NMAI059 Pravděpodobnost a statistika 1
9. přednáška

Robert Šámal

Nerovnosti, které známe z minula

- Markov:

$$X \geq 0 \Rightarrow P(X \geq a\mathbb{E}(X)) \leq \frac{1}{a}$$

- Čebyšev/Chebyshev

$$P(|X - \mathbb{E}(X)| \geq a\sigma_X) \leq \frac{1}{a^2}$$

- Chernoff ($\sigma_X = \sqrt{n}$)

$$X = \sum_{i=1}^{n} X_i, X_i = \pm 1 \Rightarrow P(|X - \mathbb{E}(X)| \geq a\sigma_X) \leq 2e^{-a^2/2}$$

Přehled

Slabý zákon velkých čísel (weak law of large numbers)

Věta

Nechť $X_1, \ldots, X_n$ jsou stejně rozdělené n.n.v. se stř. hodnotou μ a rozptylem σ^2. Označme $S_n = (X_1 + \cdots + X_n)/n$. Pak pro každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} P(|S_n - \mu| \geq \varepsilon) = 0.$$

Říkáme, že posloupnost S_n konverguje k μ v pravděpodobnosti (in probability), píšeme $S_n \xrightarrow{P} \mu$).

Dk

- $\mathbb{E}S_n = \mathbb{E}\frac{X_1+\ldots+X_n}{n} = \frac{\mathbb{E}X_1+\ldots+\mathbb{E}X_n}{n} = \frac{\mu+\ldots+\mu}{n} = \mu$
- $\mathrm{var}(S_n) = \mathrm{var}\left(\frac{X_1+\ldots+X_n}{n}\right) = \frac{\mathrm{var}(X_1)+\ldots+\mathrm{var}(X_n)}{n^2} = \frac{\sigma^2+\ldots+\sigma^2}{n^2} = \frac{\sigma^2}{n}$
- $P(|S_n - \mathbb{E}S_n| \geq a\sigma_{S_n}) \leq \frac{1}{a^2} = \frac{1}{\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)^2} = \frac{\sigma^2}{\varepsilon^2 n} \to 0$, $n\to\infty$

$\varepsilon = a\sqrt{\frac{\sigma^2}{n}}$

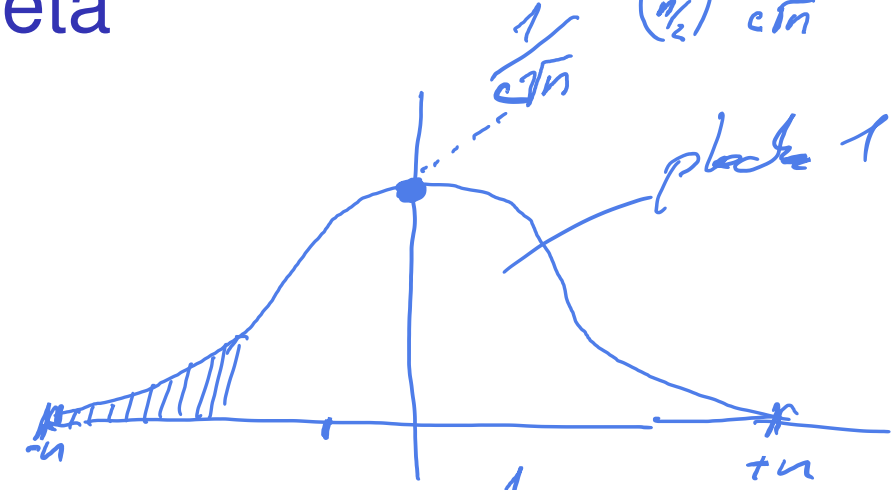
SZVČ → Centrální Limitní věta

$X_i = \pm 1$

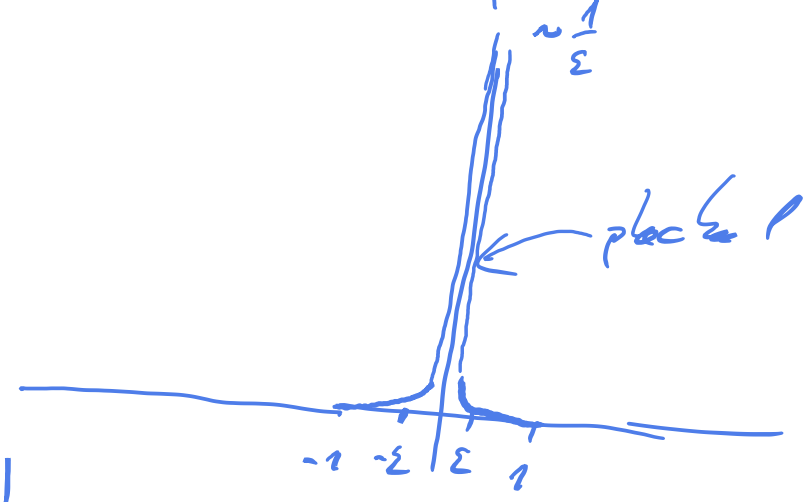
$X = X_1 + \dots + X_n$

$P(X = k) = \frac{\binom{n}{\frac{n+k}{2}}}{2^n}$

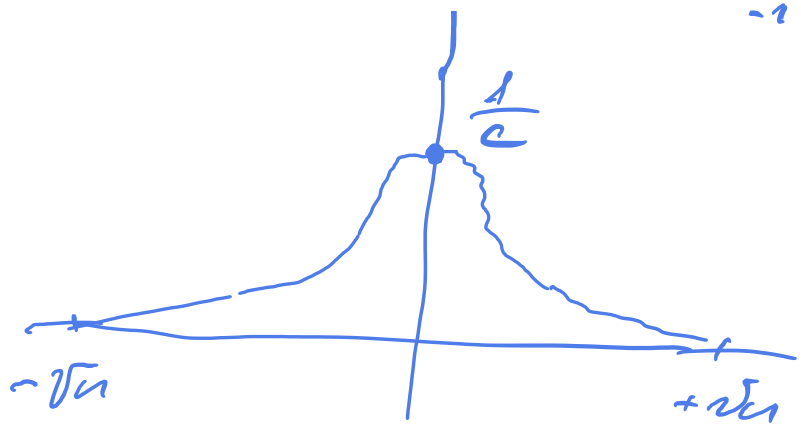
$\binom{n}{\frac{n}{2}} \doteq \frac{2^n}{c\sqrt{n}}$



