

NMAI059 Pravděpodobnost a statistika 1

11. přednáška

Robert Šámal

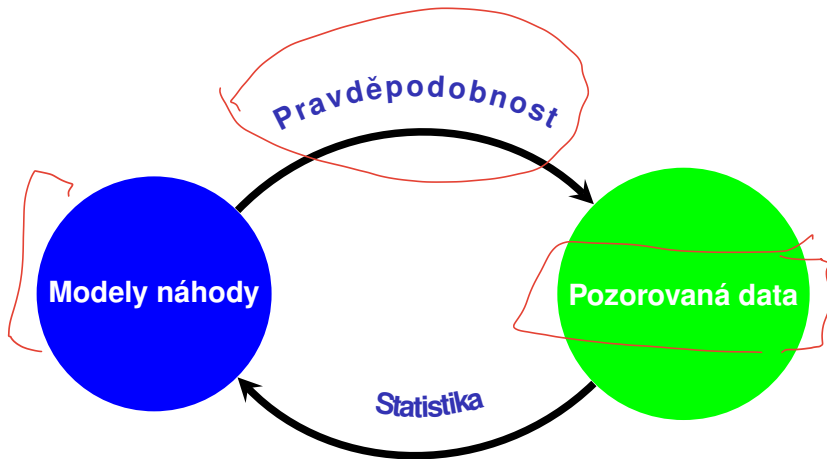
Přehled

Statistika – úvod

Statistika – bodové odhady

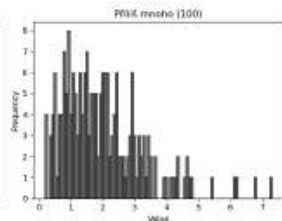
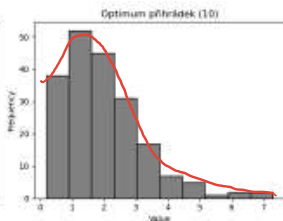
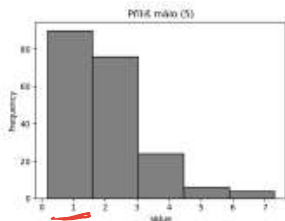
Statistika – intervalové odhady

Plán přednášky



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

Populace Ω	X	$\mathbb{R}$	Vzorek		
					
		1.86	X_1	1.86	
		1.73			
		1.89			
		2.06			
		1.71		X_2	1.71
		1.71			
		2.08		X_3	2.08
		1.91			
		1.67			
		1.87			
		1.67		X_4	1.67
		1.67			
		1.81			
		1.38			
		1.42			
	1.65				
	165				

- ▶ statistická chyba: nevyhnutelná, vzniklá tím, že vybírám vzorek a nevím nic o zbytku populace (snižuje se s velikostí vzorku, pokud je vybírán náhodně)
- ▶ systematická chyba: neuniformní náhodnost (někteří jedinci mají vyšší šanci být ve vzorku), chyby v zápisu, co
 dělat s nedokončenými dotazníky.

Náhodný výběr

Populace Ω	$\xrightarrow{X}$ $\mathbb{R}$
♀	1.86
♀	1.73
♀	1.89
♀	2.06
♀	1.71
♀	1.71
♀	2.08
♀	1.91
♀	1.67
♀	1.87
♀	1.67
♀	1.67
♀	1.81
♀	1.38
♀	1.42
♀	1.65

Vzorek	
	X_1
♀	1.86
	X_2
♀	1.71
	X_3
♀	2.08
	X_4
♀	1.67

náhodný výběr s distr. funkcí F s rozsahem (n)

nezávislé n.v. $X_1, \dots, X_n \sim F$

$\mathcal{F} = \mathcal{D}$

Parametrický model: $F \in \{F_\theta : \theta \in \Theta\}$

Tady: $N(\mu, \sigma^2)$, tj. $\theta = (\mu, \sigma)$. Asi nás zajímá hlavně μ , pak se σ nazývá rušivý parametr (nuisance parameter).

