Základy pravděpodobnosti a matematické statistiky

9. Odhady

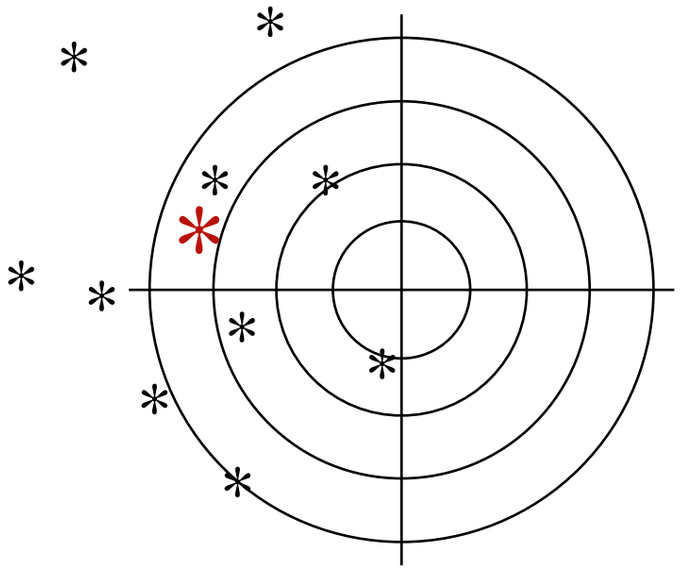
FAKULTA STROJNÍ ČVUT V PRAZE

prof. RNDr. Gejza Dohnal, CSc.

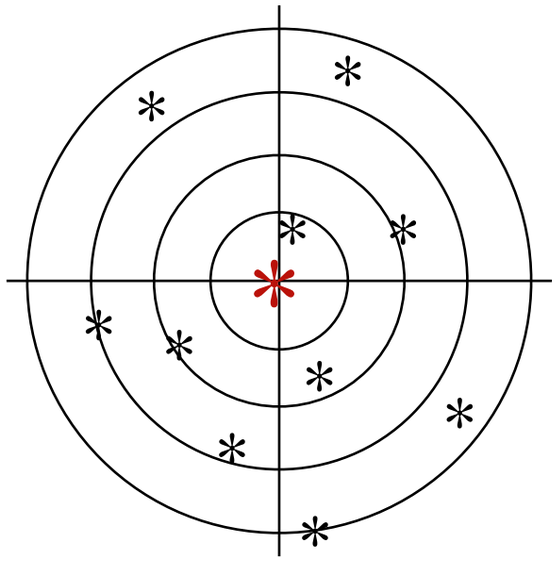
ak. rok 2021/2022

9. Odhady

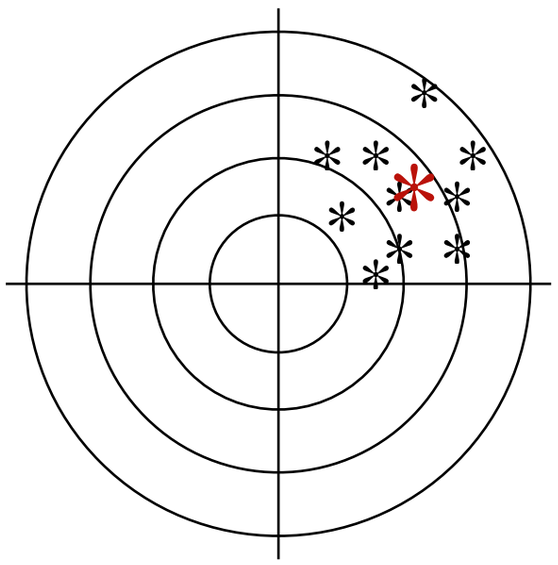
Bodové odhady



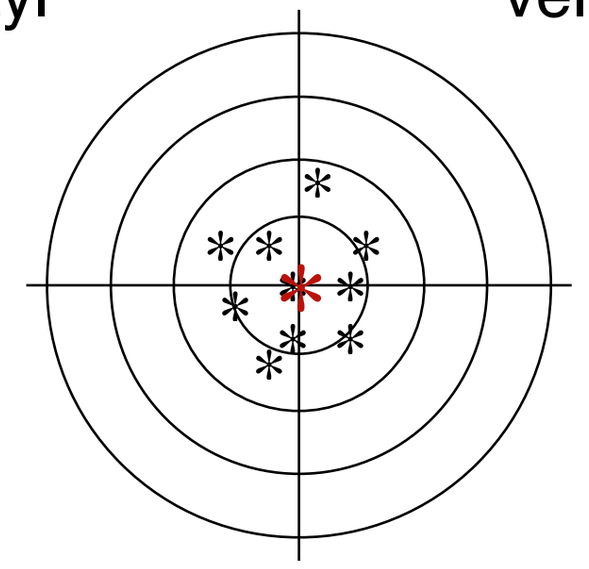
Vychýlený,
velký rozptyl



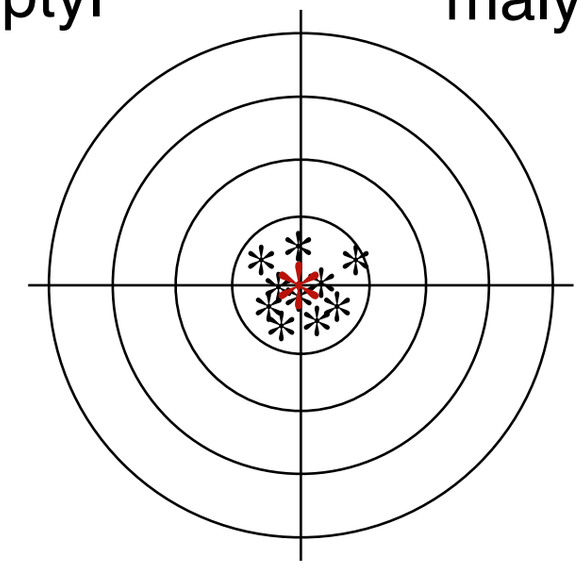
Nevychýlený,
velký rozptyl



Vychýlený,
malý rozptyl



Nevychýlený,
malý rozptyl



Nejlepší nestranný

Bodové odhady

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$
výběru $X_1, X_2, \ldots, X_n$ = bodové odhady pravděpodobnostních charakteristik

