NMAI059 Pravděpodobnost a statistika 1
8. přednáška

Robert Šámal

Přehled

Ještě k limitním větám

Statistika – úvod

Slabý zákon velkých čísel (weak law of large numbers)

Věta

Nechť $X_1, \ldots, X_n$ jsou 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

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

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

Centrální Limitní věta

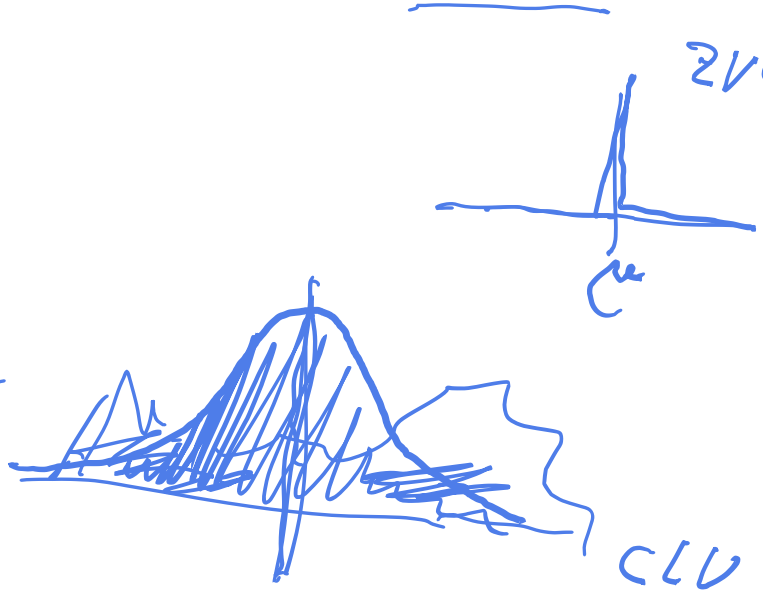
Věta

Nechť $X_1, \ldots, X_n$ jsou n.n.v. se stř. hodnotou μ a rozptylem σ^2. Označme $Y_n = ((X_1 + \cdots + X_n) - n\mu)/\sqrt{n}$.

