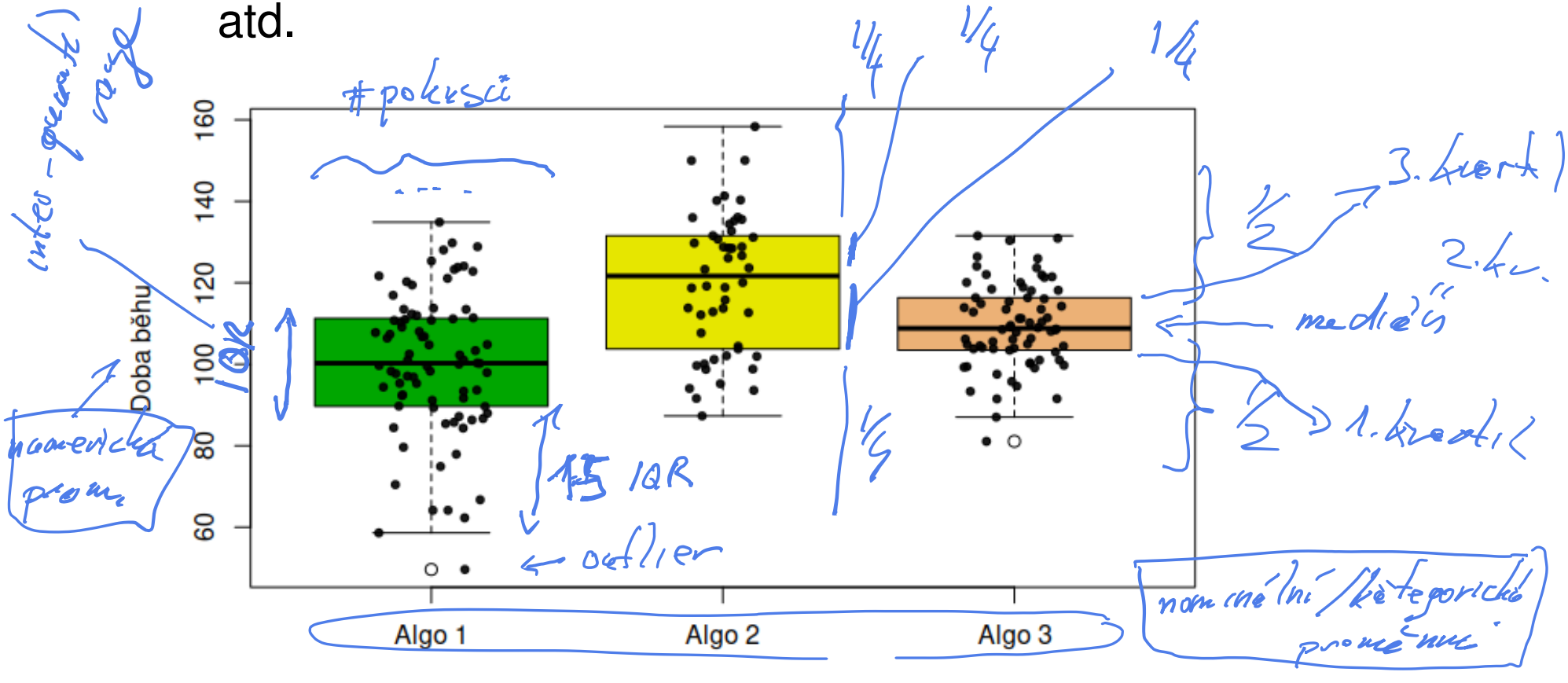
# NMAI059 Pravděpodobnost a statistika 1
# 9. přednáška

Robert Šámal

# Přehled

# Intro – explorační analýza dat (exploratory data analysis)

- posbíráme data (a dáme pozor na systémové chyby – nezávislost, nezaujatost, . . . )
- různé tabulky (třeba v Excelu a spol.)
- vhodné obrázky: histogram, krabicový diagram (boxplot), atd.



# Náhodný výběr

- s vracením
- bez vracení

$\Omega = \{$všechny $n$-tice obyvatel ČR$\}$

Pro $\omega = (\omega_1, \ldots, \omega_n)$ zvolíme $X_i = I(\omega_i$ je levák$)$.

