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 rozdělení 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 obsahující údaje o porodní hmotnosti v g (weight.C) u novorozenců. Za předpokladu, že náhodný výběr porodních hmotností u novorozenců ženského pohlaví pochází z rozdělení se střední hodnotou μ a rozptylem σ^2, vypočítejte (a) bodový odhad μ; (b) bodový odhad σ^2; (c) bodový odhad σ. Výsledné hodnoty řádně interpretujte.

Řešení příkladu 7.1
Datový soubor načteme a z načtených dat vybereme pouze řádky týkající se novorozenců ženského pohlaví a sloupec weight.C. Z výběru odstraníme řádky s chybějícími hodnotami.

```
1 data <- read.delim("17-anova-newborns.txt", sep = "\t")
2 head(data, n = 4)
```

```
  edu.M prch.N sex.C weight.C                                                    3
1     2      0     m     3470                                                    4
2     2      0     m     3240                                                    5
3     2      0     f     2980                                                    6
4     1      0     m     3280                                                    7
```

```
8 weight.CF <- na.omit(data[data$sex == "f", "weight.C"])
```

Nestranným bodovým odhadem μ je výběrový průměr M. Jeho hodnotu vypočítáme příkazem mean(). Nestranným bodovým odhadem σ^2 je výběrový rozptyl S^2. Jeho hodnotu vypočítáme příkazem var(). Asymptoticky nestranným bodovým odhadem σ je výběrová směrodatná odchylka S. Její hodnotu vypočítáme příkazem sd().

```
9  m <- mean(weight.CF) # 3012,832
10 s2 <- var(weight.CF) # 455722,4
11 s <- sd(weight.CF) # 675,0722
```

Náhodný výběr porodních hmotností u novorozenců ženského pohlaví pochází z rozdělení se střední hodnotou μ a rozptylem σ^2, kde odhad střední hodnoty $m = 3012,83\,\text{mm}$ a odhad rozptylu $s^2 = 455722,40^2\,\text{mm}^2$ (odhad směrodatné odchylky $s = 675,07\,\text{mm}$). ★

Příklad 7.2. Neřešený příklad
Načtěte datový soubor 18-more-samples-variances-clavicle.txt obsahující údaje o délce klíční kosti z pravé strany v mm (cla.L) u mužů anglické populace, řecké populace, indické populace z Amritsaru a indické populace z Varanasi. Za předpokladu, že náhodný výběr délek klíčních kostí z pravé strany u mužů řecké populace pochází z rozdělení

se střední hodnotou μ a rozptylem σ^2, vypočítejte (a) bodový odhad μ; (b) bodový odhad σ^2; (c) bodový odhad σ. Výsledné hodnoty řádně interpretujte.
Výsledky: (a) $m = 153,52\,\text{mm}$; (b) $s^2 = 83,16\,\text{mm}^2$; (c) $s = 9,12\,\text{mm}$. ★

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 rozdělení se středními hodnotami μ_1, μ_2, rozptyly σ_1^2, σ_2^2, kovariancí σ_{12} a koeficientem korelace ρ, $n \geq 2$.

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,
- $S_1 = \sqrt{S_1^2}$, $S_2 = \sqrt{S_2^2}$ … výběrové směrodatné odchylky,
- $S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ … 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 obsahující údaje o výšce mozkovny v mm (skull.pH) a morfologické výšce tváře v mm (face.H) u mužů a žen starověké egyptské populace. Za předpokladu, že náhodný výběr výšek mozkovny a morfologických výšek tváře u žen pochází z dvourozměrného rozdělení s kovariancí σ_{12} a koeficientem korelace ρ, vypočítejte (a) bodový odhad σ_{12}; (b) bodový odhad ρ. Výsledné hodnoty řádně interpretujte.

Řešení příkladu 7.3
Datový soubor načteme a z načtených dat vybereme pouze řádky týkající se žen a sloupce skull.pH a face.H. Z výběru odstraníme řádky s chybějícími hodnotami. V dalším kroku separujeme z výběru hodnoty výšky mozkovny a hodnoty morfologické výšky tváře.

```
12 data <- read.delim("05-one-sample-correlation-skull-mf.txt", sep = "\t")
13 head(data, n = 4)
```

```
   id   pop sex skull.pH face.H                                                  14
1 416 egant   m      149    123                                                  15
2 417 egant   m      140    112                                                  16
3 420 egant   m      134     NA                                                  17
4 421 egant   m      137     NA                                                  18
```

```
19 data.F <- na.omit(data[data$sex == "f", c("skull.pH", "face.H")])
20 skull.pHF <- data.F$skull.pH
21 face.HF <- data.F$face.H
```

Nestranným bodovým odhadem σ_{12} je výběrová kovariance S_{12}. Její hodnotu vypočítáme příkazem cov(). Vychýleným bodovým odhadem koeficientu korelace ρ je výběrový koeficient korelace R_{12}. Jeho hodnotu vypočítáme příkazem cor() s argumentem method = "pearson".

```
22 s12 <- cov(skull.pHF, face.HF) # 1,808192
23 r12 <- cor(skull.pHF, face.HF, method = "pearson") # 0,06417166
```

