# 6 Číselné charakteristiky, Matematická statistika, Bodové a intervalové odhady parametrů

## 6.1 Číselné charakteristiky náhodných veličin

- $F(x)$, $p(x)$, $f(x)$ ... funkcionální charakteristiky
  - obsahují veškerou informaci o chování náh.veličiny
- někdy nás zajímají pouze rysy chování náh.veličiny → *číselné charakteristiky*
  - kvantily ($x_{0.25}$, $x_{0.5}$, $x_{0.75}$, apod.)
  - střední hodnota
  - rozptyl / směrodatná odchylka
  - kovariance
  - korelace

**Kvantily vybraných spojitých rozdělení; $\alpha$-kvantil**

- $\alpha$-kvantil náh.veličiny $X$ ... $x_\alpha$
- obdoba $\alpha$-kvantilu v popisné statistice
- křivka hustoty:
  - plocha pod křivkou ... pst ... = 1
  - tuto plochu rozdělíme na 2 části
    - tmavá plocha $\alpha$
    - světlá plocha $1 - \alpha$

- $\alpha$-kvantil ... číslo, takové, že $\Pr(X \leq x_\alpha) = \alpha$
- pst, že náhodná veličina $X$ je menší nebo rovna $x_\alpha$ je rovna $\alpha$
- speciální kvantily
  - medián ... $x_{0.5}$
  - 1.kvartil ... $x_{0.25}$
  - 3.kvartil ... $x_{0.75}$
- **Standardizované normální rozdělení**
  - 
    - $X \sim N(0,1)$
    - $\alpha$-kvantil ... $u(\alpha)$

    * symetrické okolo 0 ... $u(\alpha) = -u(1-\alpha)$
    * qnorm(alpha)

- $\chi^2$ **rozdělení** s $n$ stupni volnosti
  - (Pearsonovo rozdělení)
  - $X \sim \chi^2(n)$
  - $\alpha$-kvantil ... $\chi^2_n(\alpha)$
  - nesymetrické
  - qchisq(alpha,n)
- **Studentovo rozdělení** s $n$ stupni volnosti
  - $X \sim t(n)$
  - $\alpha$-kvantil ... $t_n(\alpha)$
  - symetrické okolo 0 ... $t_n(\alpha) = -t_n(1-\alpha)$
  - qt(alpha,n)
- **Fisherovo rozdělení** s $n_1$ a $n_2$ stupni volnosti
  - (Fisherovo-Snedecorovo rozdělení)
  - $X \sim F(n_1, n_2)$
  - $\alpha$-kvantil ... $F_{n_1,n_2}(\alpha)$
  - nesymetrické, ale $F_{n_1,n_2}(\alpha) = \dfrac{1}{F_{n_1,n_2}(1-\alpha)}$
  - qf(alpha, n1, n2)

**Příklad 6.1.** Najděte medián a horní a dolní kvartil náhodné veličiny $U \sim N(0,1)$.

```
qnorm(0.5)
qnorm(0.25)
qnorm(0.75)
```

**Příklad 6.2.** Najděte dolní kvartil náhodné veličiny $X \sim N(3,5)$.

```
qnorm(0.25, 3, sqrt(5))
```

**Příklad 6.3.** Určete kvantil $\chi^2_{25}(0.025)$.

```
qchisq(0.025, 25)
```

**Příklad 6.4.** Určete kvantily $t_{30}(0.99)$ a $t_{14}(0.05)$.

```
qt(0.99, 30)
qt(0.05, 14)
```

**Příklad 6.5.** Určete kvantily $F_{5,20}(0.975)$ a $F_{2,10}(0.05)$.

```
qf(0.975, 5,20)
qf(0.05,  2,10)
```

## 6.2 Základní pojmy matematické statistiky

- popisná statistika ... datový soubor $\rightarrow$ závěry o datovém souboru
- matematická statistika ... náhodný výběr $\rightarrow$ statistiky $\rightarrow$ závěry o tvaru rozdělení a parametrech
- $X_1, \ldots, X_n$ – stoch.nezáv.náh.veličiny, které mají všechny stejné rozložení $L(\theta) \rightarrow X_1, \ldots, X_n$ ... *náhodný výběr* rozsahu $n$ z rozdělení $L(\theta)$
- číselné realizace $x_1, \ldots, x_n$ náh.výběru $X_1, \ldots, X_n$ tvoří datový soubor
- *statistika* = libovolná funkce náhodného výběru: $T = T(X_1, \ldots, X_n)$
- Statistiky – jednovýběrové:
  Nechť $X_1, \ldots, X_n$ je náhodný výběr, $n \geq 2$.

