7 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*
  $$R_{12} = \frac{S_{12}}{S_1 S_2}$$

7.1 Bodové a intervalové odhady parametrů

- $X_1 \ldots X_n$ … náhodný výběr z rozdělení $L(\theta)$ s parametrem θ.
- param. θ neznáme; chceme ho odhadnout pomocí náh. výběru
- bodovým odhadem parametru θ je nějaká vhodná statistika $T_n = T(X_1 \ldots X_n)$
- intervalovým odhadem parametru θ je interval (D, H), kde D, H jsou fce náh.výběru $D = D(X_1 \ldots X_n)$, $H = H(X_1 \ldots X_n)$ a který s dostatečně velkou pstí pokrývá hodnotu parametru θ
- typy bodových odhadů

  1. nestranný … hodnotu param. θ 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 μ, rozptylem σ^2.
  1. M je nestranný odhadem μ ... $EM = \mu$
  2. $DM = \frac{\sigma^2}{n}$
  3. S^2 je nestranným odhadem σ^2 ... $ES^2 = \sigma^2$
- $(X_1, Y_1), \dots (X_n, Y_n)$ ... náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ.
  1. $E(S_{12})$ je nestranným odhadem σ_{12} ... $E(S_{12}) = \sigma_{12}$
  2. ER_{12} je asymptoticky nestranným odhadem ρ ... $ER_{12} \approx \rho$

Příklad 7.1. 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 μ a rozptyl σ^2.

a) Určete nestranné bodové odhady neznámé střední hodnoty μ a neznámého rozptylu σ^2.

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

ad a) Vypočteme realizaci výběrového průměru

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

Vypočteme realizaci výběrového rozptylu:

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

```
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 7.2. Přírůstky cen akcií v % na burze v New Yorku u 10 náhodně vybraných společností dosáhly těchto hodnot: $10, 16, 5, 10, 12, 8, 4, 6, 5, 4$.

a) Odhadněte střední hodnotu, rozptyl a směrodatnou odchylku růstu cen akcií.

b) Odhadněte pravděpodobnost růstu cen akcií aspoň o 8.5 %.

```
x <- c(10, 16, 5, 10, 12, 8, 4, 6, 5, 4)
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>=8.5)
 pst <- sum(x>=8.5)/length(x)
pst <- 1-sum(x<8.5)/length(x)
round(pst,4)
```

Příklad 7.3. Bylo zkoumáno 9 vzorků půdy s různým obsahem fosforu (veličina X). Hodnoty veličiny Y označují obsah fosforu v obilných klíčcích (po 38 dnech), jež vyrostly na těchto vzorcích půdy.

číslo vzorku	1	2	3	4	5	6	7	8	9
X	1	4	5	9	11	13	23	23	28
Y	64	71	54	81	76	93	77	95	109

Těchto 9 dvojic hodnot považujeme za realizace náhodného výběru $(X_1, Y_1), \dots, (X_9, Y_9)$ z dvourozměrného rozdělení s kovariancí σ_{12} a koeficientem korelace ρ. Najděte bodové odhady kovariance σ_{12} a koeficientu korelace ρ.

```
x <- c(1, 4, 5, 9, 11, 13, 23, 23, 28)
y <- c(64, 71, 54, 81, 76, 93, 77, 95, 109)
cov(x,y)
cor(x,y)
```

7.1.1 INTERVALY SPOLEHLIVOSTI

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

7.1.2 Konstrukce intervalů spolehlivosti

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

Příklad 7.4. 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;

b) 90 % levostranný empirický interval spolehlivosti pro střední hodnotu životnosti;

c) 95 % pravostranný empirický interval spolehlivosti pro střední hodnotu životnosti.

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.