$|\Omega| = N^n$

$|\Omega| = N(N-1)\ldots(N-n+1)$

$n = 17$, $N =$ #obyv. ČR $= 10.6$ mil.

$X_1, \ldots, X_n$ nejsou nezávislé

$X_1, \ldots, X_n$ jsou nezávislé

(ale rozdělení je malé zanedbáme)

# Statistika – přehled

- nezávislá měření – hodnoty n.n.v. $X_1, \ldots, X_n \sim F$ náhodný výběr s distribuční funkcí $F$ s rozsahem $n$

  ($n = 20$)

- neparametrické modely: povolujeme velkou třídu $F$

- parametrické modely: $F \in \{F_\vartheta : \vartheta \in \Theta\}$
- příklady:
  - $Pois(\lambda)$ (parametr $\vartheta = \lambda$, $\Theta = \mathbb{R}^+$)
  - $U(a, b)$ (parametr $\vartheta = (a, b)$, $\Theta = \mathbb{R}^2$)
  - $N(\mu, \sigma^2)$ (parametr $\vartheta = (\mu, \sigma)$, $\Theta = \mathbb{R} \times \mathbb{R}^+$)

- Všechny modely jsou špatné, ale některé jsou užitečné.

# Zkoumané úlohy – cíle konfirmační analýzy (confirmatory data analysis)

- bodové odhady
- intervalové odhady
- testování hypotéz
- (lineární) regrese

- *statistika* – libovolná funkce náhodného výběru, tj. např. aritmetický průměr, medián, maximum, atd.
  Tj. $T = T(X_1, \ldots, X_n)$.

$T(X_1, \ldots, X_n) = \frac{1}{n} \sum X_i$

$T_2(\quad) = \max(X_1, \ldots, X_n)$

# Přehled

# Výběrový průměr a rozptyl

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{S}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

$$\widehat{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Besselova korekce

# Cíle

## Definice

*Odhad* $\hat{\Theta}_n = \hat{\Theta}_n(X_1, \ldots, X_n)$ *parametru* $\vartheta$ *je*

- *nestranný (unbiased) – pokud* $\vartheta = \mathbb{E}(\hat{\Theta}_n)$
- *asymptoticky nestranný (asymptotically unbiased) – pokud* $\vartheta = \lim_{n\to\infty} \mathbb{E}(\hat{\Theta}_n)$
- *vychýlení (bias)* $bias_\vartheta(\hat{\Theta}_n) = \mathbb{E}(\hat{\Theta}_n) - \vartheta$
- *střední kvadratická chyba (mean squared error, MSE) je* $\mathbb{E}((\hat{\Theta} - \vartheta)^2)$

## Věta

$$MSE = bias_\vartheta(\hat{\Theta}_n)^2 + var_\vartheta(\hat{\Theta}_n)$$

# Parametry výběrového momentu a rozptylu

Dk 1) $E\bar{X}_n = \frac{1}{n}E X_1 + \dots + \frac{1}{n}E X_n = \mu$ $\quad$ ZVČ $\Rightarrow$ $\bar{X}_n$ je konzistentní.

## Věta

1. $\bar{X}_n$ *je konzistentní nestranný odhad* $\mu = E X_1$
2. $\bar{S}_n$ *je konzistentní asymptoticky nestranný odhad* $\sigma^2 = \operatorname{var} X_1$
3. $\widehat{S}_n$ *je konzistentní nestranný odhad* $\sigma^2$ ✓

$$\sum_{i=1}^{n}(X_i - \mu) = n\bar{X}_n - n\mu = n(\bar{X}_n - \mu)$$

(2) $E\bar{S}_n = E\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

$$E\frac{1}{n}\sum_{i=1}^{n}\big((X_i - \mu) - (\bar{X}_n - \mu)\big)^2$$

$$E\frac{1}{n}\sum_{i=1}^{n}\big((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X}_n - \mu) + (\bar{X}_n - \mu)^2\big)$$

$$E\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - E(\bar{X}_n - \mu)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i - \mu)^2 - E(\bar{X}_n - \mu)^2$$