  1. *Výběrový průměr*
  $$M = \frac{1}{n}\sum_{i=1}^{n} X_i$$
  2. *Výběrový rozptyl*
  $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2$$
  3. *Výběrová směrodatná odchylka*
  $$S = \sqrt{S^2}$$
  4. *Výběrová distribuční funkce* $F_n(x)$ ... průměrný počet těch veličin $X_i$, pro něž platí $X_i \geq x$.

- Statistiky – dvouvýběrové:
  Nechť $(X_1, Y_1), \ldots (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozdělení. $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$ a $M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

  1. *Výběrová kovariance*
  $$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$$
  2. *Výběrový koeficient korelace*
  $$R_{12} = \frac{S_{12}}{S_1 S_2}$$

## 6.3 Bodové a intervalové odhady parametrů

- $X_1 \ldots X_n$ ... náhodný výběr z rozdělení $L(\theta)$ s parametrem $\theta$.
- $\theta$ neznáme; chceme ho odhadnout
- bodovým odhadem parametru $\theta$ je nějaká vhodná statistika $T_n = T(X_1 \ldots X_n)$

- intervalovým odhadem parametru $\theta$ je interval $(D, H)$, kde $D$, $H$ jsou fce náh.výběru $D = D(X_1 \dots X_n)$, $H = H(X_1 \dots X_n)$ a který s dostatečně velkou pstí pokrývá hodnotu parametru $\theta$

- typy bodových odhadů

  1. nestranný …hodnotu param. $\theta$ ani nepodhodnocuje, ani nenadhodnocuje … $ET_n = \theta$
  2. vychýlený …není-li odhad nestranný, je vychýlený
  3. asymptotický …s rostoucím $n$ se jeho přesnost zvětšuje

- vlastnosti bodových odhadů

- $X_1, \dots X_n$ …náh. výběr se střední hodnotou $\mu$, rozptylem $\sigma^2$.

  1. $M$ je nestranný odhadem $\mu$ … $EM = \mu$
  2. $DM = \frac{\sigma^2}{n}$
  3. $S^2$ je nestranným odhadem $\sigma^2$ … $ES^2 = \sigma^2$

- $(X_1, Y_1), \dots (X_n, Y_n)$ …náhodný výběr z dvouroz. rozložení s kovariancí $\sigma_{12}$ a koeficientem korelace $\rho$.

  1. $E(S_{12})$ je nestranným odhadem $\sigma_{12}$ … $E(S_{12}) = \sigma_{12}$
  2. $ER_{12}$ je asymptoticky nestranným odhadem $\rho$ … $ER_{12} \approx \rho$

**Příklad 6.6.** Ve 12-ti náhodně vybraných internetových obchodech byly zjištěny následující ceny deskriptoru artefaktů (v Kč): $102, 99, 106, 103, 96, 98, 100, 105, 103, 98, 104, 107$. Těchto 12 hodnot považujeme za realizace náhodného výběru $X_1, \dots, X_{12}$ z rozdělení, které má střední hodnotu $\mu$ a rozptyl $\sigma^2$.

a) Určete nestranné bodové odhady neznámé střední hodnoty $\mu$ a neznámého rozptylu $\sigma^2$.

b) Najděte výběrovou distribuční funkci $F_{12}(x)$ a nakreslete její graf.