Statistika – přehled

- ▶ neparametrické modely: povolujeme velkou třídu F
- ▶ parametrické modely: $F \in \{F_\theta : \theta \in \Theta\}$
- ▶ příklady:
 - ▶ $Pois(\lambda)$ (parametr $\theta = \lambda$, $\Theta = \underline{\mathbb{R}^+}$)
 - ▶ $U(a, b)$ (parametr $\theta = (a, b)$, $\Theta = \mathbb{R}^2$) $a < b$
 - ▶ $N(\mu, \sigma^2)$ (parametr $\theta = (\mu, \sigma)$, $\Theta = \mathbb{R} \times \mathbb{R}^+$)
- ▶ „Všechny modely jsou špatné, ale některé jsou užitečné.“
(George Box)

Zkoumané úlohy – cíle konfirmační analýzy (confirmatory data analysis)

- ▶ bodové odhady: Jaká je průměrná výška člověka?
(Chceme jedno číslo.)
- ▶ intervalové odhady: Jaká je průměrná výška člověka?
(Chceme interval.)
- ▶ testování hypotéz: Je zkoumaný lék účinný? Zrychlil se algoritmus úpravou parametrů keše?
- ▶ (lineární) regrese: Jak závisí náklon šikmé věže v Pise na čase? Jak závisí rychlost algoritmu na množství paměti pro kešování?
- ▶ *statistika* – libovolná funkce náhodného výběru, tj. např. aritmetický průměr, medián, maximum, atd.
Tj. $T = T(\underbrace{X_1, \dots, X_n}_{\text{množina}})$.

Vzorek – 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.
- ▶ dnes hlavně parametrické modely

Přehled

Statistika – úvod

Statistika – bodové odhady

Statistika – intervalové odhady

Výběrový průměr a rozptyl

Populace $\Omega \xrightarrow{X} \mathbb{R}$

♀	1.86
♀	1.73
♀	1.89
♀	2.06
♀	1.71
♀	1.71
♀	2.08
♀	1.91
♀	1.67
♀	1.87
♀	1.67
♀	1.67
♀	1.81
♀	1.38
♀	1.42
♀	1.65

*rozdíl
= 0*

Populační průměr $\mu = \mathbb{E}X$
populační rozptyl $\sigma^2 = \text{var}(X)$

*$\mu, \sigma^2 \dots$ neznámé
konstanty*

Vzorek

♀	X_1 1.86
♀	X_2 1.71
♀	X_3 2.08
♀	X_4 1.67

vektor n měř

Výběrový průměr $\bar{X}_n$

výběrový rozptyl $\hat{S}_n^0$

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

$$\hat{S}_n^0 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

*náhodné
veličiny,
které
měníme*

*pořadí
= statistiky*

Bodové odhady – Příklady

$$P_n = X_1 + X_2 + \dots + X_n$$

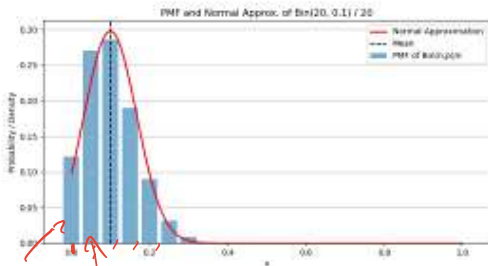
Leváci

$X \sim \text{Ber}(p)$, kde p je skutečný podíl leváků. Je to nějaká neznámá konstanta, kterou chceme odhadnout.

Měření nám dá pro vzorek velikosti n počet leváků

$P_n \sim \text{Bin}(n, p)$ – náhodnou veličinu. Odhadneme p pomocí výběrového průměru: $\hat{p} := P_n/n$.

