Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Aplikovaná statistika I

Téma 6: Bodové a intervalové odhady

Veronika Bendová

bendova.veroonika@gmail.com

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
- $X_1, \ldots, X_n$ – náhodné veličiny, které mají všechny stejné rozdělení $L(\theta)$ $\rightarrow$ dohromady tvoří **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: $T = T(X_1, \ldots, X_n)$

Jednorozměrné statistiky (jeden výběr; jeden znak)
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. *výběrový koeficient variace*
$$v = \frac{s}{m}$$

Dvourozměrné statistiky (jeden výběr; dva znaky)
Nechť $(X_1, Y_1), \dots, (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozdělení, m_1 a m_2 jsou výběrové průměry a s_1^2 a s_2^2 jsou výběrové rozptyly.

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*

$$r_{12} = \frac{s_{12}}{\sqrt{s_1^2 s_2^2}} = \frac{s_{12}}{s_1 s_2}$$

Dvouvýběrové statistiky (dva nebo více výběrů; jeden znak)
Nechť $X_{11}, \dots, X_{1n_1}$ je náhodný výběr, $X_{21}, \dots, X_{2n_2}$ je na něm nezávislý náhodný výběr, ..., $X_{k1}, \dots, X_{kn_k}$, je náhodný výběr nezávislý na všech předcházejících náhodných výběrech, n_1, n_2, ..., $n_k \geq 2$. Nechť dále s_1^2, s_2^2, ..., s_k^2 jsou výběrové rozptyly.

1. *vážený průměr dvou výběrových rozptylů* ($k = 2$)

$$S_*^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

2. *vážený průměr k výběrových rozptylů*

$$S_*^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_k - 1)S_k^2}{n_1 + n_2 + \cdots + n_k - k}$$

Bodové a intervalové odhady parametrů

- $X_1 \dots X_n$ ... náhodný výběr z rozdělení $L(\theta)$ s parametrem θ
- θ neznáme; chceme ho odhadnout
- **bodový odhad parametru** θ ... statistika $T_n = T(X_1 \dots X_n)$
- **intervalový odhad parametru** θ ... interval (D, H), který s dostatečně velkou pravděpodobností pokrývá hodnotu parametru θ; (D, H ... statistiky)

Bodový odhad parametru θ

- typy bodových odhadů
  - nestranný ... hodnotu param. θ ani nepodhodnocuje, ani nenadhodnocuje
  - vychýlený ... není-li odhad nestranný, je vychýlený
  - asymptotický ... s rostoucím n se přesnost odhadu zvětšuje
- vlastnosti bodových odhadů
  1. $X \sim N(\mu, \sigma^2)$
     - m je nestranným odhadem μ (parametr θ)
     - s^2 je nestranným odhadem σ^2 (parametr θ)
     - v je nestranným odhadem koeficientu variace $\frac{\sigma}{\mu}$ (parametr θ)
  2. $(X, Y)^T \sim \mathbf{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} = (\mu_1, \mu_2)$, $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$
     - s_{12} je nestranným odhadem σ_{12} (parametr θ)
     - r_{12} je asymptoticky nestranným odhadem ρ (parametr θ)

Dataset: 21-goldman-tigara.csv

Datový soubor 21-goldman-tigara.csv obsahuje osteometrické údaje o délce stehenní kosti (v mm) a acetabulární výšce (v mm) z pravé a levé strany u mužů a žen z aljašské populace z kmene Tigara. Data pochází ze souboru dokumentovaných skeletů (Goldman, 2006).

Popis proměnných v datasetu:

- sex ... pohlaví jedince (m - muž, f - žena);
- pop ... populace (Tigara = aljašská populace z kmene Tigara);
- femur.LR ... délka stehenní kosti z pravé strany (v mm);
- femur.LL ... délka stehenní kosti z levé strany (v mm);
- acetab.HR ... acetabulární výška z pravé strany (v mm);
- acetab.HL ... acetabulární výška z levé strany (v mm).

Příklad 6.1. Bodové odhady parametrů μ a σ^2 normálního rozdělení

Načtěte datový soubor 21-goldman-tigara.csv. Nechť náhodná veličina X popisuje *délku stehenní kosti* (v mm) z pravé strany u mužů z kmene Tigara. Za předpokladu, že náhodná veličina $X \sim N(\mu, \sigma^2)$, stanovte nestranný (bodový) odhad (a) střední hodnoty μ; (b) rozptylu σ^2; (c) směrodatné odchylky σ; (d) koeficientu variace. Všechny vypočítané hodnoty řádně interpretujte.