$S_n = \frac{X}{n}$



$Y_n = \frac{X}{\sqrt{n}}$



Gaussova

Moivre-Laplace

Centrální Limitní věta

Věta

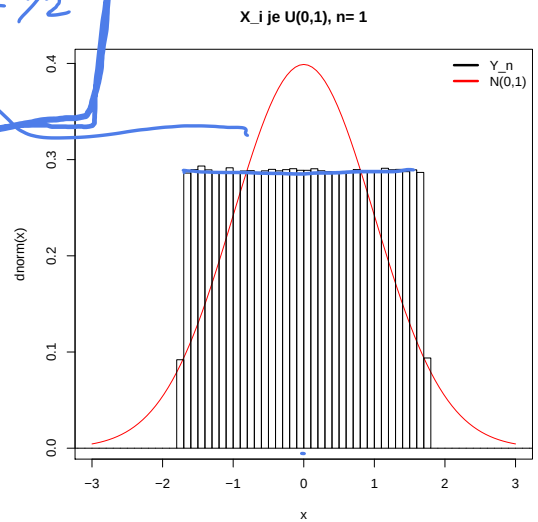
Nechť $X_1, \ldots, X_n$ jsou stejně rozdělené n.n.v. se střední hodnotou μ a rozptylem σ^2. Označme

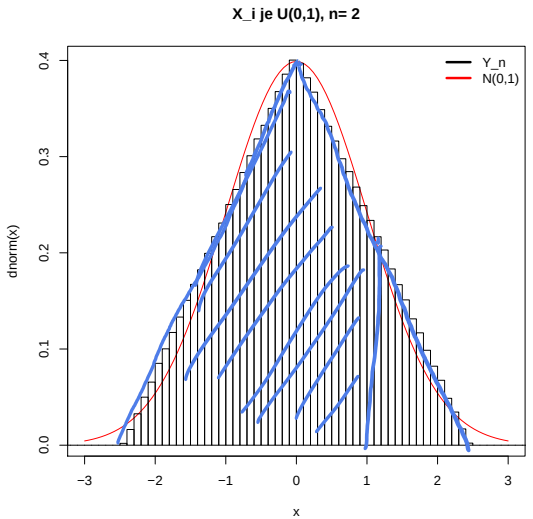
$$Y_n = ((X_1 + \cdots + X_n) - n\mu)/(\sqrt{n} \cdot \sigma).$$

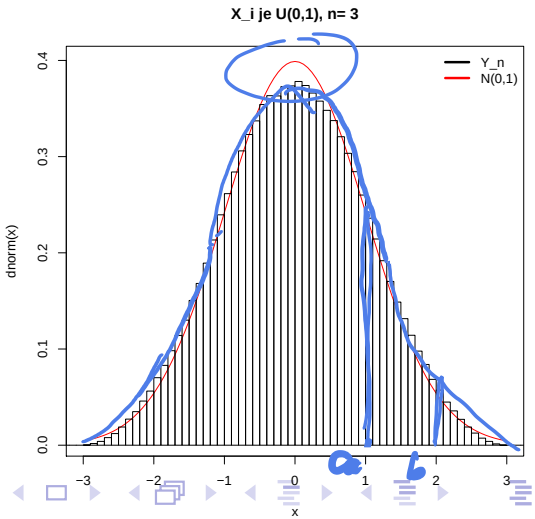
Pak $Y_n \xrightarrow{d} N(0,1)$. Neboli, pokud F_n je distribuční funkce Y_n, tak