Pravděpodobnostní charakteristiky	Výběrové charakteristiky
Střední hodnota $E(X) = \int_{-\infty}^{\infty} x f(x) dx$	Výběrový průměr $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$
Momenty $\mu_k(X) = E(X - E(X))^k$	Výběrové momenty $m_k(X) = \sum_{i=1}^{n} (x_i - \bar{X})^k$
Rozptyl $var(X) = E(X - E(X))^2$	Výběrový rozptyl $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2$
Kvantily $\tilde{x}_{100\alpha}$	Výběrové kvantily $X_{([np]+1)}$

Bodové odhady

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$ i.i.d. výběru $X_1, X_2, \ldots, X_n$.

Odhad střední hodnoty: $\hat{\mu} = \bar{X}_n$ - je nevychýlený (nestranný, unbiased)

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

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

$$\frac{1}{n^2}\left(\sum_{i=1}^{n} E(X_i - \mu)^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E(X_i - \mu)(X_j - \mu)\right) = \frac{\sum_{i=1}^{n}\mathrm{Var}(X_i)}{n^2}$$

$$= \frac{\sigma^2}{n}$$ - rozptyl konverguje k 0 pro n→∞

Bodové odhady

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$ i.i.d. výběru $X_1, X_2, \ldots, X_n$.

Odhad rozptylu: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ - je vychýlený (biased)

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

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

$$\sigma^2 = E(X_i^2) - \big(E(X_i)\big)^2 = E(X_i^2) - \mu^2 \;\Rightarrow E(X_i^2) = \sigma^2 + \mu^2$$

$$\frac{1}{n}\sigma^2 = E(\bar{X}_n^2) - \mu^2 \;\Rightarrow E(\bar{X}_n^2) = \frac{\sigma^2}{n} + \mu^2$$

Bodové odhady

Statistické charakteristiky: jsou spočteny na základě pozorování
$x_1, x_2, \ldots, x_n$ i.i.d. výběru $X_1, X_2, \ldots, X_n$.

$$E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \qquad s^2 = \frac{n}{n-1}\hat{\sigma}^2 \quad \Rightarrow \quad E(s^2) = \sigma^2$$

Tedy výběrový rozptyl $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ je nestranným odhadem rozptylu σ^2

Jak je to s rozptylem těchto odhadů?

$$\mathrm{Var}(s^2) = \mathrm{Var}\left(\frac{n}{n-1}\hat{\sigma}^2\right) = \left(\frac{n}{n-1}\right)^2 \mathrm{Var}(\hat{\sigma}^2) \geq \mathrm{Var}(\hat{\sigma}^2)$$

Tedy:

rozptyl základního souboru $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ je vychýlený, ale má menší rozptyl

výběrový rozptyl $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ je nevychýlený, ale má větší rozptyl

Bodové odhady

Odhadujeme neznámý parametr θ. Najdeme odhadovou statistiku $\hat{\theta}(X_1, X_2, \ldots, X_n)$ tak, aby splňovala některou (nejlépe všechny) z následujících vlastností:

- Nestrannost: $E\big(\hat{\theta}(X_1, X_2, \ldots, X_n)\big) = \theta$
- Vydatnost: $\mathrm{Var}\big(\hat{\theta}(X_1, X_2, \ldots, X_n)\big)$ je minimální
- Konzistence: $\lim_{n\to\infty} \mathrm{Var}\big(\hat{\theta}(X_1, X_2, \ldots, X_n)\big) = 0$

Jak se hledá odhadová statistika?

- Metoda maximální věrohodnosti (maximalizuje věrohodnostní funkci)
- Momentová metoda (srovnává teoretické a výběrové momenty)
- Metoda nejmenších čtverců (minimalizuje čtvercovou odchylku)

Bodové odhady

Metoda maximální věrohodnosti (Maximum likelihood)

- pozorováním i.i.d. výběru $X_1, X_2, \ldots, X_n$ dostaneme pozorování $x_1, x_2, \ldots, x_n$
- uvažujme sdruženou hustotu náhodného vektoru $(X_1, X_2, \ldots, X_n)$, která závisí na neznámém parametru $f(x_1, x_2, \ldots, x_n; \theta)$
- s touto funkcí budeme nadále zacházet jako s funkcí neznámé θ a budeme ji nazývat věrohodnostní funkcí: $l(\theta; x_1, x_2, \ldots, x_n)$
- hledáme $\hat{\theta} = \arg\max_{\theta} l(\theta; x_1, \ldots, x_n)$ a nazveme je maximálně věrohodným odhadem parametru θ.

Často namísto funkce $l(\theta; x_1, x_2, \ldots, x_n)$ maximalizujeme logaritmickou věrohodnostní funkci $L(\theta; x_1, x_2, \ldots, x_n) = \ln(l(\theta; x_1, x_2, \ldots, x_n))$.