$\operatorname{var}(X_i)$

$$= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \xrightarrow{n\to\infty} \sigma^2$$

$\operatorname{var}\bar{X}_n = \operatorname{var}\frac{X_1 + \dots + X_n}{n}$

$$= \frac{1}{n^2}\operatorname{var} X_i = \frac{\sigma^2}{n}$$

(3) $E\widehat{S}_n = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2$

# Metoda momentů

- $m_r(\vartheta) := \mathbb{E}(X^r)$ pro $X \sim F_\vartheta$ … $r$-tý moment
- $\widehat{m_r(\vartheta)} := \frac{1}{n}\sum_{i=1}^{n} X_i^r$ pro náhodný výběr $X_1, \ldots, X_n$ z $F_\vartheta$ … $r$-tý výběrový moment

## Věta

*$\widehat{m_r(\vartheta)}$ je nestranný konzistentní odhad pro $m_r(\vartheta)$*

- Odhad metodou momentů je řešení soustavy rovnic

$$m_r(\vartheta) = \widehat{m_r(\vartheta)} \qquad r = 1, \ldots, k.$$

# Metoda momentů – příklady

motivace: náhodná fronta, dozvíme se naše pořadí

1) $X_1, \dots, X_n$ n.v. z $Ber(p)$

$m_1(\vartheta) = \mathbb{E}X_1 = p$

$\widehat{m_1(\vartheta)} = \frac{1}{n}(X_1 + \dots + X_n) = \overline{X_n}$

2) $X_1, \dots, X_n$ n. výběr z $U(0, a)$

$m_1(\vartheta) = \mathbb{E}X_1 = \frac{a}{2}$

$\widehat{m_1(\vartheta)} = \overline{X_n}$

$m_1(\vartheta) = m_1(\vartheta)$

$\frac{a}{2} = \overline{X_n}$

$a = 2\overline{X_n}$

statistika $\hat{a} = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} X_i$ je náš odhad par. $a$

# Metoda maximální věrohodnosti (maximal likelyhood, ML)

- náh. výběr $X = (X_1, \ldots, X_n)$ z modelu s parametrem $\vartheta$
- možný výsledek $x = (x_1, \ldots, x_n)$
- … sdružená pravděpodobnostní funkce $p_X(x; \vartheta)$
- … sdružená hustota $f_X(x; \vartheta)$
- *věrohodnost (likelyhood)* $L(x; \vartheta)$ značí $p_X$ nebo $f_X$
- normálně: máme pevné $\vartheta$, a $L(x; \vartheta)$ je funkce $x$
- teď: máme pevné $x$ a $L(x; \vartheta)$ je funkce $\vartheta$

**Metoda MV (ML):**
volíme takové $\vartheta$, pro které je $L(x; \vartheta)$ maximální

# Metoda maximální věrohodnosti (maximal likelyhood, ML)

- **Metoda MV (ML):**
  volíme takové $\vartheta$, pro které je $L(x;\vartheta)$ maximální
- definujeme také $\ell(x;\vartheta) = \log(L(x;\vartheta))$
- díky nezávislosti je

$$L(x;\vartheta) = p_X(x;\vartheta) = p_X(x_1,\dots,x_n;\vartheta) = \prod_{i=1}^{n} p_{X_i}(x_i;\vartheta)$$