$$\lim_{n\to\infty} F_n(x) = \Phi(x) \quad \text{pro každé } x \in \mathbb{R}.$$

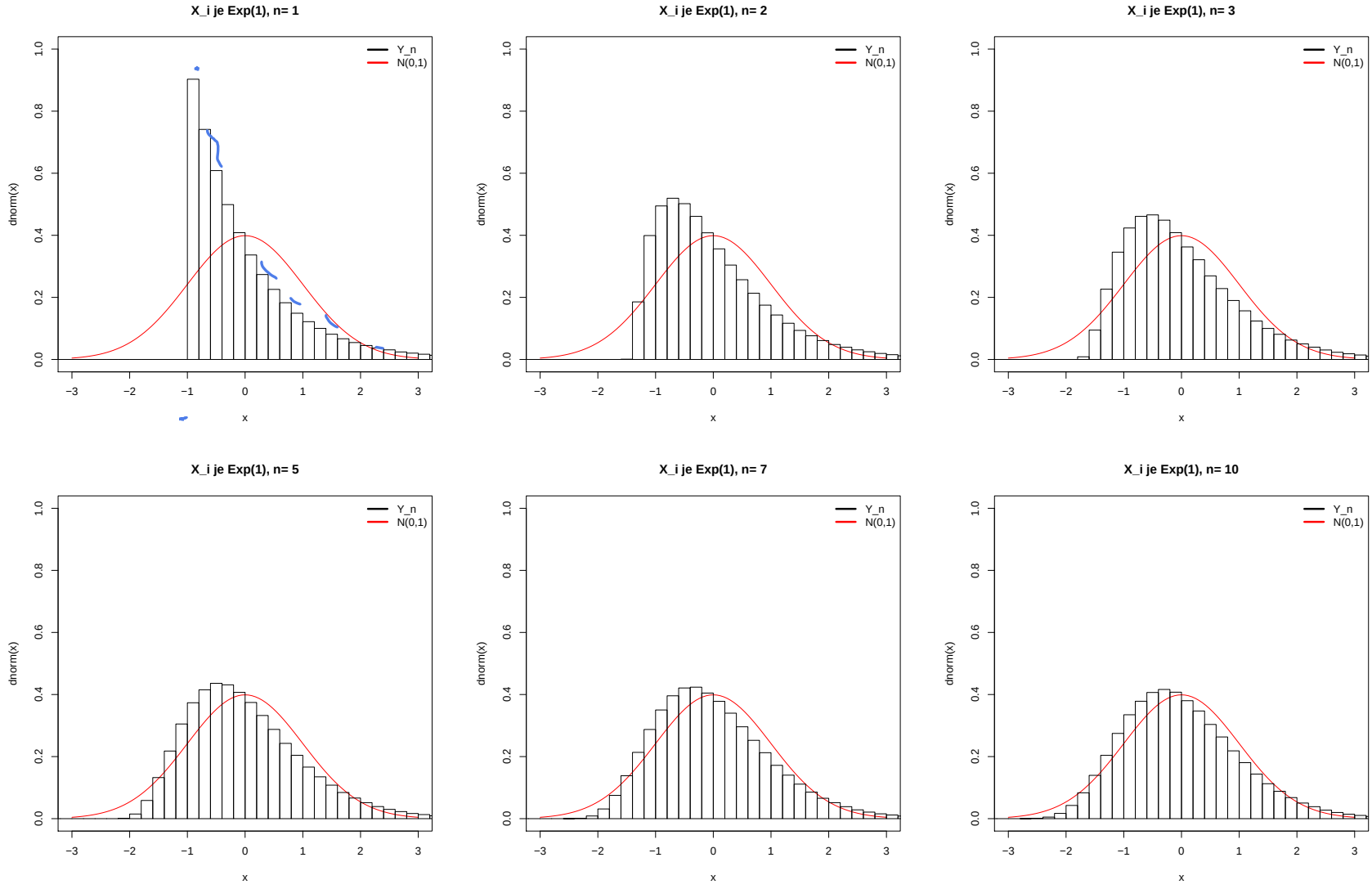
Říkáme, že posloupnost Y_n konverguje k $N(0,1)$ v distribuci (in distribution).







CLV další ukázka



Bonus: Momentová vytvořující funkce

Definice

Pro náhodnou veličinu X označíme

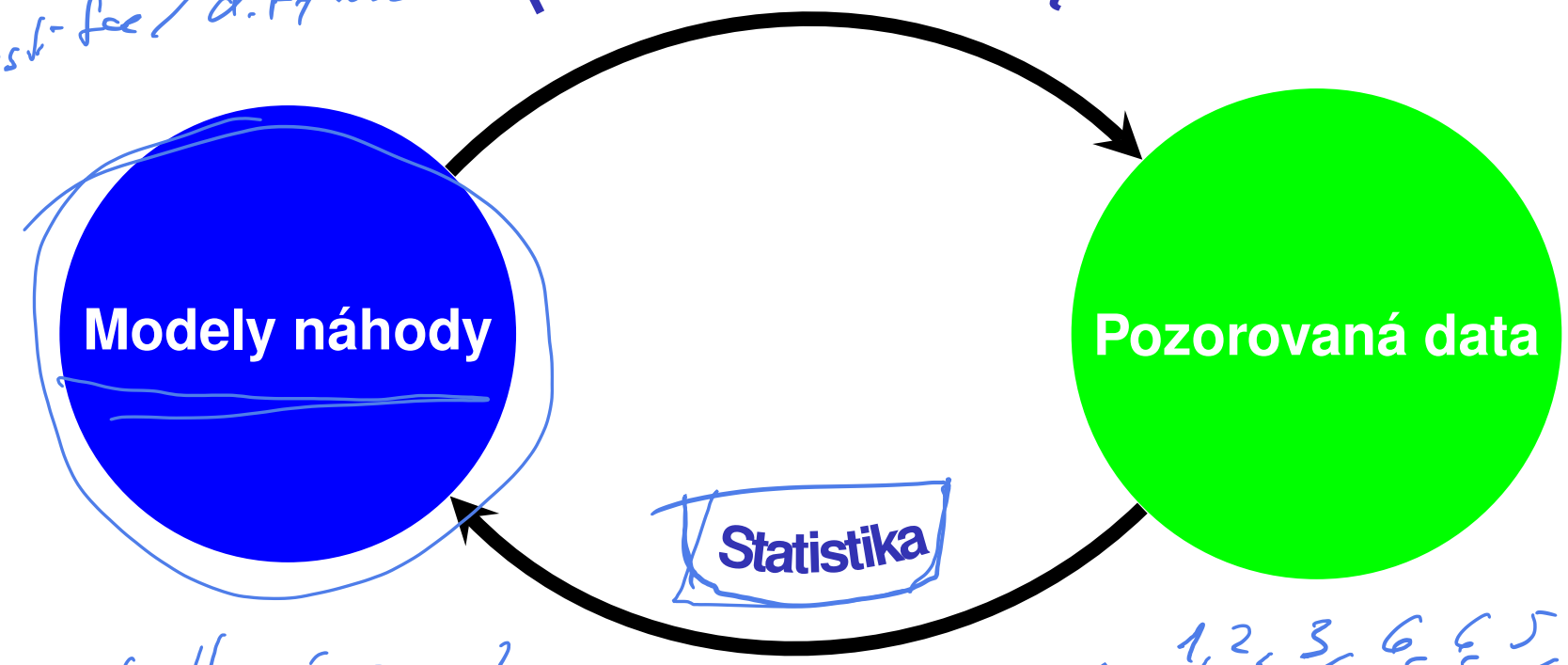
$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

Funkci $M_X(t)$ nazýváme momentová vytvořující funkce (moment generating function).

