

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

nezáleží!

Nechť X_1, \dots, X_n jsou n.n.v. se stř. hodnotou μ a rozptylem σ^2 .

Označme $S_n = \underbrace{(X_1 + \dots + X_n)/n}$. Pak pro každé $\varepsilon > 0$ platí

sample mean
 \bar{X}_n

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

*↑ nezávislé
↑ skutečnost*

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

Centrální Limitní věta

$$Y_n = \frac{(X_1 - \mu) + (X_2 - \mu) + \dots}{\sqrt{n}}$$

Věta

Nechť X_1, \dots, X_n jsou n.n.v. se stř. hodnotou μ a rozptylem σ^2 .

Označme $Y_n = ((X_1 + \dots + X_n) - n\mu) / \sqrt{n}$.

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

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

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

$$E X_i = \mu$$

$$E (X_i - \mu) = 0$$

$$\text{var } Y_n = \sigma^2$$

... centrovaní



Momentová vytvořující funkce

Definice

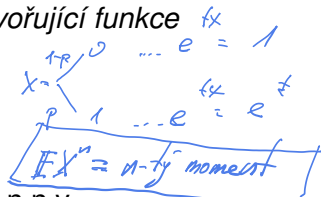
Pro náhodnou veličinu X označíme

$$\begin{aligned} E e^{t(X+t)} &= E e^{tX} \cdot E e^{t} \\ &= E e^{tX} \cdot E e^{t} \end{aligned}$$

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

$$\begin{aligned} &= E \sum_{n=0}^{\infty} \frac{(tX)^n}{n!} \\ &= E \sum_{n=0}^{\infty} X^n \cdot \frac{t^n}{n!} \end{aligned}$$

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



▶ $M_{\text{Bern}(p)}(t) = p \cdot e^t + (1 - p)$

▶ $M_X(t) = \sum_{n=0}^{\infty} \mathbb{E}(X^n) \frac{t^n}{n!}$

▶ $M_{X+Y}(t) = M_X(t)M_Y(t)$, jsou-li X, Y n.n.v.

▶ $M_{\text{Bin}(n,p)} = (pe^t + 1 - p)^n$

▶ $M_{N(0,1)} = e^{t^2/2}$

▶ $M_{\text{Exp}(\lambda)} = \frac{1}{1-t/\lambda}$

▶ Pokud $M_X(t) = M_Y(t)$ na intervalu $(-a, a)$ pro nějaké $a > 0$, tak je $X = Y$ s.j.

Přehled

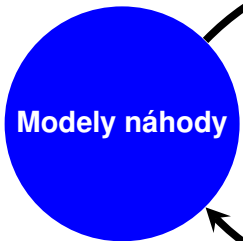
Ještě k limitním větám

Statistika – úvod

Plán přednášky $X \sim \text{Pois}(1)$ #emvili

postup - prostor diskre. $\dots P_X$
 náh. veličiny \leftarrow spoj. $\dots f_X$

Pravděpodobnost



diskre. \uparrow spoj. \uparrow
 $X + Y$

$X = \text{unif}\{1, 2, \dots, 6\}$
 $Y \sim N(0, 1)$



1. ilustrace – počet leváků

Leváci: 2 = 12%

Prav. 15 = 88%

$$\Omega = \{ \text{obgv. } \bar{c}R \}$$

$\omega \in \Omega$: $X(\omega) = \begin{cases} 1 & \text{je levák} \\ 0 & \text{jinak} \end{cases}$

máme náhodný vzorek?

$$EX = \bar{x} = \mu$$

($n=17$)

S_1, \dots, S_n moži. studentů k č. se učily ($\Omega = \{ \text{obgv. } \bar{c}R \}$)

$$X_i = X(S_i)$$

$$EX_i = EX = \mu$$

provedeme kouř. měření x_1, \dots, x_n

$$\frac{x_1 + \dots + x_n}{n} = 12\% = \bar{x} = \mu = \frac{2}{17}$$

2. ilustrace – doba běhu programu

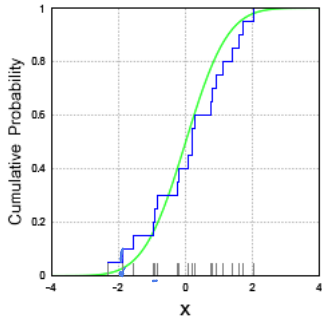
- ▶ $X_1, \dots, X_n \sim F$ n.n.v., F je jejich distribuční funkce
- ▶ **Definice:** *Empirická distribuční funkce (empirical CDF)* je definována

indik

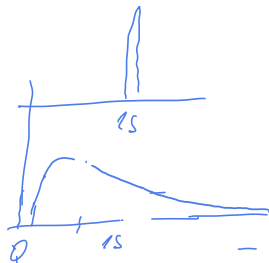
$$\hat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n},$$

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

(Obrázek vytvořil wiki-editor nagualdesign.)



$F = \phi$



Empirická distribuční funkce – vlastnosti

Věta

Pro pevné x platí

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

Důkaz.