Bodové odhady

Příklad: Odhad parametru Poissonova rozdělení $p(k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$

Máme pozorování $k_1, k_2, \ldots, k_n$ a vytvoříme věrohodnostní funkci

$$l(\lambda; k_1, \ldots, k_n) = \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!}\mathrm{e}^{-\lambda} = \mathrm{e}^{-n\lambda}\lambda^{\sum_{i=1}^{n} k_i} \prod_{i=1}^{n} \frac{1}{k_i!}$$

Vytvoříme logaritmickou věrohodnostní funkci:

$$L(\lambda; k_1, \ldots, k_n) = -n\lambda + \ln(\lambda)\sum_{i=1}^{n} k_i + \ln\left(\prod_{i=1}^{n} \frac{1}{k_i!}\right)$$

a hledáme maximum:

$$\frac{\mathrm{d}L(\lambda; k_1, \ldots, k_n)}{\mathrm{d}\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} k_i$$

$$-n + \frac{1}{\hat{\lambda}}\sum_{i=1}^{n} k_i = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{\sum_{i=1}^{n} k_i}{n}$$

= ML odhad λ (MLE λ)

Bodové odhady

Momentová metoda

- pozorováním i.i.d. výběru $X_1, X_2, \ldots, X_n$ z nějakého rozdělení s d.f. $F(x;\theta)$, kde $\theta = (\theta_1, \ldots, \theta_k)$ je k neznámých parametrů
- Spočteme k teoretických momentů $\mu_1, \ldots, \mu_k$ a k výběrových momentů $m_1, \ldots, m_k$
- porovnáním těchto momentů dostaneme k rovnic, z nichž vyjádříme k odhadů neznámých parametrů.

Bodové odhady

Příklad: Odhad parametrů normálního rozdělení $\mathrm{N}(\mu, \sigma^2)$ momentovou metodou.

$$\mu_1 = E(X) = \mu, \qquad \mu_2 = E(X^2) = \sigma^2 + \mu^2$$

$$m_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

odtud:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

Intervalové odhady

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$ výběru $X_1, X_2, \ldots, X_n$.

Intervalový odhad je interval $(\underline{\theta}(X), \overline{\theta}(X))$ takový, že

$$P((\theta \in (\underline{\theta}(X), \overline{\theta}(X))) = 1-\alpha$$

- Předpokládáme nějaké rozdělení výběru - zpravidla normální

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

$$\frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} \sim N(0, 1) \qquad \frac{\bar{X}_n - \mu}{s}\sqrt{n} \sim t(n-1)$$

$$P(\underline{\mu} \le \mu \le \overline{\mu}) = P(-\underline{\mu} \ge -\mu \ge -\overline{\mu}) = P(\bar{X} - \underline{\mu} \ge \bar{X} - \mu \ge \bar{X} - \overline{\mu})$$

$$= P\left(\frac{\bar{X} - \underline{\mu}}{\sigma}\sqrt{n} \ge \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \ge \frac{\bar{X} - \overline{\mu}}{\sigma}\sqrt{n}\right)$$

$$= P\left(\frac{\bar{X} - \overline{\mu}}{\sigma}\sqrt{n} \le Z \le \frac{\bar{X} - \underline{\mu}}{\sigma}\sqrt{n}\right) = 1 - \alpha \ = P(u_{\alpha/2} \le Z \le u_{1-\alpha/2})$$

Intervalové odhady

$$\frac{\bar{X} - \overline{\mu}}{\sigma}\sqrt{n} = u_{\alpha/2} \qquad \overline{\mu} = \bar{X} - \frac{\sigma}{\sqrt{n}}u_{\alpha/2} = \bar{X} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}$$

$$\frac{\bar{X} - \underline{\mu}}{\sigma}\sqrt{n} = u_{1-\alpha/2} \qquad \underline{\mu} = \bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}$$

$$P(\,\bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2} \le \mu \le \bar{X} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}\,) = 1 - \alpha$$

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

$$\frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} \sim N(0,\, 1)$$

$$\frac{\bar{X}_n - \mu}{s}\sqrt{n} \sim t(n-1)$$

$$= P(\,\frac{\bar{X} - \overline{\mu}}{\sigma}\sqrt{n} \le Z \le \frac{\bar{X} - \underline{\mu}}{\sigma}\sqrt{n}\,) = 1 - \alpha \;\; = P(\,u_{\alpha/2} \le Z \le u_{1-\alpha/2})$$

Intervalové odhady

$$\frac{\bar{X} - \overline{\mu}}{\sigma}\sqrt{n} = u_{\alpha/2} \qquad \overline{\mu} = \bar{X} - \frac{\sigma}{\sqrt{n}}u_{\alpha/2} = \bar{X} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}$$

$$\frac{\bar{X} - \underline{\mu}}{\sigma}\sqrt{n} = u_{1-\alpha/2} \qquad \underline{\mu} = \bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}$$

$$P(\, \bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}\,) = 1 - \alpha$$

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

$$\frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} \sim N(0, 1)$$

$$\frac{\bar{X}_n - \mu}{s}\sqrt{n} \sim t(n-1)$$

$$P(\, \bar{X} - \frac{s}{\sqrt{n}}t_{1-\alpha/2}(n-1) \leq \mu \leq \bar{X} + \frac{s}{\sqrt{n}}t_{1-\alpha/2}(n-1)\,) = 1 - \alpha$$