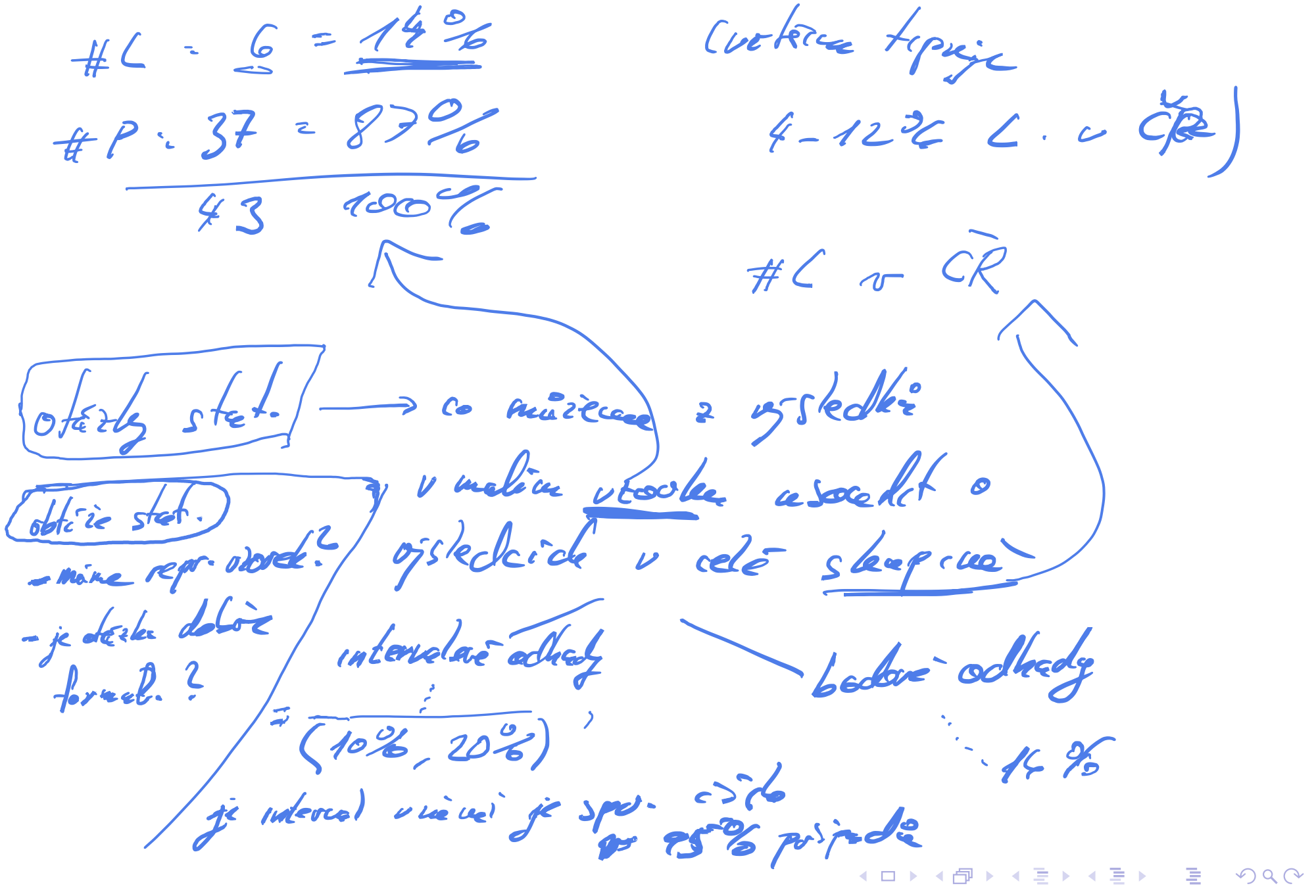
- $M_X(t) = \sum_{n=0}^{\infty} \mathbb{E}\left(X^n\right) \frac{t^n}{n!}$.
- $M_{Bern(p)}(t) = p \cdot e^t + (1-p)$.
- $M_{X+Y}(t) = M_X(t) M_Y(t)$, jsou-li X, Y n.n.v.
- $M_{Bin(n,p)} = (pe^t + 1 - p)^n$
- $M_{N(0,1)} = e^{t^2/2}$
- $M_{Exp(\lambda)} = \frac{1}{1-t/\lambda}$
- Pokud $M_X(t) = M_Y(t)$ na intervalu $(-a, a)$ pro nějaké $a > 0$, tak je $X = Y$ s.j.