Slabý zákon velkých čísel.

$$\left. \begin{aligned} Y_1 &= I(X_1 \leq x) \\ &\vdots \\ Y_n &= I(X_n \leq x) \end{aligned} \right\}$$
$$S_n = \overline{Y}_n = \frac{Y_1 + \dots + Y_n}{n} = \widehat{F}_n(x)$$

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

$$Y_i \sim \text{Ber}(\mu)$$

$$\text{var } Y_i = \mu(1-\mu)$$

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



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

Věta

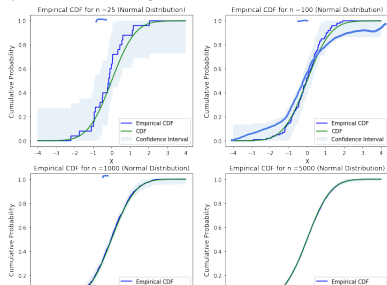
Nechť $X_1, \dots, X_n \sim F$ jsou n.n.v., \hat{F}_n jejich empirická distribuční funkce. Nechť $\mathbb{E}(X_i)$ je konečná. Zvolme $\alpha \in (0, 1)$ a označme

$$\varepsilon = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}. \text{ Pak platí}$$

$$\varepsilon = \frac{\rho(\alpha)}{\sqrt{n}} \quad \leftarrow$$

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

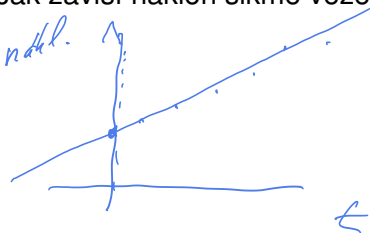
(Obrázek vytvořil wiki-editor Sivaji12331, $\alpha = 0.05$.)



→ $F = \Phi$
 poz. 1 ... ε klesá vs \sqrt{n}
 poz. 2 ... $f = F'$ je citlivější na malé změny v F

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. $X_1, \dots, X_n \sim F$
- ▶ O F předpokládáme, že patří do nějakého *modelu* – množiny vhodných distr. funkcí.
- ▶ parametrické/neparametrické modely

$F \in \{\text{všechny distr. funkce}\}$

$F \in \{\text{---}\}$
↳ konečnou str. hodn.

$\{F_\theta, \theta \in \Theta\}$
↑ parametr množ.

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

$$\theta = \lambda \in \mathbb{R}^+$$

 $(0, \infty)$

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

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

$$\{\text{Pois}(\lambda)\}$$

Zkoumané úlohy – cíle

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

určíte θ \rightarrow odp. odhady
určíte $g(\theta)$ pro něj: h, g

odp. je interval

$$\theta \in (a, b)$$

hypotéze: mince je spravedlivá

Ano

Ne.

existuje nej. možná
konst. θ , my ji určíme
pomocí náhodných měření

Metody

klasické

Bayesovské

θ je náhodný

Bodové odhady a jejich vlastnosti

pozorování $X = (X_1, \dots, X_n)$

X_i n.n.v. s d.f. F

$\hat{\theta}_n$ je konzistentní (consistent)
 $\hat{\theta}_n \xrightarrow{P} \vartheta$

odhad je n.v. $\hat{\theta} = g(X)$
 (estimator) vs. estimate
 $\hat{\theta}_n$ závisí na ϑ

pro funkce g

velké theta
 malé theta

chyba odhadu (estimation error)

$$\tilde{\theta}_n := \hat{\theta}_n - \vartheta$$

odhad \uparrow skutečnost

$$\text{bias } b_{\vartheta}(\hat{\theta}_n) = E \tilde{\theta}_n = E \hat{\theta}_n - \vartheta$$

n.v.
 "dopadne pokrývají
 "gicok"

$\hat{\theta}_n$ je nezažijí (unbiased) $\Leftrightarrow b_{\vartheta}(\hat{\theta}) = 0$
 a asympt. nezažijí $\Leftrightarrow b_{\vartheta}(\hat{\theta}) \rightarrow 0$

pro ϑ je
 konstanta
 $\neq 0$

$\boxed{\text{Pr}} \quad X_1, \dots, X_n \sim \text{Ber}(\vartheta)$

odhad č.1 $\frac{1}{2}$

odhad č.2

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



konvergenz $\forall \varepsilon > 0$
 $P(|\hat{\theta}_n - \vartheta| > \varepsilon) \rightarrow 0$

odhad č.3

$X_1 = \bar{X}_1$

$E \hat{\theta}_n = \begin{cases} 1) \frac{1}{2} \\ 2) \mu \\ 3) \mu \end{cases} \left. \begin{matrix} \text{bias} = \frac{1}{2} - \mu \\ \text{bias} = 0 \end{matrix} \right\}$

----- je konzistentní ZVC

----- $\hat{\theta}_n = X_1 = \begin{cases} 0 \\ 1 \end{cases}$

2. a 3. jsou nezávislé

takže pokud $0 < \mu < 1$
 $\varepsilon < \min(\mu, 1-\mu)$

$P(|\hat{\theta}_n - \vartheta| > \varepsilon) = 1$