Tedy intervalový odhad je ve tvaru: $(\, \bar{X} - SE,\, \bar{X} + SE\,)$

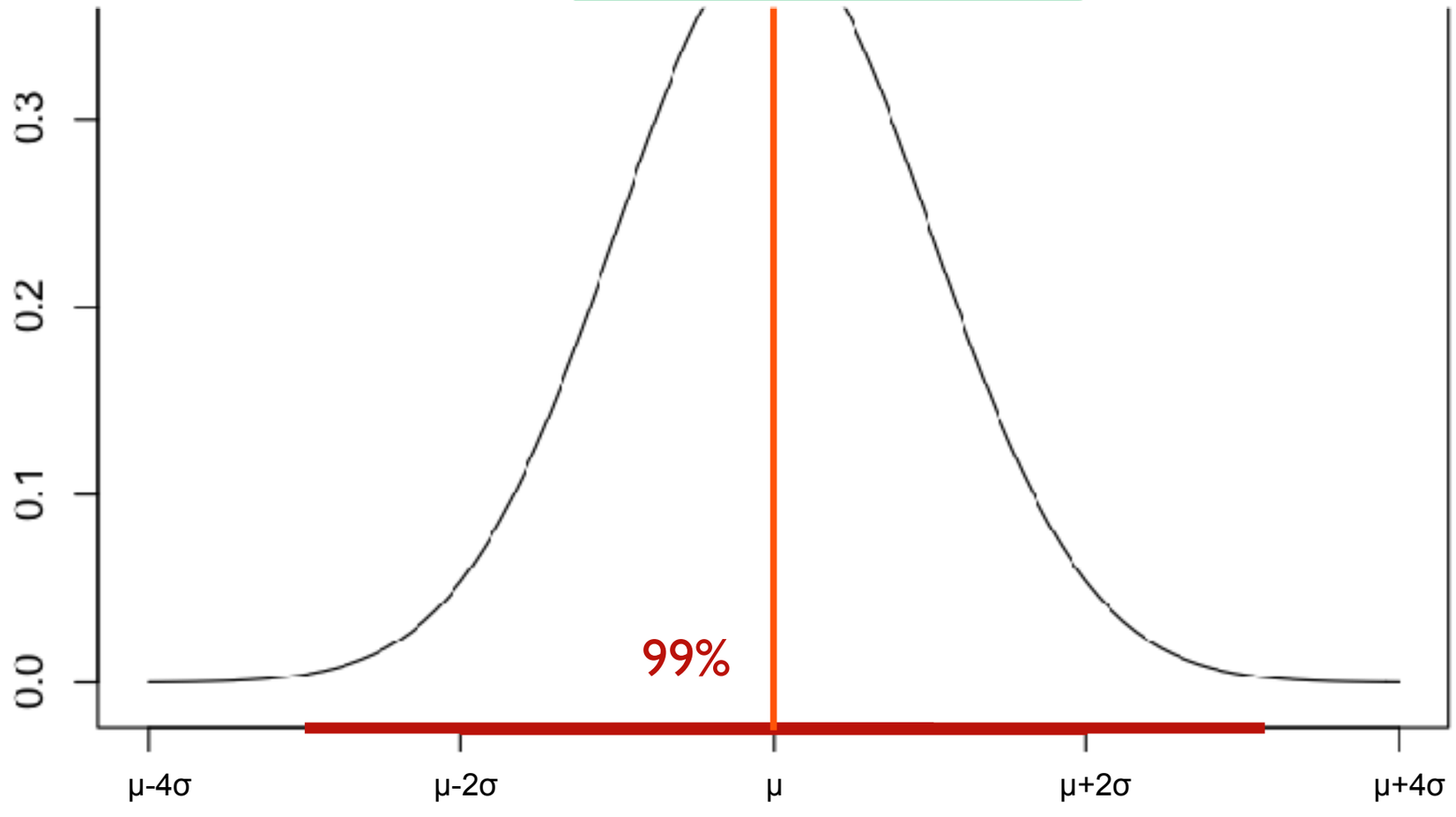
kde SE je tzv. standardní chyba.