Bodový odhad $\hat{p}$ je náhodná veličina! $\leftarrow$



$$P_{20} = 0$$
$$P_{20} = 1$$

$$20$$

$$P_{20} = 3$$

$$\hat{p} = \frac{3}{20} = 0.15$$

Bodové odhady – Příklady

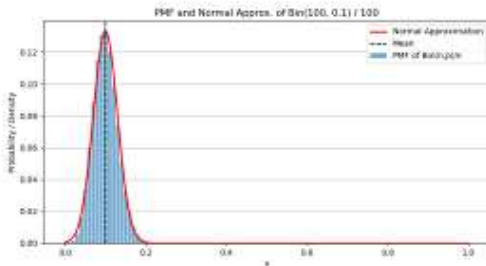
Leváci

$X \sim \text{Ber}(p)$, kde p je skutečný podíl leváků. Je to nějaká neznámá konstanta, kterou chceme odhadnout.

Měření nám dá pro vzorek velikosti n počet leváků

$P_n \sim \text{Bin}(n, p)$ – náhodnou veličinu. Odhadneme p pomocí výběrového průměru: $\hat{p} := P_n/n$.

Bodový odhad $\hat{p}$ je náhodná veličina!



Bodové odhady – Příklady

Leváci

$X \sim \text{Ber}(p)$, kde p je skutečný podíl leváků. Je to nějaká neznámá konstanta, kterou chceme odhadnout.

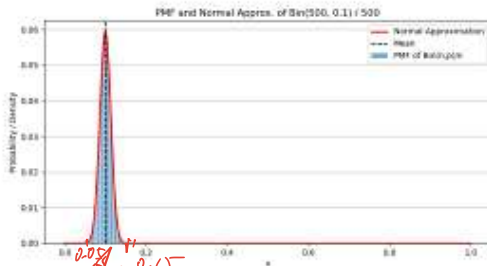
Měření nám dá pro vzorek velikosti n počet leváků

$P_n \sim \text{Bin}(n, p)$ – náhodnou veličinu. Odhadneme p pomocí výběrového průměru: $\hat{p} := P_n/n$.

Bodový odhad $\hat{p}$ je náhodná veličina!

$$E P_n = n p$$

$$E \hat{p} = \frac{1}{n} E P_n = p$$



$$E \hat{p} = p$$

Bodové odhady – Příklady

Výška $X \sim N(\mu, \sigma^2)$ – opět, μ, σ jsou neznámé parametry.

Odhadneme μ pomocí výběrového průměru

$$\hat{\mu} := \bar{X}_n \sim N(\mu, \sigma^2/n).$$

Bodový odhad $\hat{\mu}$ je opět náhodná veličina!

Bodové odhady – Cíle

Definice

Odhad $\hat{\theta} = \hat{\theta}_n = T(\overbrace{X_1, \dots, X_n}^{\text{nezřejmý}})$ parametru θ je

- ▶ nevychýlený/nestranný (unbiased) – pokud $\theta = \mathbb{E}(\hat{\theta}_n)$
- ▶ asymptoticky nevychýlený (asymptotically unbiased) – pokud $\theta = \lim_{n \rightarrow \infty} \mathbb{E}(\hat{\theta}_n)$
- ▶ konzistentní – pokud $\hat{\theta}_n \xrightarrow{P} \theta$ ↗ $\forall \varepsilon > 0 \quad P(|\hat{\theta}_n - \theta| > \varepsilon) \rightarrow 0$
- ▶ vychýlení (bias) $\text{bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$ ↖
- ▶ střední kvadratická chyba (mean squared error, MSE) je $\mathbb{E}((\hat{\theta}_n - \theta)^2)$ ↖ $\mathbb{E}(\hat{\theta}_n - \theta) =$
- ▶ standardní chyba (standard error) se je $\sqrt{\text{var}(\hat{\theta}_n)} = \sigma(\hat{\theta}_n)$ ↖ $\uparrow$ n.v. – násobek

Věta

$$\underline{P(|\hat{\theta}_n - \theta| > \varepsilon) \leq \frac{MSE}{\varepsilon^2}}$$