Řešení příkladu 6.1

```r
data <- read.delim(...) # nacteni datoveho souboru
data.M <- data[..., ...] # vyber sloupcu femur.LR a acetab.HR pro muze z kmene Tigara
data.M <- na.omit(...) # odstraneni NA udaju
femur.LR <- data.M$... # vyber sloupce femur.LR z tabulky data.M
acetab.HR <- data.M$... # vyber sloupce acetab.HR z tabulky data.M
n <- length(...) # pocet udaju o delce stehenni kosti u muzu z kmene Tigara

m.LR  <- mean() # bodovy odhad stredni hodnoty mu
s2.LR <- var() # bodovy odhad rozptylu sigma^2
s.LR <-  sd() # bodovy odhad sm. odchylky sigma
v.LR <- ... # bodovy odhad koef. variace sigma / mu
tab <- data.frame(...) # souhrnna tabulka vysledku
```

```
      m     s2     s    v
1 427.9 539.17 23.22 0.05
```

Interpretace výsledků: Nestranný odhad střední hodnoty μ je .... mm. Nestranný odhad rozptylu σ^2 (resp. směrodatné odchylky σ) je .... mm^2 (resp. .... mm). Nestranný odhad koeficientu variace je .... Délka stehenní kosti z pravé strany u mužů z kmene Tigara se pohybuje okolo hodnoty .... mm se směrodatnou odchylkou .... mm. Směrodatná odchylka představuje .... % aritmetického průměru.

Příklad 6.2. Bodové odhady parametrů $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ normálního rozdělení

Načtěte datový soubor 21-goldman-tigara.csv. Nechť náhodná veličina X popisuje *délku stehenní kosti* (v mm) z pravé strany a náhodná veličina Y popisuje *acetabulární výšku* (v mm) z pravé strany u mužů z kmene Tigara. Za předpokladu, že náhodný vektor $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, stanovte (a) nestranný (bodový) odhad vektoru středních hodnot $\boldsymbol{\mu}$; (b) nestranný (bodový) odhad kovariance σ_{12}; (c) asymptoticky nestranný (bodový) odhad korelačního koeficientu ρ; (d) nestranný (bodový) odhad varianční matice $\boldsymbol{\Sigma}$.

Řešení příkladu 6.2

```
m.HR  <- ... # bodovy odhad stredni hdonoty
s2.HR <- ... # bodovy odhad rozptylu sigma^2
s.HR <- ... # bodovy odhad sm. odchylky
s12 <- cov(..., ...) # bodovy odhad kovariance sigma12
r12 <- cor(..., ...) # bodovy odhad korel. koef. rho
s12 <- ... # bodovy odhad kovariance sigma12 (prepis vzorce)
r12 <- ... # bodovy odhad korel. koef. rho (prepis vzorce)
tab <- data.frame(...) # souhrnna tabulka vysledku
```

```
   m.LR  m.HR  s2.LR s2.HR  s.LR s.HR   s12  r12
1 427.9 51.93 539.17  9.77 23.22 3.13 42.11 0.58
```

```
   m.LR   m.HR   s2.LR s2.HR   s.LR s.HR    s12   r12
1 427.9 51.93 539.17   9.77 23.22 3.13 42.11 0.58
```

Interpretace výsledků: Nestranný odhad vektoru středních hodnot $\boldsymbol{\mu}$ =(................ , ...)^T mm. Nestranný odhad kovariance σ_{12} je .... Asymptoticky nestranný odhad korelačního koeficientu ρ je .... Nestranný odhad varianční matice $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ je matice $\begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix}$, kde s_1^2 =........................... mm^2, s_2^2 = ... mm^2 a $s_{12} = r_{12}s_1s_2$ = ... Délka stehenní kosti z pravé strany mužů z kmene Tigara se pohybuje okolo hodnoty .... mm se směrodatnou odchylkou .... mm. Acetabulární výška z pravé strany se pohybuje okolo hodnoty .... mm se směrodatnou odchylkou .... mm. Mezi délkou stehenní kosti a acetabulární výškou z pravé strany existuje .... stupeň .... závislosti (r_{12} = ...).

Dataset: 21-goldman-shells.csv
Datový soubor 21-goldman-shells.csv obsahuje osteometrické údaje o délce kyčelní kosti z pravé a levé strany u mužů a žen ze tří japonských populací(Tsugumo Shell Mound, Yoshigo Shell Mound a Yasaki Shell Mound). Data pochází ze souboru dokumentovaných skeletů (Goldman, 2006).

Popis proměnných v datasetu:

- sex . . . pohlaví jedince (m - muž, f - žena);
- pop . . . populace (tsg = Tsugumo Shell Mound, yos = Yoshigo Shell Mound, yas = Yasaki Shell Mound);
- iblade.LR . . . délka kyčelní kosti z pravé strany (v mm);
- iblade.LL . . . délka kyčelní kosti z levé strany (v mm).