Intervalové odhady

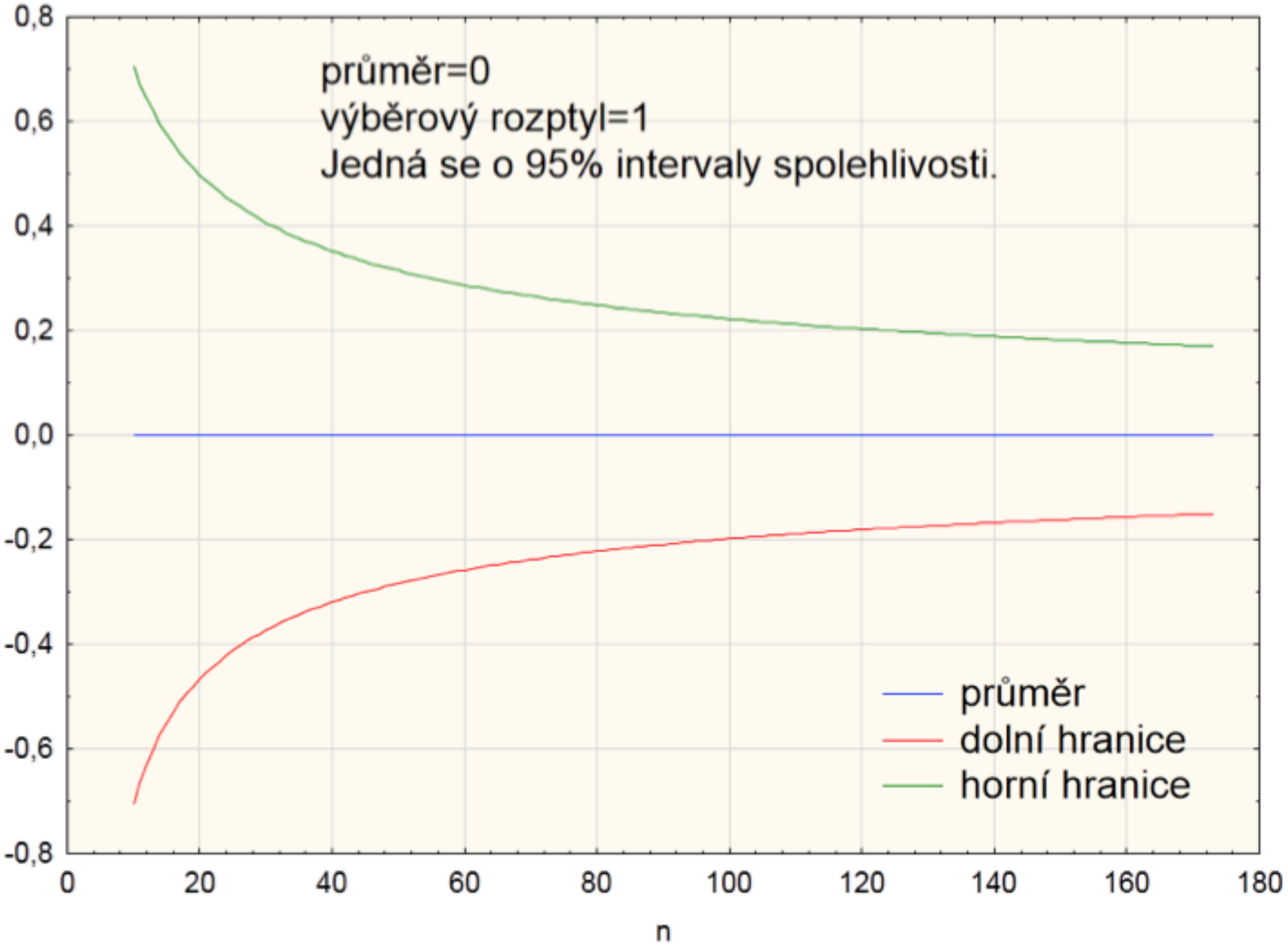
Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Int. spolehl. -90,000%	Int. spolehl. +90,000%	Referenční konstanta	t	SV	p
spotřeba l/100	13,91429	0,555888	14	0,148567	13,65118	14,17739	12,50000	9,519502	13	0,000000

Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Int. spolehl. -95,000%	Int. spolehl. +95,000%	Referenční konstanta	t	SV	p
spotřeba l/100	13,91429	0,555888	14	0,148567	13,59333	14,23525	12,50000	9,519502	13	0,000000

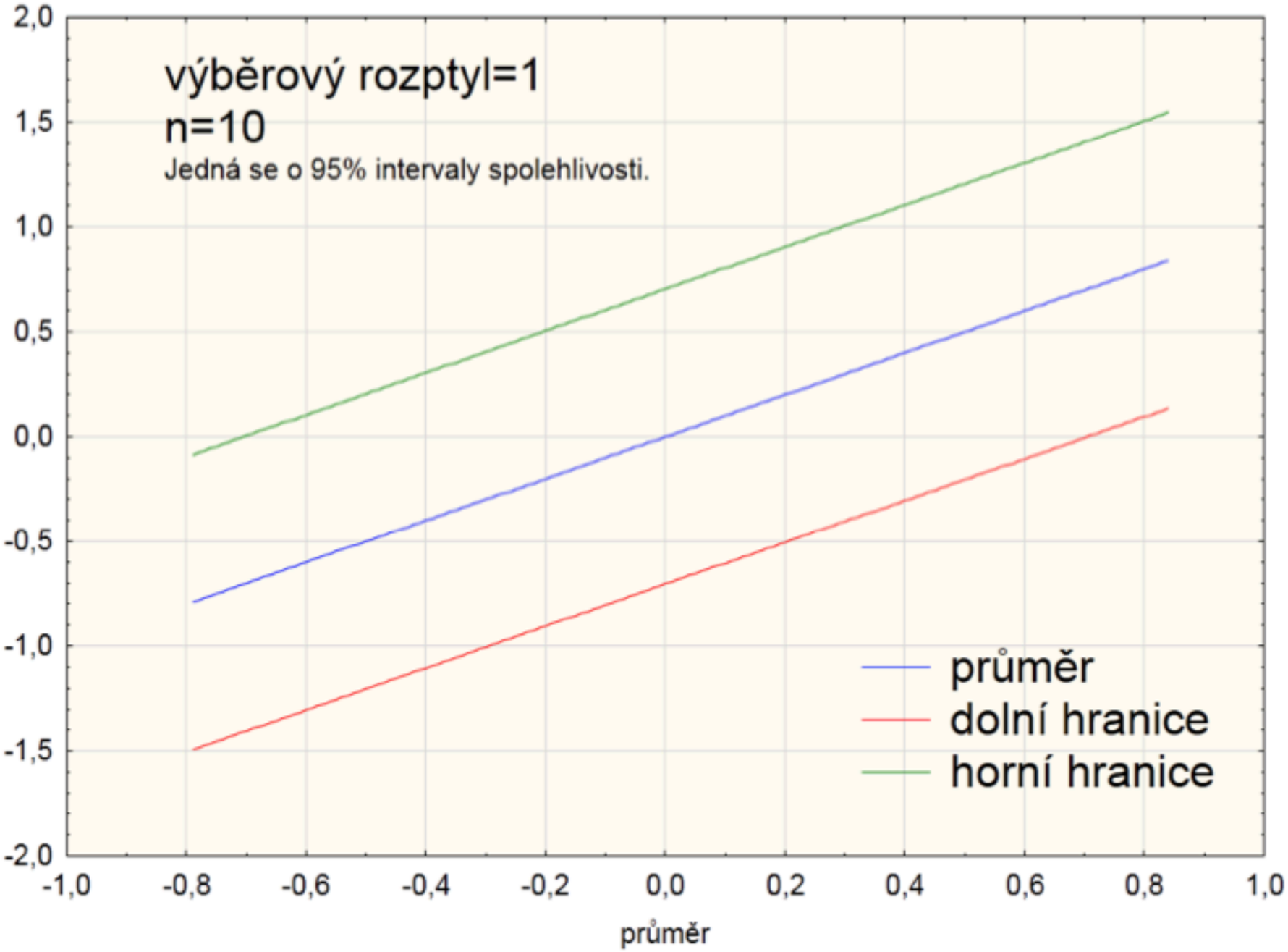
Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Int. spolehl. -99,000%	Int. spolehl. +99,000%	Referenční konstanta	t	SV	p
spotřeba l/100	13,91429	0,555888	14	0,148567	13,46676	14,36181	12,50000	9,519502	13	0,000000