Věta

$$MSE(\hat{\theta}_n) = \text{bias}(\hat{\theta}_n)^2 + \text{var}(\hat{\theta}_n)$$

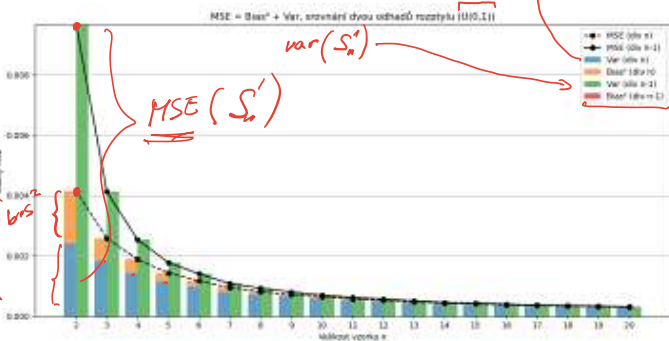
Vlastnosti výběrového průměru a rozptylu

sluč Z.V.C.

$$E \frac{X_1 + \dots + X_n}{n} = \frac{EX_1 + \dots + EX_n}{n} = EX = \mu$$

Věta

1. $\hat{\mu} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ je konzistentní nestranný odhad μ $SE = \sigma(\bar{X}_n)$
2. $\hat{S}_n^0 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ je konzistentní asymptoticky nevychýlený odhad σ^2
3. $\hat{S}_n^1 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ je konzistentní nevychýlený odhad σ^2



Metoda momentů

- ▶ $m_r(\theta) := \mathbb{E}(X^r)$ pro $X \sim F_\theta$... r -tý moment r = 1, 2, ...
- ▶ $\widehat{m_r(\theta)} := \frac{1}{n} \sum_{i=1}^n X_i^r$ pro náhodný výběr $X_1, \dots, X_n$ z F_θ ...
 r -tý výběrový moment

Věta

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

(důk: sled 2.0.8.)

Metoda momentů: Volíme takové θ , které řeší soustavu rovnic

$$\underline{m_r(\theta)} = \underline{\widehat{m_r}} \quad r = 1, \dots, k.$$

(Typicky k bude počet reálných čísel, která tvoří parametr θ .)

Metoda momentů – příklady

Leváci $X_i \sim \text{Ber}(p)$

$$p = \theta \quad P_i = X_1, \dots, X_n$$

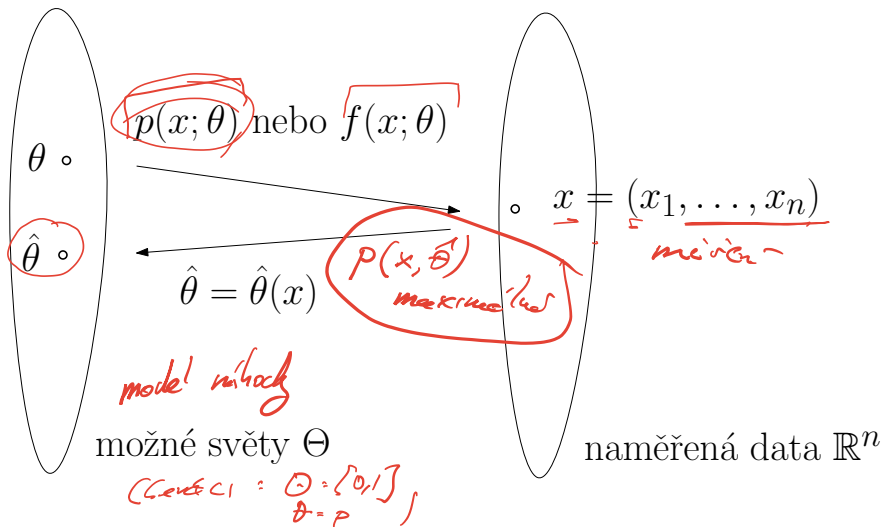
$$m_1(\theta) = EX_i = p$$

$$\widehat{m_1(\theta)} = \overline{X_n} = \frac{P_n}{n}$$

$$m_1 = \widehat{m_1}$$

$$p \rightarrow \frac{P_n}{n} \leftarrow \text{volume } \hat{p} = \frac{P_n}{n}$$