Příklad 6.3. Dvouvýběrové statistiky
Načtěte datový soubor 21-goldman-shells.csv. Vypočítejte vážený průměr výběrových rozptylů délek kyčelních kostí z levé strany u mužů (a) z populací Yoshigo Shell Mound a Yasaki Shell Mound; (b) ze všech tří uvedených populací.

Řešení příkladu 6.3

```
data <- read.delim(...) # nacteni datoveho souboru
ib.Tg <- data[..., ...] # vyber sloupce iblade.LL pro muze z pop. Tsugumo s.m.
ib.Yo <- data[..., ...] # vyber sloupce iblade.LL pro muze z pop. Yoshigo s.m.
ib.Ya <- data[..., ...] # vyber sloupce iblade.LL pro muze z pop. Yasaki s.m.
ib.Tg <- na.omit(...) # odstraneni NA hodnot z vektoru ib.Tg
ib.Yo <- ... # odstraneni NA hodnot z vektoru ib.Yo
ib.Ya <- ... # odstraneni NA hodnot z vektoru ib.Ya

n.Tg <- length(...) # pocet hodnot delek kycelnich kosti (Tsugumo s.m.)
n.Yo <- ... # pocet hodnot delek kycelnich kosti (Yoshigo s.m.)
n.Ya <- ... # pocet hodnot delek kycelnich kosti (Yasaki s.m.)

s2.Tg <- var(...) # odhad rozptylu sigma^2 (Tsugumo s.m.)
s2.Yo <- ... # odhad rozptylu sigma^2 (Yoshigo s.m.)
s2.Ya <- ... # odhad rozptylu sigma^2 (Yasaki s.m.)

sh.YoYa <- ... # (a) vazeny prumer vyb. rozptylu dvou pop. (prepis vzorce)
sh.TgYoYa <- ... # (b) vazeny prumer vyb. rozptylu tri pop. (prepis vzorce)
tab <- data.frame(...) # sumarizacni tabulka
```

```
  sh.YoYa sh.TgYoYa
1    16.5     24.35
```

Interpretace výsledků: Vážený průměr výběrových rozptylů délek kyčelních kostí z levé strany mužů z populací Yoshigo Shell Mound a Yasaki Shell Mound $s^2_{YoYa*} =$ .... mm^2. Vážený průměr výběrových rozptylů délek kyčelních kostí z levé strany mužů všech tří japonských populací $s^2_* =$ .... mm^2.

Intervalové odhady parametrů

Nechť $\alpha \in (0, 1)$; koeficient α nazýváme **riziko**; koeficient $(1 - \alpha)$ nazýváme **spolehlivost**

- Interval (D, H) ... $100(1 - \alpha)\%$ oboustranný IS pro param. θ
- Interval (D, ∞)... $100(1 - \alpha)\%$ levostranný IS pro param. θ
- Interval $(-\infty, H)$... $100(1 - \alpha)\%$ pravostranný IS pro param. θ

Tvary intervalů spolehlivosti pro $X \sim N(\mu, \sigma^2)$

1. IS pro μ, když σ^2 známe
   a. Oboustranný:
   $$(d, h) = (m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, m - \frac{\sigma}{\sqrt{n}} u_{\alpha/2})$$
   b. Levostranný:
   $$(d, \infty) = (m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty)$$
   c. Pravostranný:
   $$(-\infty, h) = (-\infty, m - \frac{\sigma}{\sqrt{n}} u_{\alpha})$$

   u_α je α kvantil standardizovaného normálního rozložení ... qnorm(alpha,0,1).

2. IS pro μ, když σ^2 neznáme

   a. Oboustranný:
   $$(d, h) = (m - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1), m - \frac{s}{\sqrt{n}} t_{\alpha/2}(n-1))$$

   b. Levostranný:
   $$(d, \infty) = (m - \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1), \infty)$$

   c. Pravostranný:
   $$(-\infty, h) = (-\infty, m - \frac{s}{\sqrt{n}} t_{\alpha}(n-1))$$

   $t_\alpha(n-1)$ je α kvantil studentova rozdělení o $n-1$ stupních volnosti ... qt(alpha, n-1).

3. IS pro σ^2, když μ neznáme

   a. Oboustranný:
   $$(d, h) = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} \right)$$

   b. Levostranný:
   $$(d, \infty) = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)}, \infty \right)$$

   c. Pravostranný:
   $$(-\infty, h) = \left(-\infty, \frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)} \right)$$

   $\chi^2_\alpha(n-1)$ je α kvantil χ^2 rozdělení o $n-1$ stupních volnosti... qchisq(alpha,n-1).