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

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

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



Momentová vytvořující funkce

Definice

Pro náhodnou veličinu X označíme

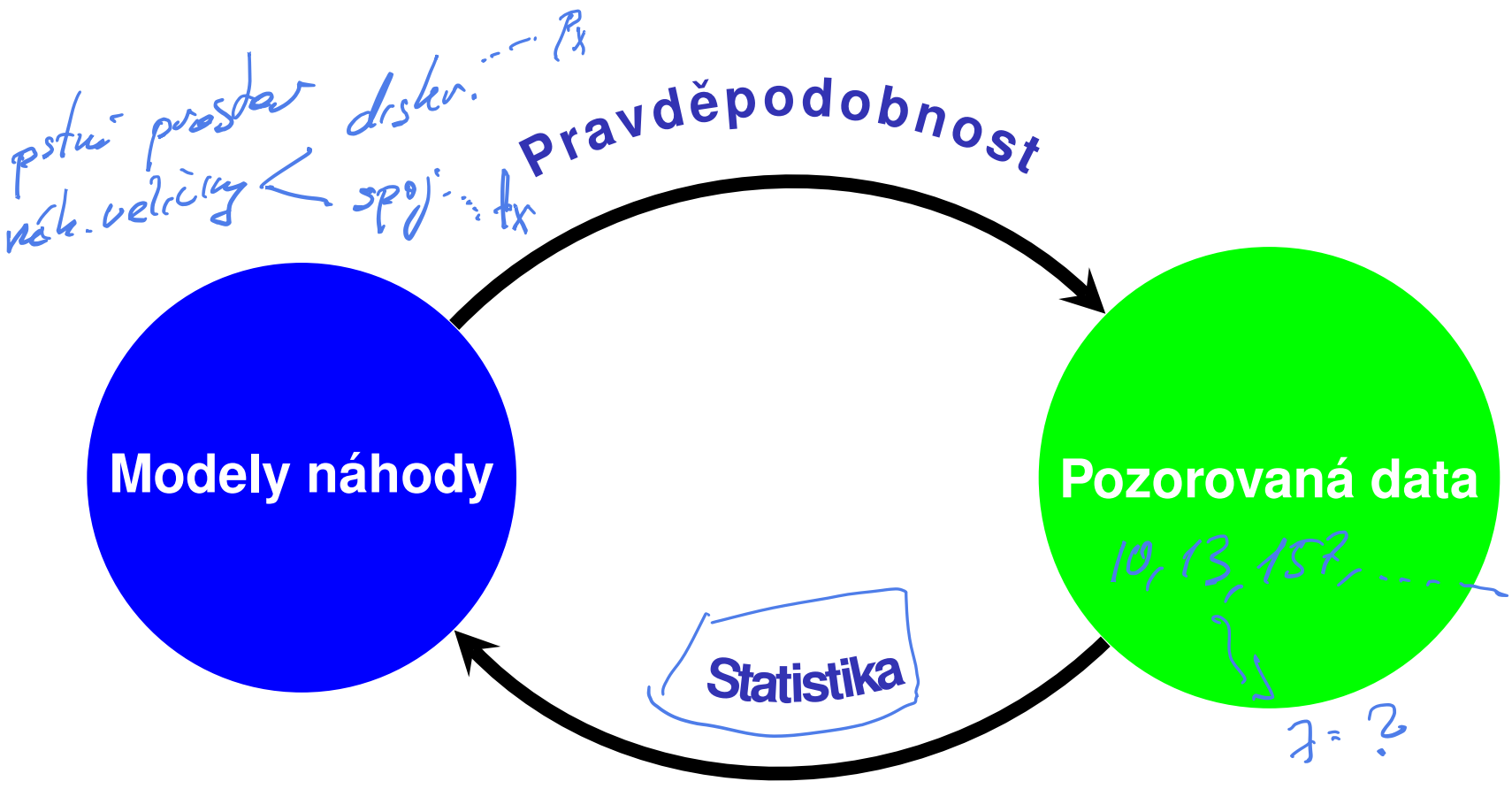
$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