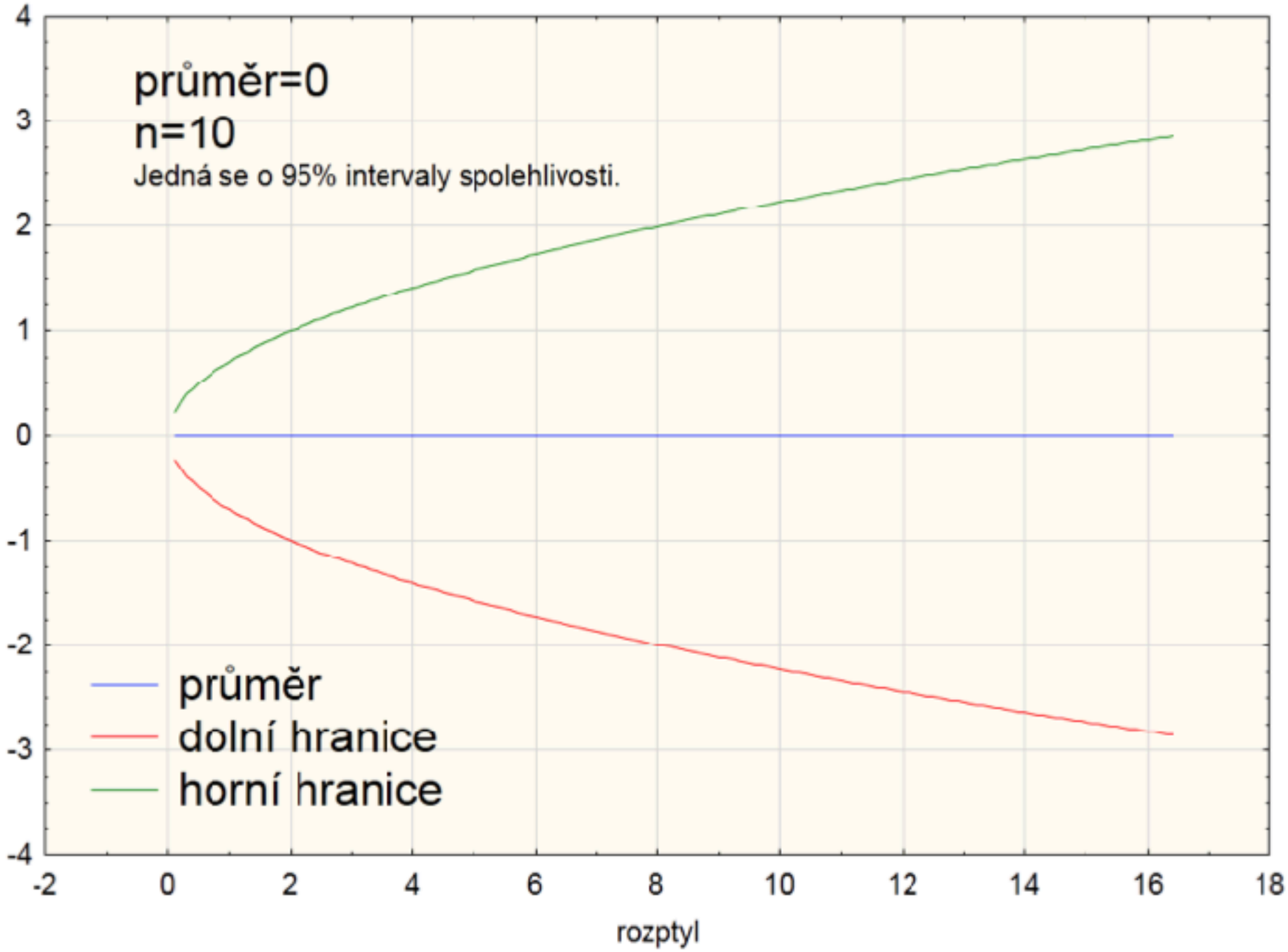
Intervalové odhady



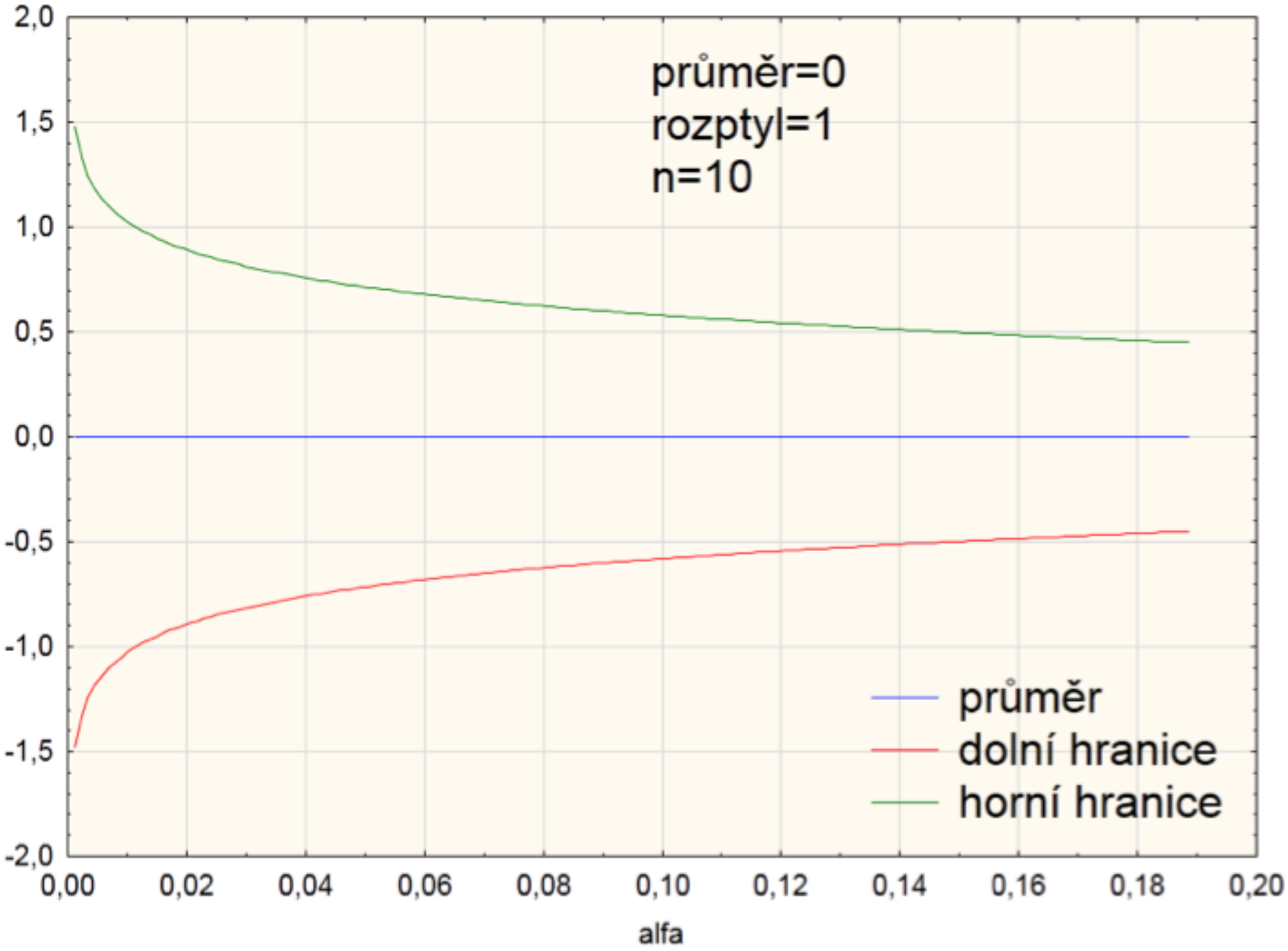
Intervalové odhady



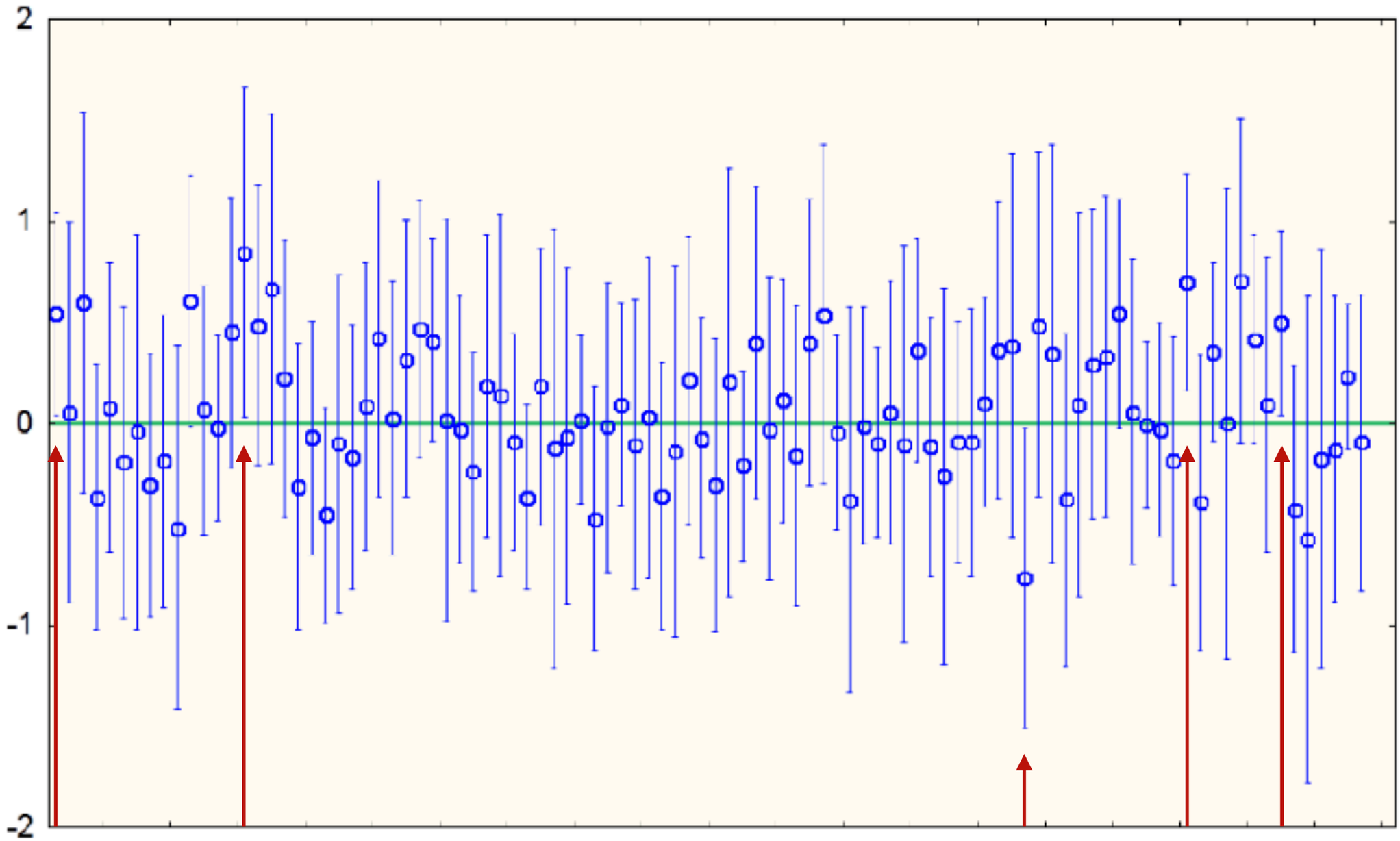
Intervalové odhady



Intervalové odhady



Intervalové odhady



Simulační příklad: N=100, μ=0, α=0,05, (5 intervalů mimo)

Intervalové odhady

Příklad: **Test spotřeby automobilu**

Deklarovaná průměrná spotřeba = 12,5l/100km

$$\bar{X} = \frac{1}{14}\sum_{i=1}^{14} x_i = \frac{194,8}{14} = 13,914$$

$$s^2 = \frac{1}{13}\left[\sum_{i=1}^{14} x_i^2 - 14.\bar{x}^2\right] = \frac{4,017}{13} = 0,309$$

$$SE = \sqrt{\frac{0,309}{14}} t_{0,95}(13) = 0,15.2,16 = 0,32$$

Tedy intervalový odhad je $(13{,}59;\ 14{,}23)$.

Protože $12{,}5 \notin (13{,}59;\ 14{,}23)$, můžeme tvrdit, že naměřená spotřeba se od deklarované *statisticky významně liší*.

n	*spotřeba (l/100)*
1	12,8
2	13,5
3	14,2
4	13,6
5	14,1
6	14,5
7	13,6
8	13,9
9	14,3
10	15,1
11	13,7
12	13,4
13	13,9