Přehled

Plán přednášky



1. ilustrace – počet leváků

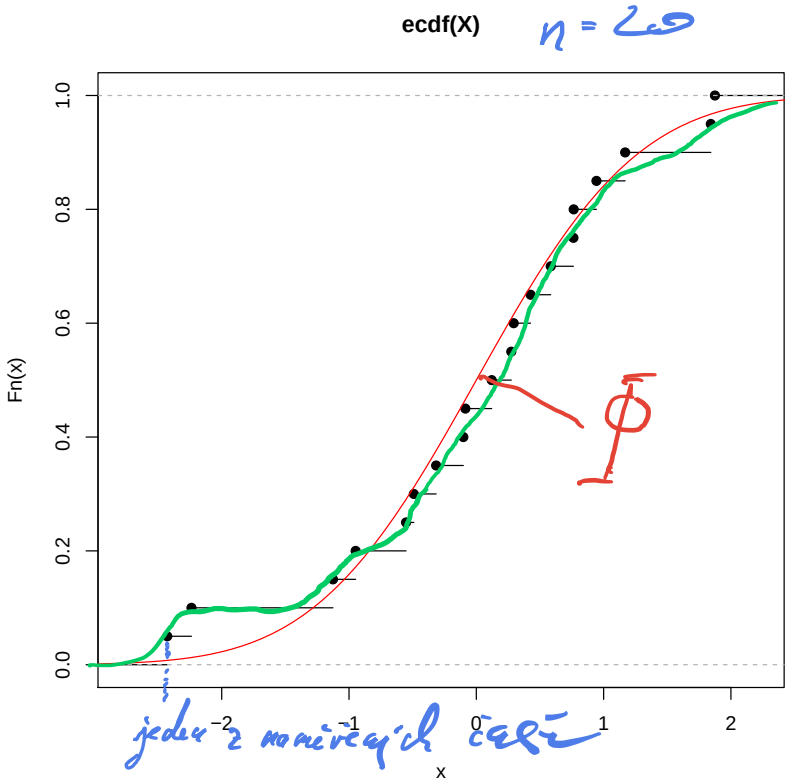


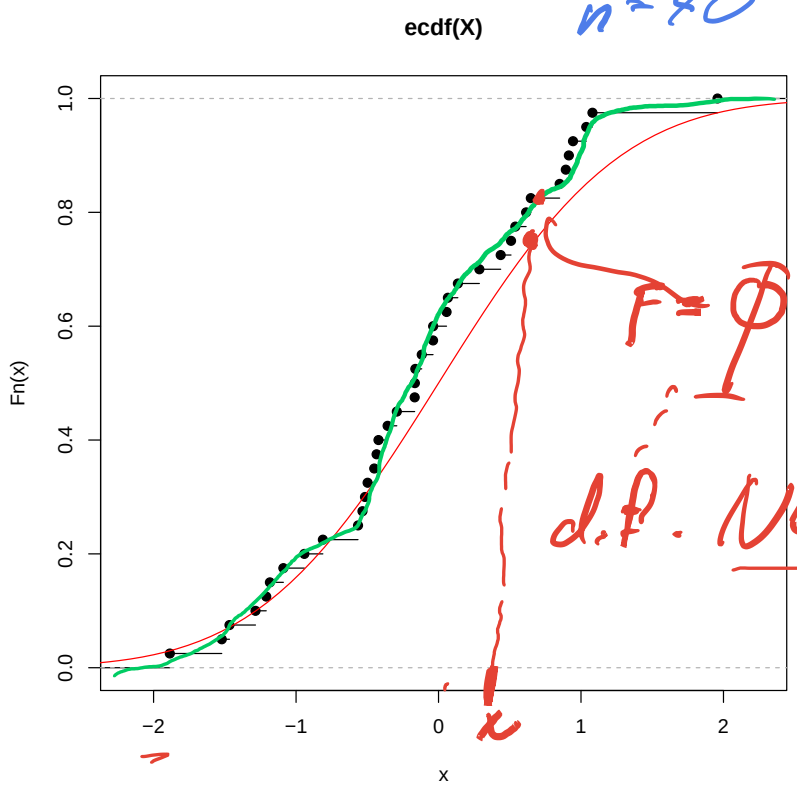
2. ilustrace – doba běhu programu

- $X_1, \ldots, X_n \sim F$ n.n.v., F je jejich distribuční funkce
- **Definice:** *Empirická distribuční funkce (empirical CDF)* je definována

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n},$$

kde $I(X_i \leq x) = 1$ pokud $X_i \leq x$ a 0 jinak.





Empirická distribuční funkce – vlastnosti

$n \cdot \hat{F}_n(x) \sim Bin(n, F(x))$

Věta

Pro pevné x platí

- $\mathbb{E}(\widehat{F}_n(x)) = F(x)$
- $var(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$
- $\widehat{F}_n(x)$ *konverguje k* $F(x)$ *v pravděpodobnosti, píšeme* $\widehat{F}_n(x) \xrightarrow{P} F(x)$.

Důkaz.

Slabý zákon velkých čísel. □

$X_{ni} \rightsquigarrow S_n$

$I(X_i \le x)$

$\Rightarrow \hat{F}_n(x)$

$var X_i' = p(1-p)$

$X_i' \sim Ber(p)$ $p = F(x)$

Tj. $\mathbb{E}\hat{F}_n(x) = \mathbb{E}S_n = \mathbb{E}I(X_i \le x) = P(X_i \le x) = F(x)$