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

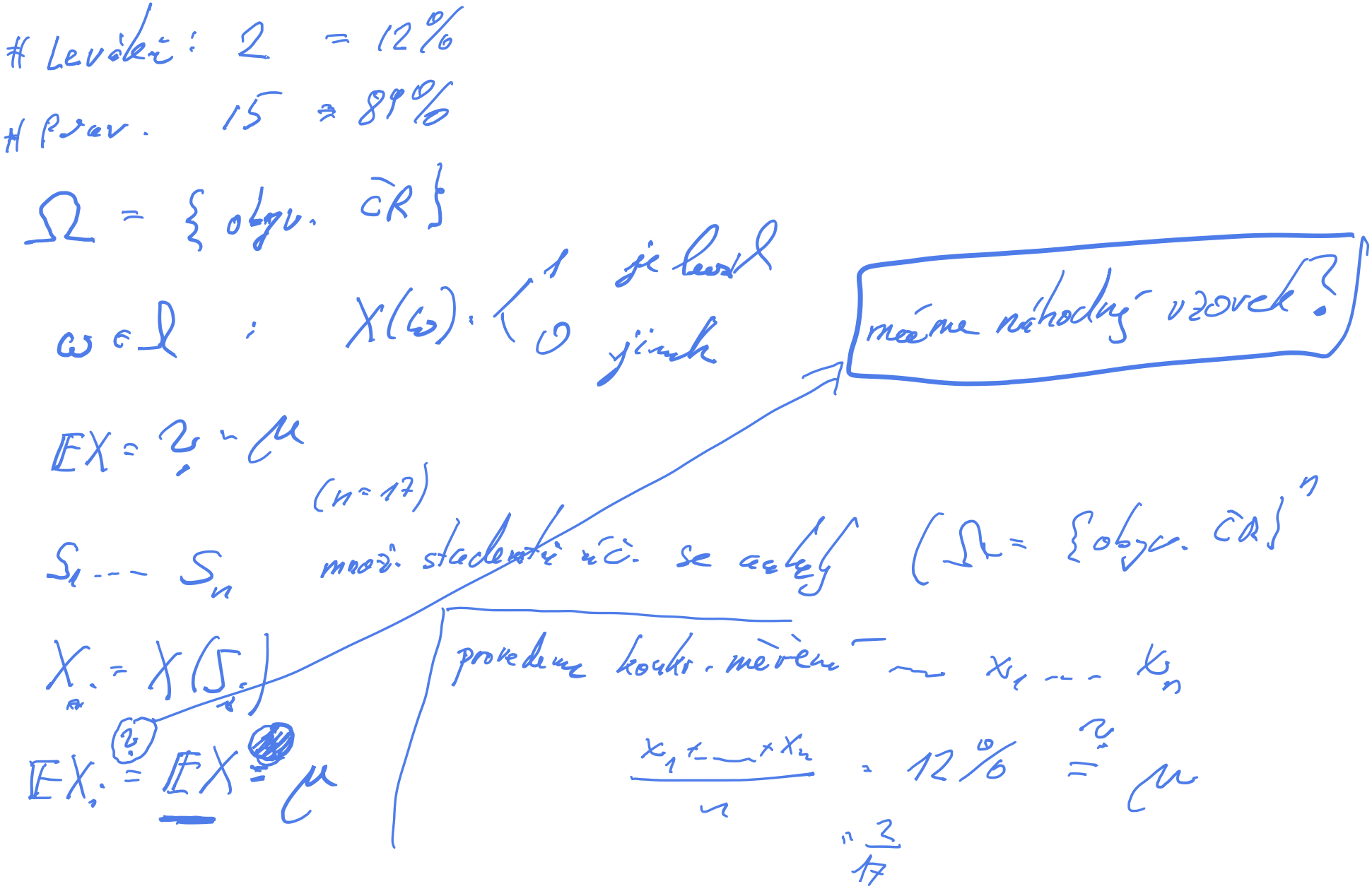
- $M_{Bern(p)}(t) = p \cdot e^t + (1-p)$.
- $M_X(t) = \sum_{n=0}^{\infty} \mathbb{E}\left(X^n\right) \frac{t^n}{n!}$.
- $M_{X+Y}(t) = M_X(t) M_Y(t)$, jsou-li X, Y n.n.v.
- $M_{Bin(n,p)} = (pe^t + 1 - p)^n$
- $M_{N(0,1)} = e^{t^2/2}$
- $M_{Exp(\lambda)} = \frac{1}{1 - t/\lambda}$
- Pokud $M_X(t) = M_Y(t)$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