Metoda maximální věrohodnosti (maximal likelihood, ML)



► náh. výběr $X = (X_1, \dots, X_n)$ z modelu s parametrem θ

► možný výsledek $x = (x_1, \dots, x_n)$

► odrušená pravděpodobnostní funkce $p(x, \theta)$

Metoda maximální věrohodnosti (maximal likelihood, ML)

model měření

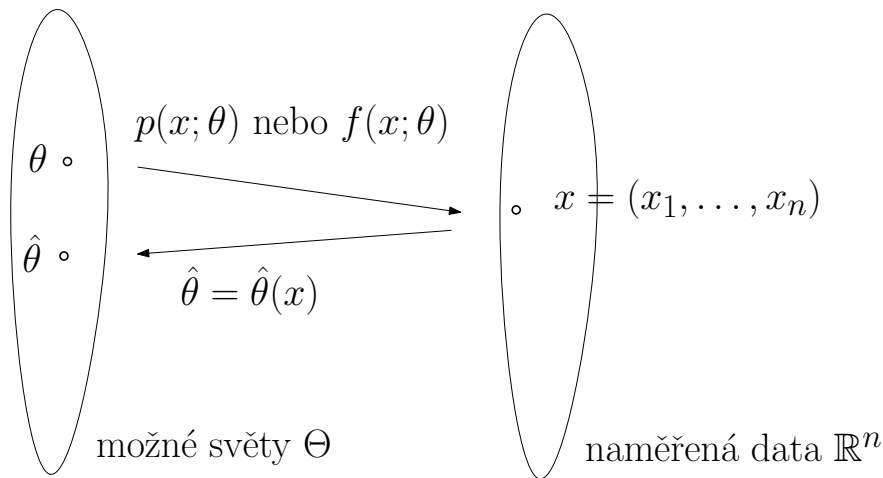
- ▶ náh. výběr $X = (X_1, \dots, X_n)$ z modelu s parametrem θ
- ▶ možný výsledek $x = (\underline{x}_1, \dots, \underline{x}_n) \in \mathbb{R}^n$ *realizace* $\leftarrow$ *to is a realization*
- ▶ ... sdružená pravděpodobnostní funkce $\underline{p}_X(x; \theta)$
- ▶ ... sdružená hustota $\underline{f}_X(x; \theta)$
- ▶ *věrohodnost (likelihood)* $\underline{L}(x; \theta)$ značí $\underline{p}_X$ nebo $\underline{f}_X$
- ▶ normálně: máme pevné θ , a $L(x; \theta)$ je funkce x
- ▶ teď: máme pevné x a $L(x; \theta)$ je funkce θ

Metoda MV (ML):

volíme takové θ , pro které je $L(x; \theta)$ maximální

$\hat{\theta}_{ML}$

Metoda maximální věrohodnosti (maximal likelihood, ML)



► náh. výběr $X = (X_1, \dots, X_n)$ z modelu s parametrem θ

► možný výsledek $x = (x_1, \dots, x_n)$

► odručená pravděpodobnostní funkce $p(x; \theta)$

Metoda maximální věrohodnosti (maximal likelihood, ML)