14	14,2

Intervalové odhady

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$ výběru $X_1, X_2, \ldots, X_n$.

Intervalový odhad je interval $(\underline{\theta}(X), \overline{\theta}(X))$ takový, že

$$P((\theta \in (\underline{\theta}(X), \overline{\theta}(X))) = 1-\alpha$$

- Předpokládáme nějaké rozdělení výběru - zpravidla normální

- Příklad 2: odhad rozptylu s^2

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}$$

Intervalové odhady

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$ výběru $X_1, X_2, \ldots, X_n$.

Intervalový odhad je interval $(\underline{\theta}(X), \overline{\theta}(X))$ takový, že

$$P((\theta \in (\underline{\theta}(X), \overline{\theta}(X))) = 1-\alpha$$

- Předpokládáme nějaké rozdělení výběru - zpravidla normální

- Příklad 3: odhad pravděpodobnosti p
  - provedeme posloupnost n nezávislých alternativních pokusů
  - spočteme relativní četnost p případů, v nichž nastal sledovaný jev

$$\bar{X} - \sqrt{\frac{p(1-p)}{n}}\, u_{1-\alpha/2} \leq p \leq \bar{X} + \sqrt{\frac{p(1-p)}{n}}\, u_{1-\alpha/2}$$

Intervalové odhady

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$ výběru $X_1, X_2, \ldots, X_n$.

Intervalový odhad je interval $(\underline{\theta}(X), \overline{\theta}(X))$ takový, že

$$P((\theta \in (\underline{\theta}(X), \overline{\theta}(X))) = 1-\alpha$$

- Předpokládáme nějaké rozdělení výběru - zpravidla normální

- Příklad 3: odhad pravděpodobnosti p
  - při šetření jsme zjistili, že ze 100 dotázaných respondentů 43 se chystá volit stranu mírného pokroku v mezích zákona
  - má tato strana šanci vyhrát ve volbách?

$$\bar{X} - \sqrt{\frac{p(1-p)}{n}}\, u_{1-\alpha/2} \leq p \leq \bar{X} + \sqrt{\frac{p(1-p)}{n}}\, u_{1-\alpha/2}$$

$\bar{X} = 0,43$
$u_{0,975} = 1,96$

Tedy intervalový odhad je (0,33; 0,53) => SMPvMZ má statisticky významnou šanci získat nadpoloviční většinu.