4. IS pro σ^2, když μ známe

   - existuje, ale neprobíráme ho, neboť není příliš využitelný v praxi

Tvary intervalů spolehlivosti pro $X \sim \mathrm{Alt}(p)$

a. Oboustranný:

$$(d, h) = \left(\widehat{p} - u_{1-\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/N}\ ;\ \widehat{p} - u_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/N}\right),$$

b. Levostranný:

$$(d, \infty) = \left(\widehat{p} - u_{1-\alpha}\sqrt{\widehat{p}(1-\widehat{p})/N}\ ;\ \infty\right),$$

c. Pravostranný:

$$(-\infty, h) = \left(-\infty,\ ;\ \widehat{p} - u_{\alpha}\sqrt{\widehat{p}(1-\widehat{p})/N}\right),$$

u_α je α kvantil standardizovaného normálního rozložení ... qnorm(alpha,0,1).

Příklad 6.4. Intervalové odhady parametrů normálního rozdělení

Načtěte datový soubor 21-goldman-tigara.csv. Nechť náhodná veličina X popisuje *délku stehenní kosti* (v mm) z pravé strany u mužů z kmene Tigara. Za předpokladu, že náhodná veličina $X \sim N(\mu, \sigma^2)$, stanovte (a) 95% intervalový odhad střední hodnoty μ; (c) 99% levostranný intervalový odhad rozptylu σ^2; (c) 90% pravostranný intervalový odhad směrodatné odchylky σ.

Řešení příkladu 6.4.

```
alpha <- ... # koeficient alpha pro (a)
dh.mu <- ... # dolni hranice 95% IS pro par. mu (prepis vzorce)
hh.mu <- ... # horni hranice 95% IS pro par. mu (prepis vzorce)

alpha <- ... # koeficient alpha pro (b)
D.sig2 <- ... # dolni hranice 99% levostr. IS pro par. sigma^2 (prepis vzorce)

alpha <- ... # koeficient alpha pro (c)
H.sig2 <-... # horni hranice 90% pravostr. IS pro par. sigma^2 (prepis vzorce)
H.sig <- sqrt(...) # odmocneni h. hranice pro sigma^2 -> h. hranice pro sigma
tab <- data.frame(...) # souhrnna tabulka vysledku
```

```
   dh.mu hh.mu    D.mu H.sig
1 418.09 437.7 297.83  28.9
```

Interpretace výsledků: 95% empirický interval spolehlivosti pro střední hodnotu μ má tvar .... To znamená, že .... $< \mu <$ .... s pravděpodobností 95 %. V 95 případech ze sta bude střední hodnota délky stehenní kosti z pravé strany u mužů z kmene Tigara nabývat hodnoty z intervalu .... mm.

99% levostranný empirický interval spolehlivosti pro rozptyl σ^2 má tvar .... To znamená, že $\sigma^2 >$ .... s pravděpodobností 99 %. V 99 případech ze sta bude rozptyl délky stehenní kosti z pravé strany u mužů z kmene Tigara větší / menší než .... mm^2.

90% pravostranný empirický interval spolehlivosti pro směrodatnou odchylku σ má tvar .... To znamená, že $\sigma <$ .... s pravděpodobností 90 %. V 90 případech ze sta bude směrodatná odchylka délky stehenní kosti z pravé strany u mužů z kmene Tigara větší / menší než .... mm.

Příklad 6.5. Bodový a intervalový odhad parametru p alternativního rozdělení
Načtěte datový soubor 17-anova-newborns-2.txt. Mějme náhodnou veličinu X popisující *ženské pohlaví* novorozenců. Za předpokladu, že náhodná veličina $X \sim \mathrm{Alt}(p)$, kde p je pravděpodobnost narození holčičky, stanovte (a) bodový odhad parametru p; (b) 95% intervalový odhad parametru p.

Řešení příkladu 6.5

```
data <- read.delim(...) # nacteni datoveho souboru
data <- na.omit(...) # odstraneni NA hodnot
sex <- ... # vyber sloupce sex.C z datove tabulky
N <- length(...) # pocet udaju o pohlavi
p <- ... # bodovy odhad parametru p
alpha <- ... # koeficient alpha
dh.p <- ... # dolni hranice 95% IS pro parametr p (prepis vzorce)
hh.p <-... # horni hranice 95% IS pro parametr p (prepis vzorce)
tab <- data.frame(...) # sumarizacni tabulka vysledku
```

```
       p  dh.p   hh.p
1 0.4794 0.453 0.5057
```

Interpretace výsledků: Bodový odhad pravděpodobnosti narození holčičky je .... K narození holčičky dojde s pravděpodobností .... %. 95% empirický IS pro pravděpodobnost narození holčičky p má tvar .... To znamená, že .... $< p <$ .... s pravděpodobností 95 %. Pravděpodobnost narození holčičky se pohybuje v rozmezí ....% – ....% s pravděpodobností 95 %.