► Metoda MV (ML):

volíme takové θ , pro které je $L(x; \theta)$ maximální

► definujeme také $\ell(x; \theta) = \log(L(x; \theta))$

► díky nezávislosti je

$$x = (x_1, \dots, x_n)$$

$$\max \rightarrow L(x; \theta) = \prod_{i=1}^n L(x_i; \theta)$$

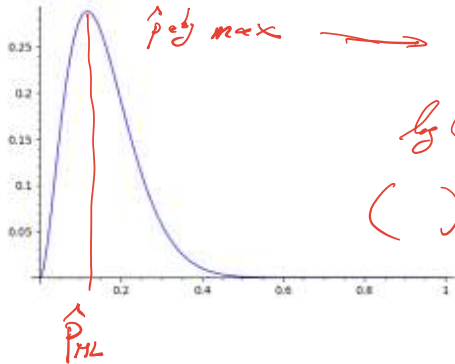
$$\Downarrow \ell(x; \theta) = \log \prod_{i=1}^n L(x_i; \theta) = \sum_{i=1}^n \log L(x_i; \theta)$$

$$\ell(x; \theta)' = 0 \rightarrow \sum_{i=1}^n \frac{L'(x_i; \theta)}{L(x_i; \theta)}$$

derivative
wrt θ

ML – leváci

```
plot(binomial(17, 0.2)*p^2*(1-p)^15, [0, 1])
```



n meren-
 k levek

$n-k$

$\hat{p}_{HL} \text{ max} \rightarrow P(P_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$

$$\lg(\quad) = \lg(\binom{n}{k}) + k \cdot \lg p + (n-k) \lg(1-p)$$

$$(\quad)' = 0$$

$$\Rightarrow \frac{k}{p} + \frac{n-k}{1-p} \cdot (-1)$$

$$\frac{n-k}{1-p} = \frac{k}{p}$$

$$\} p = \frac{k}{n}$$

Přehled

Statistika – úvod

Statistika – bodové odhady

Statistika – intervalové odhady

Intervalové odhady

- místo jednoho čísla s nejistým významem vypočítáme z dat interval $[D, H]$

Definice

Nechť D, H jsou n.v. které závisí na náhodném výběru $X = (X_1, \dots, X_n)$. Tyto n.v. určují intervalový odhad, interval spolehlivosti též konfidenční interval o spolehlivosti $1 - \alpha$ ($1 - \alpha$ confidence interval), pokud

0.05

$$P(\underline{D} \leq \theta \leq \underline{H}) \geq 1 - \alpha.$$

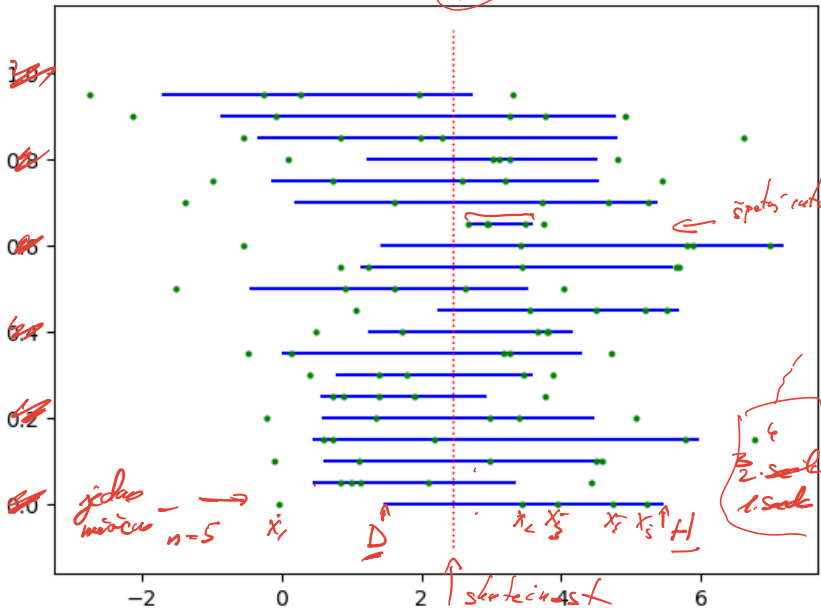
obousměrné
odhady

Číslo α se nazývá koeficient spolehlivosti. Někdy také uvažujeme jednostranné intervalové odhady (případ $D = -\infty$, resp. $H = +\infty$). Můžeme také požadovat silnější vlastnosti: $P(\theta > H) = P(\theta < D) = \alpha/2$.