Přehled

Plán přednášky



1. ilustrace – počet leváků



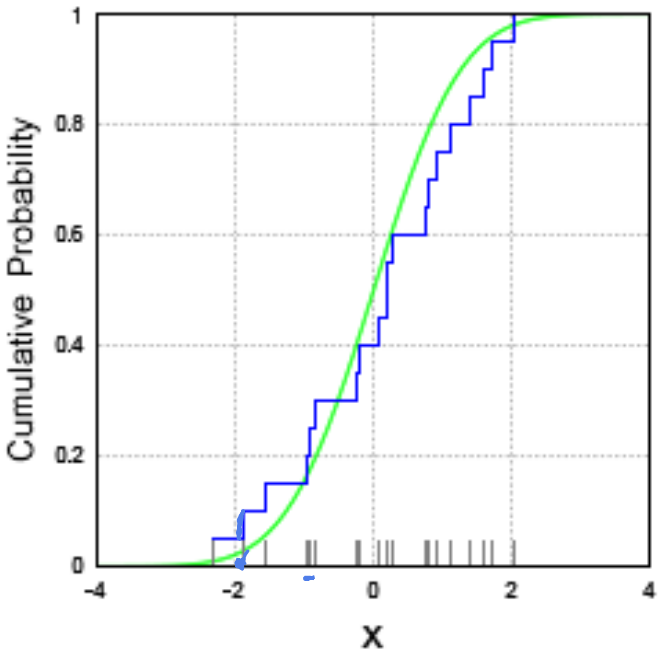
2. ilustrace – doba běhu programu

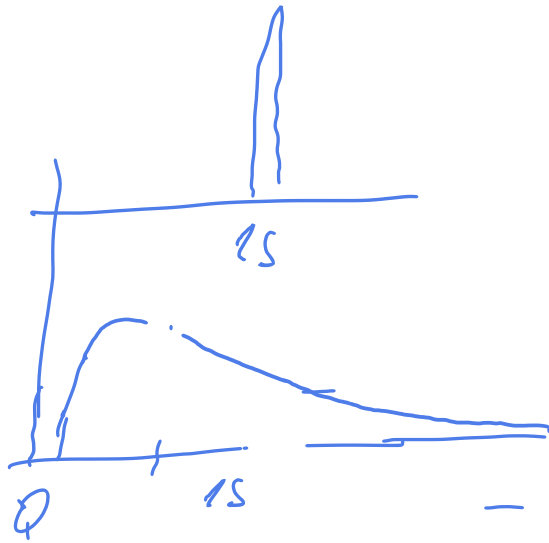
- $X_1, \ldots, X_n \sim F$ n.n.v., F je jejich distribuční funkce
- **Definice:** *Empirická distribuční funkce (empirical CDF)* je definována

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

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

(Obrázek vytvořil wiki-editor nagualdesign.)





Empirická distribuční funkce – vlastnosti

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. □

$Y_1 = I(X_1 \le x)$

$Y_n = I(X_n \le x)$

$S_n = \bar{Y}_n = \frac{Y_1 + \dots + Y_n}{n} = \widehat{F}_n(x)$

$\mathbb{E} S_n = \mathbb{E} Y_i = F(x) = \mu$

$Y_i \sim Bern(\mu)$

$var\, Y_i = \mu(1-\mu)$

$var\, S_n = \frac{\mu(1-\mu)}{n}$

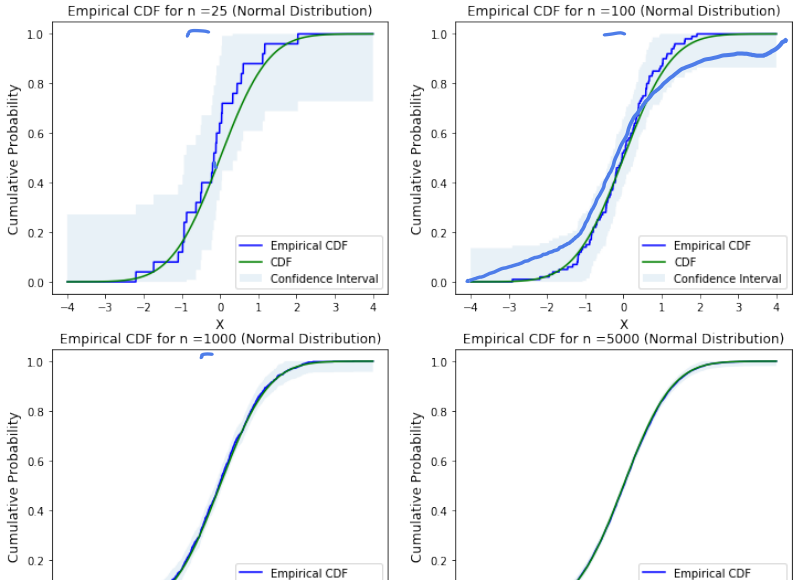
Empirická distribuční funkce – Dvoretzky-Kiefer-Wolfowitz (DKW)