ad a) Vypočteme realizaci výběrového průměru

$$m = \frac{1}{12}(102 + 99 + \cdots + 107) = 101.75\,\text{Kč}$$

Vypočteme realizaci výběrového rozptylu:

$$s^2 = \frac{1}{11}\left[(102 - 101.75)^2 + (99 - 101.75)^2 + \cdots + (107 - 101.75)^2\right] = 12.39\,\text{Kč}^2$$

```
x <- c(96, 98, 98, 99, 100, 102, 103, 103, 104, 105, 106, 107)
n <- length(x)
(m <- mean(x))
(s2 <- var(x))

# Vyberova distribucni funkce
t  <- unique(sort(x))
y  <- sort(x)
```

```
nt <- length(t)

cetnost <- NULL
for(i in 1:nt){
  cetnost[i] <- sum(y<=t[i])}
Fx <- cetnost/n
t(round(Fx, digits=4))

# graf vyberove distribucni funkce
x <- c(min(t)-1,t, max(t)+1)
y <- c(0,Fx,1)
plot(x, y, type='n', xlab='x', ylab='F(x)',
     main='Vyberova␣distribucni␣funkce')
abline(h=seq(0,1,by=0.1), col='grey85')
abline(v=seq(95, 108,by=2), col='grey85')
lines(x,y, type='s', col='red', lwd=2)
arrows(96,0,95,0, col='red', lwd=2, length=0.1)
arrows(107,1,108,1, col='red', lwd=2, length=0.1)
```

**Příklad 6.7.** Z archivních materiálů (Schmidt, 1888) máme k dispozici původní kraniometricé údaje o výšce horní části tváře (v mm) u 13 mužů bantuské populace. Hodnoty výšky horní části tváře jsou 67, 67, 63, 68, 70, 70, 75, 74, 80, 77, 77, 67, 64.

a) Odhadněte střední hodnotu, rozptyl a směrodatnou odchylku výšky horní části tváře.

b) Odhadněte pravděpodobnost že výška tváře bantuského muže bude vyšší než 72 mm.

```
x <- c(67, 67, 63, 68, 70, 70, 75, 74, 80, 77, 77, 67, 64)
x <- sort(x)
n <- length(x)
s2 <- var(x)
s  <- sd(x)
Tab <- data.frame(m=m, s2=s2, s=s, row.names='akcie')
round(Tab, digits=2)

# P(X>=70)
pst <- sum(x>=70)/length(x)
pst2 <- 1-sum(x<70)/length(x)
round(pst,4)
round(pst2,4) # 0.5385
```

**Poznámka:** Dodělat analogicky pro zbylé populace a dát jako procvičovací příklady.

**Příklad 6.8.** Máme k dispozici antropometrické údaje mladých dospělých lidí, převážně studentů vysokých škol z Brna a Ostravy, konkrétně údaje o šířce hlavy (head.W), šířce tváře (bizyg.W) a šířce dolní čelisti (bigo.W). Dále máme u každého studenta uveden údaj o pohlaví (sex), přičemž v databázi máme celkem 75 mužů a 100 žen. Zaměřme se na údaje týkající se mužů. Najděte bodové odhady kovariance $\sigma_{1,2}$ a korelace $\rho$ pro náhodné proměnné $X_1$ …šířka hlavy a $X_2$ …šířka tváře.

```
data <- read.delim('16-anova-head.txt', sep='\t')
muzi <- data[data$sex=='m',]
head.w <- muzi$head.W
bizyg.w <- muzi$bizyg.W
cov(head.w, bizyg.w)  # 31.83
cor(head.w, bizyg.w)  # 0.6785
plot(head.w, bizyg.w)
```

### 6.3.1 INTERVALY SPOLEHLIVOSTI

- $X_1 \dots X_n$ ...náh.výběr z rozdělení $L(\theta)$, $\theta$ je parametr, $\alpha \in (0,1)$
- interval $(D, H)$ ... $100(1-\alpha)\%$ oboustranný IS pro param. $\theta$
- interval $(D, \infty)$... $100(1-\alpha)\%$ levostranný IS pro param. $\theta$
- interval $(-\infty, H)$... $100(1-\alpha)\%$ pravostranný IS pro param. $\theta$
- $\alpha$ se nazývá *riziko*, $(1-\alpha)$ se nazývá *spolehlivost*.