Náhodný výběr výšek mozkovny a morfologických výšek tváře u žen starověké egyptské populace pochází z dvourozměrného normálního rozdělení s kovariancí σ_{12} a koeficientem korelace ρ, kde odhad kovariance $s_{12} = 1,8082$ a odhad koeficientu korelace $r_{12} = 0,0642$. Mezi výškou mozkovny a morfologickou výškou tváře u žen starověké egyptské populace existuje velmi nízký stupeň přímé lineární závislosti (viz kapitola 3, sekce 3.4.5 a tabulka 3.2). ★

Příklad 7.4. Neřešený příklad
Načtěte datový soubor 19-more-samples-correlations-skull.txt obsahující údaje o výšce nosu v mm (nose.H) a šířce nosu v mm (nose.B) u mužů bantuské, čínské, malajské, německé a peruánské populace. Za předpokladu, že náhodný výběr výšek nosu a šířek nosu u mužů peruánské populace pochází z dvourozměrného rozdělení s kovariancí σ_{12} a koeficientem korelace ρ, vypočítejte (a) bodový odhad σ_{12}; (b) bodový odhad ρ. Výsledné hodnoty řádně interpretujte.
Výsledky: (a) $s_{12} = 0,6990$; (b) $r_{12} = 0,1371$; nízký 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 rozdělení se střední hodnotou μ_1 a rozptylem σ^2, druhý z rozdělení 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 obsahující údaje o výšce horní části tváře v mm (upface.H) u mužů bantuské, čínské, malajské, německé a peruánské populace. Za předpokladu, že náhodný výběr výšek horní části tváře u mužů malajské populace pochází z rozdělení se střední hodnotou μ_1 a rozptylem σ^2 a náhodný výběr výšek horní části tváře u mužů čínské populace pochází z rozdělení se střední hodnotou μ_2 a rozptylem σ^2, vypočítejte (a) bodový odhad $\mu_1 - \mu_2$; (b) bodový odhad σ^2. Výsledné hodnoty řádně interpretujte.

Řešení příkladu 7.5
Datový soubor načteme. Z načtených dat vybereme pouze řádky týkající se mužů malajské populace a sloupec upface.H. Z výběru odstraníme řádky s chybějícími hodnotami. Stejný postup zopakujeme pro muže čínské populace.

```
24 data <- read.delim("15-anova-means-skull.txt", sep = "\t")
25 head(data, n = 4)
```

```
  id pop sex upface.H                                                    26
1  1 nem   m       73                                                    27
2  2 nem   m       73                                                    28
3  3 nem   m       67                                                    29
4  4 nem   m       75                                                    30
```

```
31 upface.HM <- na.omit(data[data$pop == "mal", "upface.H"])
32 upface.HC <- na.omit(data[data$pop == "cin", "upface.H"])
```

Nestranným bodovým odhadem $\mu_1 - \mu_2$ je rozdíl výběrových průměrů $M_1 - M_2$. Hodnoty výběrových průměrů vypočítáme příkazem mean().

```
33 m1 <- mean(upface.HM) # 70,13043
34 m2 <- mean(upface.HC) # 72
35 m1 - m2 # -1,869565
```

Nestranným bodovým odhadem σ^2 je vážený průměr výběrových rozptylů S_*^2. Výpočet provedeme pomocí vzorce $S_*^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$, kde hodnoty výběrových směrodatných odchylek S_1 a S_2 vypočítáme příkazem sd() a rozsahy náhodných výběrů n_1 a n_2 zjistíme příkazem length().

```
36 s1 <- sd(upface.HM) # 4,952459
37 s2 <- sd(upface.HC) # 4,563281
38 n1 <- length(upface.HM) # 69
39 n2 <- length(upface.HC) # 18
40 sh2 <- ((n1 - 1) * s1 ^ 2 + (n2 - 1) * s2 ^ 2) / (n1 + n2 - 2) # 23,78619
```

Náhodný výběr výšek horní části tváře u mužů malajské populace pochází z rozdělení se střední hodnotou μ_1, jejíž odhad je o $1,87\,\text{mm}$ menší než odhad střední hodnoty μ_2 rozdělení náhodného výběru výšek horní části tváře u mužů čínské populace. Odhad rozptylu σ^2 obou výběrů $s_*^2 = 23,79\,\text{mm}^2$. ★

Příklad 7.6. Neřešený příklad
Načtěte datový soubor 01-one-sample-mean-skull-mf.txt obsahující údaje o délce mozkovny v mm (skull.L) u mužů a žen starověké egyptské populace. Za předpokladu, že náhodný výběr délek mozkovny u mužů pochází z rozdělení se střední hodnotou μ_1 a rozptylem σ^2 a náhodný výběr délek mozkovny u žen pochází z rozdělení se střední hodnotou μ_2 a rozptylem σ^2, vypočítejte (a) bodový odhad $\mu_1 - \mu_2$; (b) bodový odhad σ^2. Výsledné hodnoty řádně interpretujte.
Výsledky: (a) $m_1 - m_2 = 7,50\,\text{mm}$; (b) $s_*^2 = 39,95\,\text{mm}^2$. ★

7.4 Vybrané statistiky pro alespoň 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 rozdělení se střední hodnotou μ_1 a rozptylem σ^2 až r-tý z rozdělení 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, z nichž alespoň jedna je nenulová.