$$\ell(x;\vartheta) = \sum_{i=1}^{n} \log p(x_i;\vartheta)$$

max

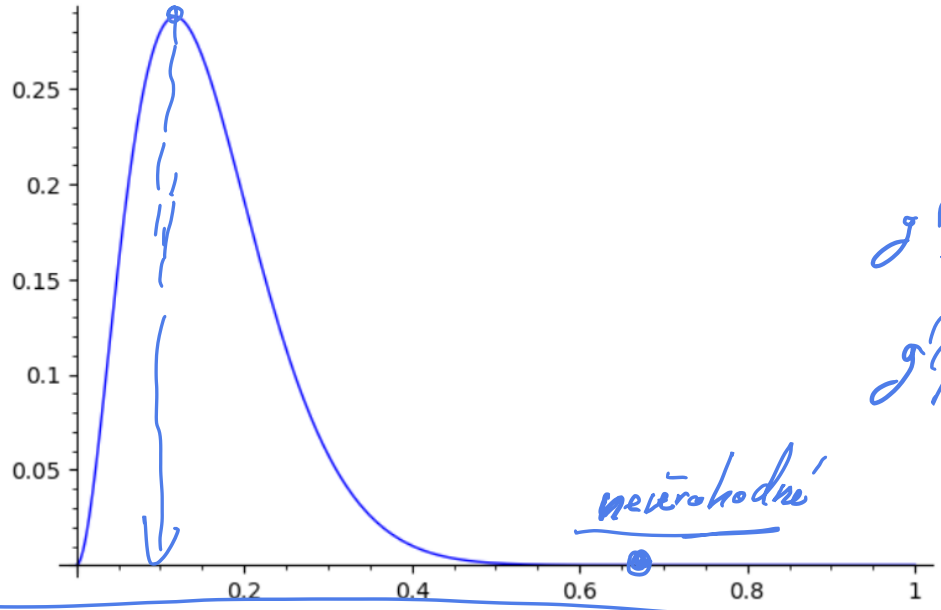
max

# ML – leváci

$n = 17$, 2 leváci

$P_p(\sum X_i = 2) = \binom{17}{2} p^2 (1-p)^{15}$

```
plot(binomial(17,2)*p^2*(1-p)^15, [0,1])
```




nevěrohodné

max. $\ln p^2 (1-p)^{15}$

$g(p) = 2 \ln p + 15 \ln (1-p)$

$g'(p) = \frac{2}{p} - \frac{15}{1-p} = 0$

$\frac{1-p}{p} = \frac{15}{2}$

$\frac{1}{p} - 1 = \frac{15}{2}$

$\frac{1}{p} = \frac{17}{2}$ $\quad p = \frac{2}{17}$

$x_1 \dots x_{17} = 1, 0, 0, 1, 0, \dots$ – 0

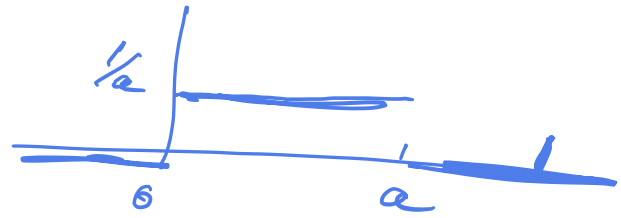
$P(X_1 = x_1) = p$, $P(X_2 = x_2) = 1-p$, ...

$L(x, p) = p(1-p)(1-p) \cdot p (1-p) \cdots$

$\ell(x, p) = \ln p + \ln (1-p) + \dots$

$X_1 \dots X_n \sim U(0, a)$

$x_1 \dots x_n \in \mathbb{R}^+$




$$L(x; a) = f_X(x; a) = \prod_{i=1}^{n} f_{X_i}(x_i; a) = \frac{1}{a^n}$$ pokud $x_i \in \langle 0, a \rangle$

$$= 0$$ jinak

mez. max. věr.: volme $a$ tak, aby $0 \le x_1 \dots x_n \le a$

$a \ge \max(x_1 \dots x_n)$

2) $\frac{1}{a^n}$ co nejv.

$a$ co nejmenší

$\hat{a} = \max(x_1, \dots, x_n)$

# Přehled

# Intervalové odhady

- místo jednoho čísla s nejistým významem vypočítáme z dat interval $[\hat{\Theta}^-, \hat{\Theta}^+]$

## Definice

*Nechť $\hat{\Theta}^-$, $\hat{\Theta}^+$ jsou n.v. které závisí na náhodném výběru $X = (X_1, \ldots, X_n)$. Tyto n.v. určují intervalový odhad, též konfidenční interval o spolehlivosti $1 - \alpha$ ($1 - \alpha$ confidence interval), pokud*

$$P(\hat{\Theta}^- \leq \vartheta \leq \hat{\Theta}^+) \geq 1 - \alpha.$$

# Intervalové odhady normální náhodné veličiny