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

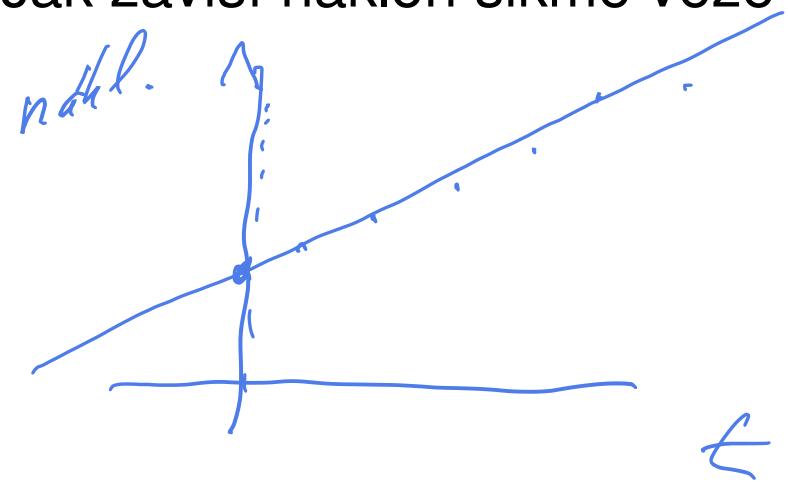
$$P(\widehat{F}_n(x) - \varepsilon \leq F(x) \leq \widehat{F}_n(x) + \varepsilon) \geq 1 - \alpha.$$

(Obrázek vytvořil wiki-editor Sivaji12331, $\alpha = 0.05$.)



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

$F \in \{$ všechny distr. fce $\}$

$F \in \{ -\!\!''\!\!- $ s konečnou stř. hodn. $\}$

$\{F_\vartheta, \vartheta \in \Theta\}$ — parametr, množ.