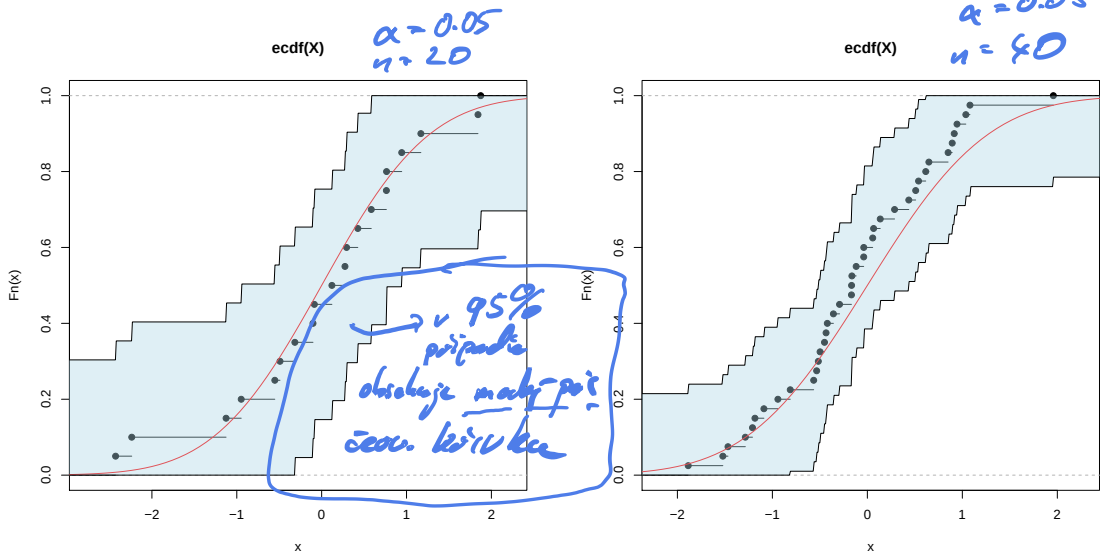
$var(\hat{F}_n(x)) = \frac{var X_i'}{n} = \frac{p(1-p)}{n}$

Empirická distribuční funkce – Dvoretzky-Kiefer-Wolfowitz (DKW)

Věta

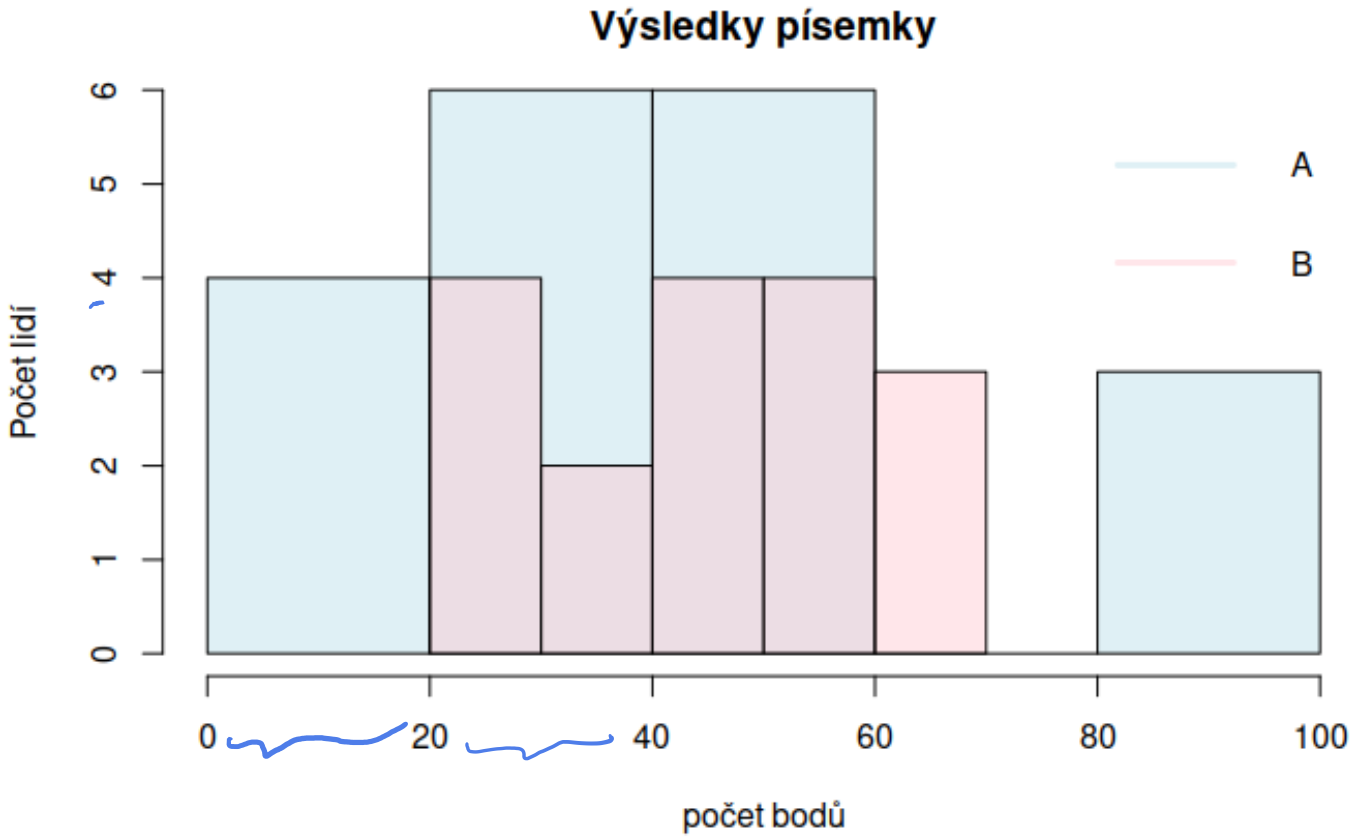
Nechť $X_1, \ldots, X_n \sim F$ *jsou n.n.v.,* $\widehat{F}_n$ *jejich empirická distribuční funkce. Nechť* $\mathbb{E}(X_i)$ *je konečná. Zvolme* $\alpha \in (0,1)$ *a označme* $\varepsilon = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}$. *Pak platí*

$$P\left(\widehat{F}_n(x) - \varepsilon \leq F(x) \leq \widehat{F}_n(x) + \varepsilon\right) \geq 1 - \alpha.$$



