 3m_3 = 408,5723$, hodnota váženého průměru výběrových rozptylů $s_*^2 = 76,0811$. ★

Příklad 7.8. Neřešený příklad
Načtěte datový soubor 19-more-samples-correlations-skull.txt. Pro interorbitální šířku (intorb.B) německé, bantuské, malajské, čínské a peruánské populace stanovte (a) hodnotu bodového odhadu $\sum_{j=1}^{r} c_j \mu_j$, kde $c_1 = 1$, $c_2 = 2$, $c_3 = -1$, $c4 = 0$, $c_5 = -1$; (b) hodnotu bodového odhadu σ^2.
Výsledky: (a) $m_1 + 2m_2 - m_3 + 0m_4 - m_5 = 30,8\,\text{mm}$; (b) $s_*^2 = 5,6\,\text{mm}^2$. ★