$F \in \{$ d.f. $N(\mu, \sigma^2)\}$

$\vartheta = (\mu, \sigma) \in \mathbb{R} \times (0, \infty)$

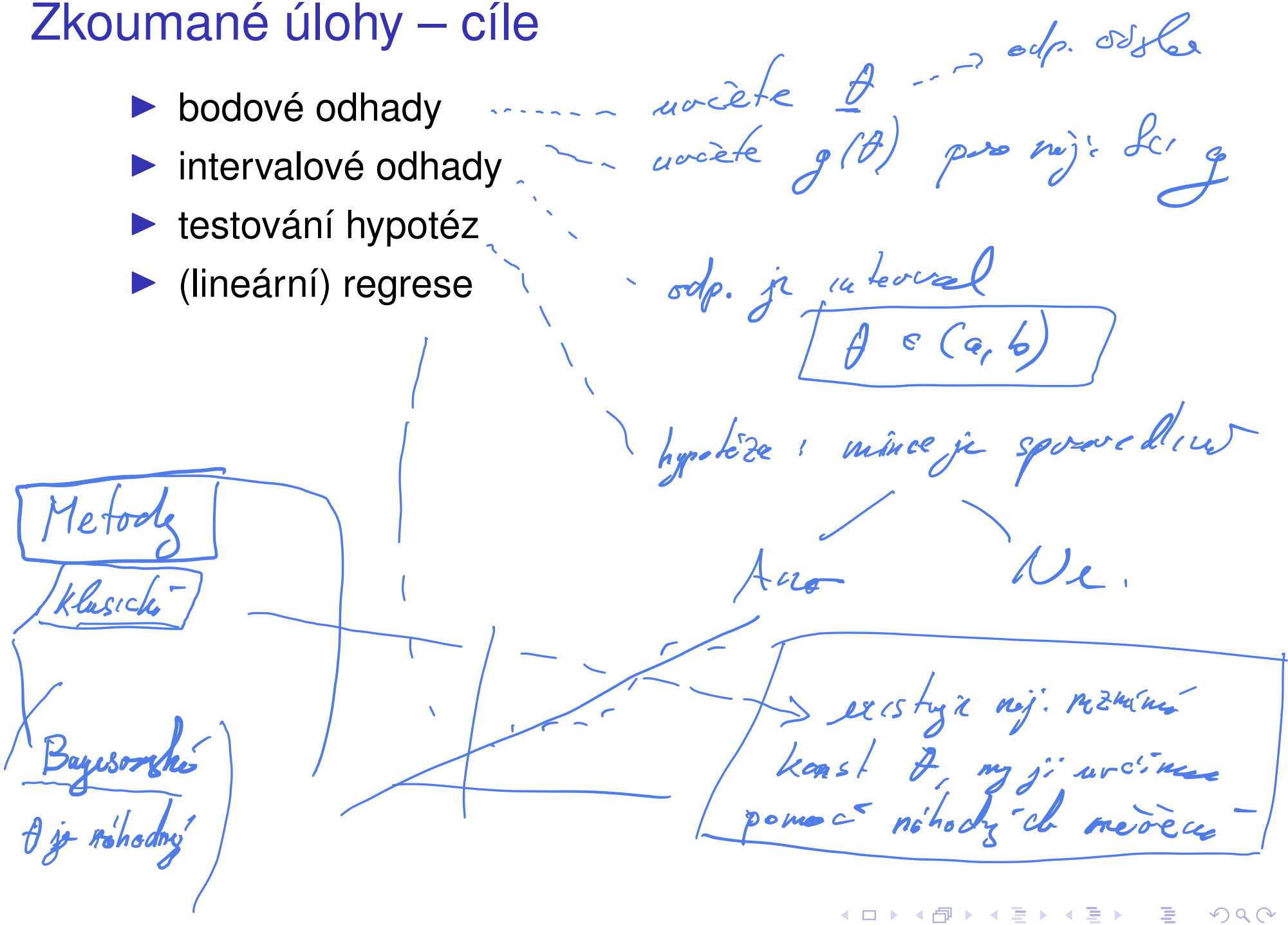
$F \in \{$ d.f. $Exp(\lambda)\}$

$\vartheta = \lambda \in \mathbb{R}^+$ $(0, \infty)$

$\{ Pois(\lambda) \}$

Zkoumané úlohy – cíle

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



Bodové odhady a jejich vlastnosti

pozorování $X = (X_1, \dots, X_n)$

X_i n.v. s d.f. F_ϑ, $\vartheta \in \Theta$

$\hat{\Theta}_n$ je **konzistentní** (consistent) $\Leftrightarrow$ $\hat{\Theta}_n \xrightarrow{P} \vartheta$ $\forall \vartheta$

odhad je n.v. $\hat{\Theta}_n = g(X)$ pro vhodnou g

(**estimator**) vs. estimate

závisí na ϑ

velké hodnoty, malé hodnoty

chyba odhadu (estimation error) $\tilde{\Theta}_n := \hat{\Theta}_n - \vartheta$

odhad, skutečnost

n.v. "dopadne pokaždé jinak"

bias $b_\vartheta(\hat{\Theta}_n) = \mathbb{E}\tilde{\Theta}_n = \mathbb{E}\hat{\Theta}_n - \vartheta$

$\hat{\Theta}_n$ je **nezkreslený** (unbiased) $\Leftrightarrow b_\vartheta(\hat{\theta}) = 0$ $\forall \vartheta$

pro pevné ϑ konstanta

asympt. nezkreslený $\Leftrightarrow b_\vartheta(\hat{\theta}) \to 0$ $\forall \vartheta$

Př. $X_1, \dots, X_n \sim Be(\vartheta)$

odhad č. 1 $\frac{1}{2}$

odhad č. 2 $\frac{X_1 + \dots + X_n}{n} = \bar{X}_n$

odhad č. 3 $X_1 = \overline{X_1}$

$\hat{\vartheta}_n$

konzistence

$\forall \varepsilon > 0$

$P(|\vartheta_n - \vartheta| > \varepsilon) \to 0$

$E\hat{\vartheta}_n =$ 1) $\frac{1}{2}$ $\quad bias = \frac{1}{2} - \mu$

2) μ } $bias = 0$

3) μ

..... je konzistentní – ZVČ

..... $\vartheta_n = X_1 = \begin{cases} 0 \\ 1 \end{cases}$

Z 2. a 3. jsou nezaujaté

takže pokud $0 < \mu < 1$

$\varepsilon < \min(\mu, 1-\mu)$

$P(|\vartheta - \vartheta| > \varepsilon) = 1$