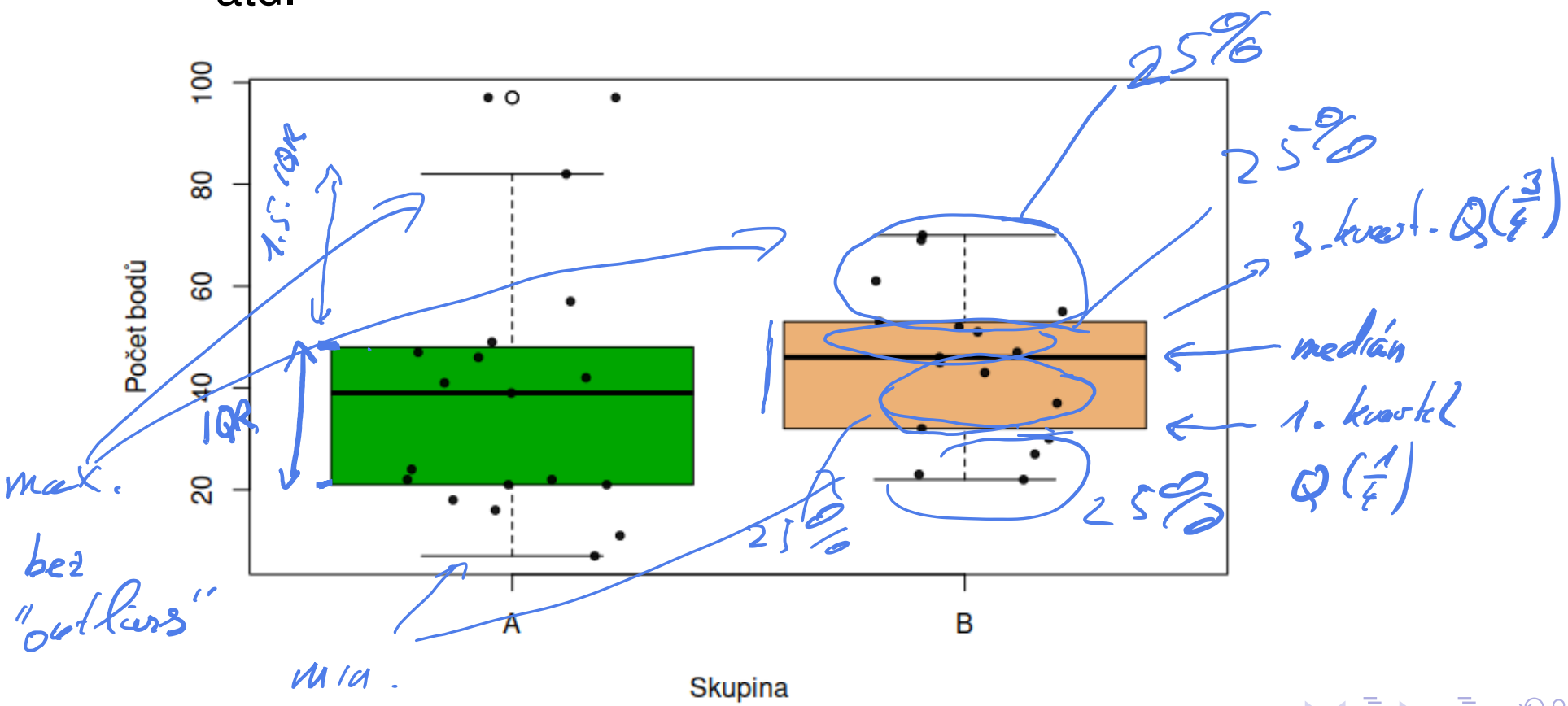
Intro – explorační analýza dat (exploratory data analysis)