Definice statistik

- $\sum_{j=1}^r c_j M_j$ … lineární kombinace r výběrových průměrů,
- $S_*^2 = \frac{\sum_{j=1}^r (n_j-1)S_j^2}{n-r}$ … vážený průměr r 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 obsahující údaje o délce klíční kosti z pravé strany v mm (cla.L) u mužů anglické populace, řecké populace, indické populace z Amritsaru a indické populace z Varanasi. Za předpokladu, že náhodný výběr délek klíčních kostí z pravé strany u mužů řecké populace, resp. indické populace z Amritsaru, resp. indické populace z Varanasi pochází z rozdělení se střední hodnotou μ_1, resp. μ_2, resp. μ_3 a rozptylem σ^2, vypočítejte (a) bodový odhad $\sum_{j=1}^r c_j \mu_j$, kde $c_1 = -2$, $c_2 = 2$, $c_3 = 3$; (b) bodový odhad σ^2. Výsledné hodnoty řádně interpretujte.

Řešení příkladu 7.7
Datový soubor načteme. Z načtených dat vybereme pouze řádky týkající se mužů řecké populace a sloupec cla.L. Z výběru odstraníme řádky s chybějícími hodnotami. Stejný postup zopakujeme pro muže indické populace z Amritsaru a pro muže indické populace z Varanasi.

```
data <- read.delim("18-more-samples-variances-clavicle.txt", sep = "\t")
head(data, n = 4)
```

```
  pop sex cla.L
1 gre   m   147
2 gre   m   148
3 gre   m   160
4 gre   m   140
```

```
cla.LG <- na.omit(data[data$pop == "gre", "cla.L"])
cla.LA <- na.omit(data[data$pop == "ind1", "cla.L"])
cla.LV <- na.omit(data[data$pop == "ind2", "cla.L"])
```

V tomto příkladu pracujeme s $r = 3$ nezávislými náhodnými výběry. Nestranným bodovým odhadem $\sum_{j=1}^{r} c_j \mu_j$ je lineární kombinace tří výběrových průměrů $\sum_{j=1}^{r} c_j M_j$. Poznamenejme, že při definování vektoru mj je třeba striktně dodržet pořadí výběrových průměrů podle populací (nejprve řecká populace, pak indická populace z Amritsaru a nakonec indická populace z Varanasi). Jedině tak bude pořadí výběrových průměrů odpovídat pořadí příslušných koeficientů c_1, c_2 a c_3.

```
cj <- c(-2, 2, 3)
mj <- c(mean(cla.LG), mean(cla.LA), mean(cla.LV)) # 153,5213; 145,5667; 141,4938
sum(cj * mj) # 408,5723
```

Nestranným bodovým odhadem σ^2 je vážený průměr tří výběrových rozptylů S_*^2, který vypočítáme přepisem vzorce $S_*^2 = \frac{\sum_{j=1}^{r}(n_j-1)S_j^2}{n-r}$.

```
r <- 3
sj <- c(sd(cla.LG), sd(cla.LA), sd(cla.LV)) # 9,118961; 8,733432; 8,220209
nj <- c(length(cla.LG), length(cla.LA), length(cla.LV)) # 94; 120; 81
n <- sum(nj) # 295
sh2 <- sum((nj - 1) * sj ^ 2) / (n - r) # 76,08107
```

Lineární kombinace výběrových průměrů délek klíčních kostí z pravé strany u mužů řecké populace, indické populace z Amritsaru a indické populace z Varanasi $-2m_1 + 2m_2 + 3m_3 = 408,57\,\text{mm}$, vážený průměr výběrových rozptylů těchto tří populací $s_*^2 = 76,08\,\text{mm}^2$.

★

Příklad 7.8. Neřešený příklad
Načtěte datový soubor 19-more-samples-correlations-skull.txt obsahující údaje o interorbitální šířce nosu v mm (interorb.B) u mužů bantuské, čínské, malajské, německé a peruánské populace. Za předpokladu, že náhodný výběr interorbitálních šířek nosu u mužů německé, resp. bantuské, resp. malajské, resp. čínské, resp. peruánské populace pochází z rozdělení se střední hodnotou μ_1, resp. μ_2, resp. μ_3, resp. μ_4, resp. μ_5 a rozptylem σ^2, vypočítejte (a) bodový odhad $\sum_{j=1}^{r} c_j \mu_j$, kde $c_1 = 1$, $c_2 = 2$, $c_3 = -1$, $c_4 = 0$, $c_5 = -1$; (b) bodový odhad σ^2. Výsledné hodnoty řádně interpretujte.
Výsledky: (a) $m_1 + 2m_2 - m_3 - m_5 = 30,77\,\text{mm}$; (b) $s_*^2 = 5,61\,\text{mm}^2$. ★