Než pokročíme dále, uvědomte si, co je v uvedených výrazech náhodné!

σ

podíl
řádek, kde
neprobíhá
 $\propto$
 $\frac{1}{20}$



Hledaný interval budeme typicky uvažovat ve tvaru $[x - \delta, x + \delta]$ (pro vhodně určené x a δ).

Tento interval pro zkrácení zapisujeme jako $x \pm \delta$.

Většina našich metod bude založena na následujícím přístupu:

► Máme nevychýlený bodový odhad $\hat{\theta}$ pro parametr θ

► $\hat{\theta}$ má (přibližně) normální rozdělení

► Pak $\hat{\theta} \pm z_{\alpha/2} \cdot se$ je $(1 - \alpha)$ -CI, kde

$$z_{\alpha/2} := \Phi^{-1}(1 - \alpha/2),$$

$$se := \sigma(\hat{\theta}),$$

konfidenční int.

$$E\hat{\theta} = \theta$$

$$\hat{\theta} = \bar{X}_n \dots \text{Cenová}$$

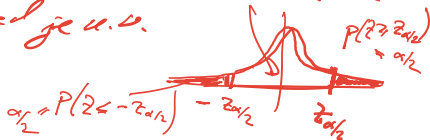
$$P(z_{\alpha/2} \leq Z \leq z_{\alpha/2})$$

$$1 - \alpha \quad \varphi \cdot \frac{1}{2}$$

$$P(Z \geq z_{\alpha/2}) = \alpha/2$$

$$se = \sigma(\hat{\theta}),$$

pro $Z \sim N(0,1)$



Důkaz.

Provedme standardizaci $Z = \text{stand}(\hat{\theta}) = \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})}$. Z vlastností normálního rozdělení víme, že $Z \sim N(0, 1)$, můžeme tedy pro jeho popis použít distribuční funkci Φ — nezávisle na hodnotě θ , se.

Z volby čísla $z_{\alpha/2}$ víme, že $P(Z \leq z_{\alpha/2}) = 1 - \alpha/2$, ze symetrie také platí $P(Z \leq -z_{\alpha/2}) = \alpha/2$, tudíž

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

Když dosadíme definici Z a ekvivalentně upravíme nerovnici, zjistíme, že také

$$P(\hat{\theta} - z_{\alpha/2} \cdot \sigma(\hat{\theta}) \leq \theta \leq \hat{\theta} + z_{\alpha/2} \cdot \sigma(\hat{\theta})) = 1 - \alpha.$$

$$\hat{\theta} - z_{\alpha/2} \cdot se \leq \theta \leq \hat{\theta} + z_{\alpha/2} \cdot se$$



Intervalové odhady normální n.v. se známým rozptylem

Normální veličiny $N(\theta, \sigma^2)$ Máme náhodný výběr

$X_1, \dots, X_n \sim N(\theta, \sigma^2)$. Přitom θ neznáme, ale σ ano.

(Realistický příklad je teploměr se známou chybou měření, kterým měříme neznámou teplotu.) Z Věty ?? víme, že $\bar{X}_n$ je nevychýlený odhad. Z kapitoly o normálním rozdělení víme, že $\bar{X}_n$ má normální rozdělení. Z důkazu Věty ?? víme, že $\text{se} = \sigma(\bar{X}) = \sigma/\sqrt{n}$.

Závěr je, že

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \text{ je } (1 - \alpha)\text{-CI.}$$

Jako konkrétní příklad: pokud naměříme hodnoty TODO

$$z_{0.025} = 1.96.$$

Naše první statistická úloha bude vlastně cvičení na práci s normální veličinou. Větu formulujeme obecně, ale jako realistický příklad si můžeme představit fyzikální měření – měříme hmotnost opakovaným měřením na stejné váze, jejíž

Intervalové odhady normální náhodné veličiny