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



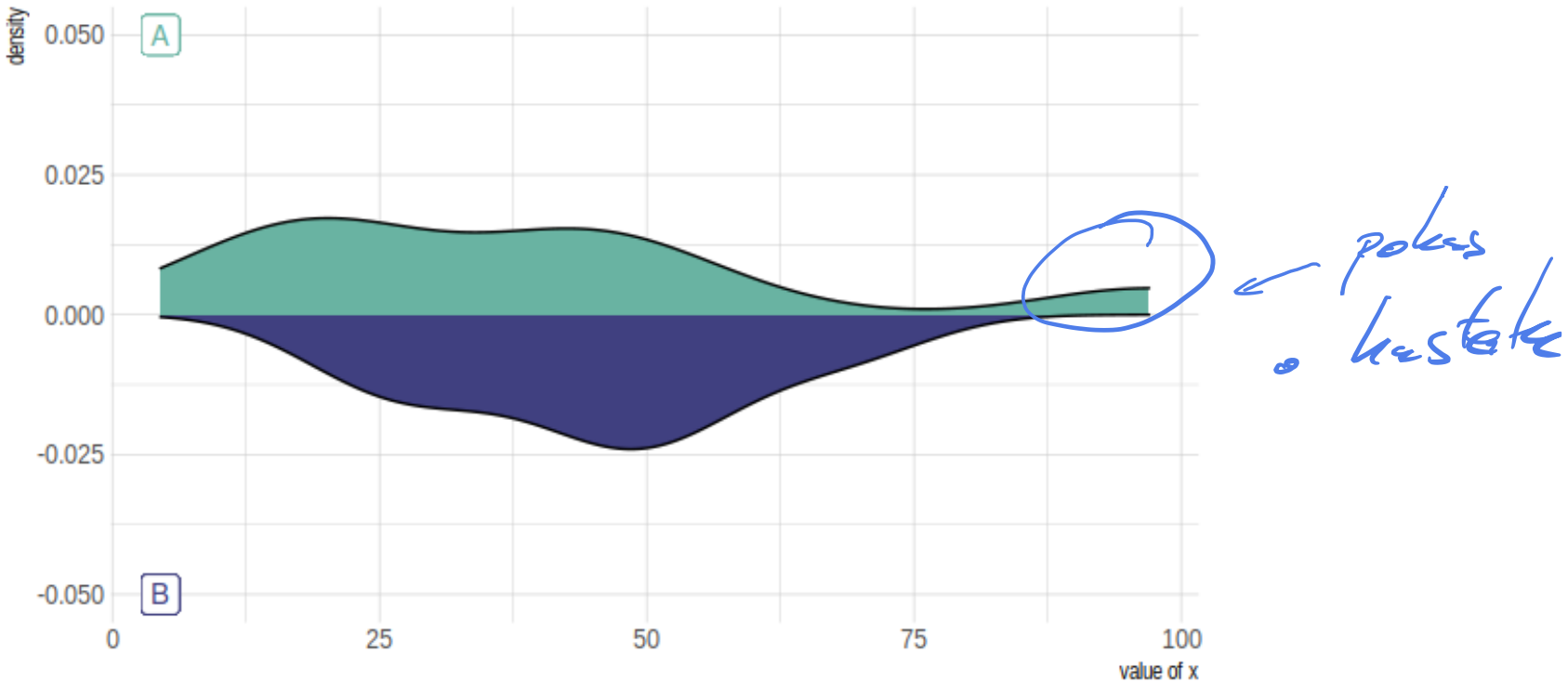
Intro – explorační analýza dat (exploratory data analysis)

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



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$)$.

Statistika – přehled

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

- 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.“
  (George Box)

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. $T = T(X_1, \ldots, X_n)$.

Další typy zkoumaných problémů

- Je zkoumaný lék účinný?
- Je naše mince, kostka spravedlivá?
- Předpokládáme, že výška člověka je normálně rozdělená. Jaké je μ, σ?
- Předpokládáme, že výška člověka je normálně rozdělená. V jakém vztahu je průměrná výška mužů a žen? Praváků a leváků?
- Jak závisí náklon šikmé věže v Pise na čase?

Zkoumané úlohy – předpoklady

- Vždy předpokládáme, že máme nezávislá měření – hodnoty n.n.v. $X_1, \ldots, X_n \sim F$
- O F předpokádáme, že patří do nějakého *modelu* – množiny vhodných distr. funkcí.
- parametrické/neparametrické modely

Zkoumané úlohy – cíle

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

Přehled

Výběrový průměr a rozptyl

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

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

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

Cíle

Definice

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

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

Věta

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

Parametry výběrového momentu a rozptylu

Věta

1. $\bar{X}_n$ *je konzistentní nestranný odhad* μ
2. $\bar{S}_n$ *je konzistentní asymptoticky nestranný odhad* μ
3. $\widehat{S}_n$ *je konzistentní nestranný odhad* μ

Metoda momentů

- $m_r(\vartheta) := \mathbb{E}\left(X^r\right)$ 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_ϑ … 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

Metoda maximální věrohodnosti (maximal likelihood, ML)

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

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

Metoda maximální věrohodnosti (maximal likelihood, ML)

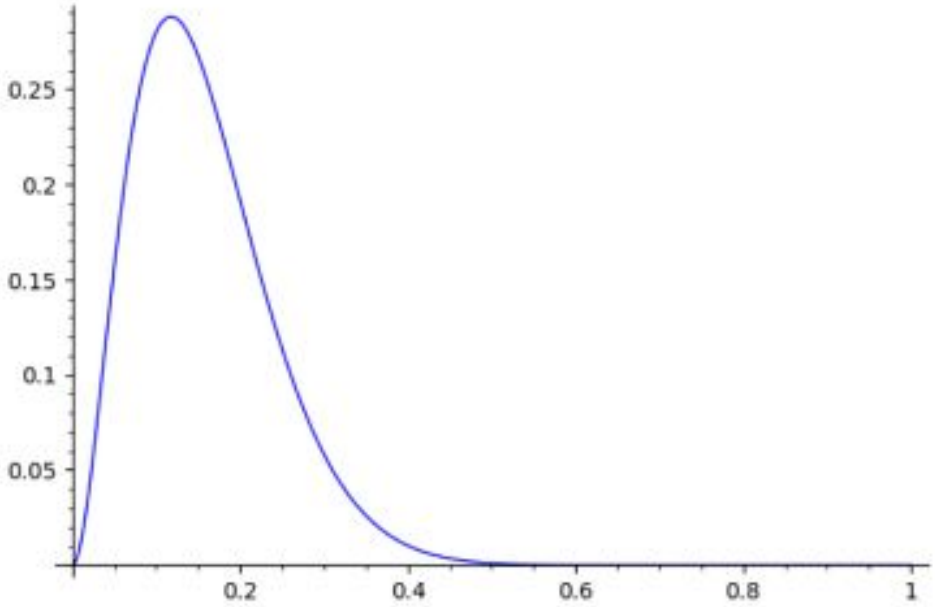
- **Metoda MV (ML):**
  volíme takové ϑ, pro které je $L(x;\vartheta)$ maximální
- definujeme také $\ell(x;\vartheta) = \log(L(x;\vartheta))$
- díky nezávislosti je

$L(x;\vartheta) =$

$\ell(x;\vartheta) =$

ML – leváci

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



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