### 6.3.2 Konstrukce intervalů spolehlivosti

- konečný tvar IS pro param. $\theta$ odvozujeme z příslušné pivotovy statistiky
- pivotová statistika = statistika, jejíž rozdělení je známé a nezávisí na parametru $\theta$
  - používá se také k testování hypotéz
- příklad odvození IS z pivotovy statistiky viz studijní materiály

**Příklad 6.9.** Vezměte data z příkladu 7.3. Vypočítejte

- 95 % empirický interval spolehlivosti pro střední hodnotu délky šířky čelisti u mužů. (106.4945; 109.1322
- 95 % pravostranný empirický interval spolehlivosti pro střední hodnotu šířky dolní čelisti u mužů $(-\infty; 109.1352)$.
- 95 % levostranný empirický interval spolehlivosti pro střední hodnotu šířky dolní čelisti u mužů $(106.4914; \infty)$.

```
data <- read.delim('16-anova-head.txt', sep='\t')
muzi <- data[data$sex=='m',]
head.w <- muzi$head.W
bizyg.w <- muzi$bizyg.W
cov(head.w, bizyg.w)  # 31.83
cor(head.w, bizyg.w)  # 0.6785
plot(head.w, bizyg.w)
```

```
data <- read.delim('16-anova-head.txt', sep='\t')
muzi <- data[data$sex=='m',]
head.w <- muzi$head.W
bizyg.w <- muzi$bizyg.W
bigo.w <- muzi$bigo.W

m <- mean(bigo.w)
s <- sd(bigo.w)
alpha <- 0.05
n <- length(bigo.w)
dh <- m-s/sqrt(m)*qt(1-alpha/2, n-1)
hh <- m-s/sqrt(m)*qt(alpha/2, n-1)

round(dh, 4)
```

```
round(hh, 4)

dh <- m-s/sqrt(n)*qt(1-alpha, n-1)
dh

hh <- m-s/sqrt(n)*qt(alpha, n-1)
hh
```

**Příklad 6.10.** Při kontrolních zkouškách životnosti 16 žárovek byl stanoven odhad $m = 3000\,h$ střední hodnoty jejich životnosti. Z dřívějších zkoušek je známo, že životnost žárovky se řídí normálním rozdělením se směrodatnou odchylkou $\sigma = 20\,h$. Vypočtěte

a) 99 % empirický interval spolehlivosti pro střední hodnotu životnosti $(2987.1; 3012.9)$;

b) 90 % levostranný empirický interval spolehlivosti pro střední hodnotu životnosti $(29993.6; \infty)$;

c) 95 % pravostranný empirický interval spolehlivosti pro střední hodnotu životnosti $(-\infty; 3008.2)$.

ad a)

$$d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha} = 3000 - \frac{20}{\sqrt{16}} 2.57583 = 2987.1$$
$$h = m - \frac{\sigma}{\sqrt{n}} u_{\alpha} = 3000 + \frac{20}{\sqrt{16}} 2.57583 = 3012.9$$

```
m <- 3000
s <- 20
n <- 16

# a)
alpha <- 0.01
(dh <- m-s/sqrt(n)*qnorm(1-alpha/2))
(hh <- m-s/sqrt(n)*qnorm(alpha/2))
```

2987 h a 6 min $< \mu <$ 3012 h a 54 min s pravděpodobností 0.99.

ad b)

$$d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha} = 3000 - \frac{20}{\sqrt{16}} 1.28155 = 2993.6$$

```
alpha <- 0.1
(dh <-m-s/sqrt(n)*qnorm(1-alpha))
```

2993 h a 36 min $< \mu$ s pravděpodobností 0.9.

ad c)

$$h = m - \frac{\sigma}{\sqrt{n}} u_{\alpha} = 3000 + \frac{20}{\sqrt{16}} 1.95996 = 3008.2$$

```
  alpha <- 0.05
(hh <- m-s/sqrt(n)*qnorm(alpha))
```

3009 h a 48 min $> \mu$ s pravděpodobností 0.95.