Základy pravděpodobnosti a matematické statistiky

9. Odhady statistických charakteristik

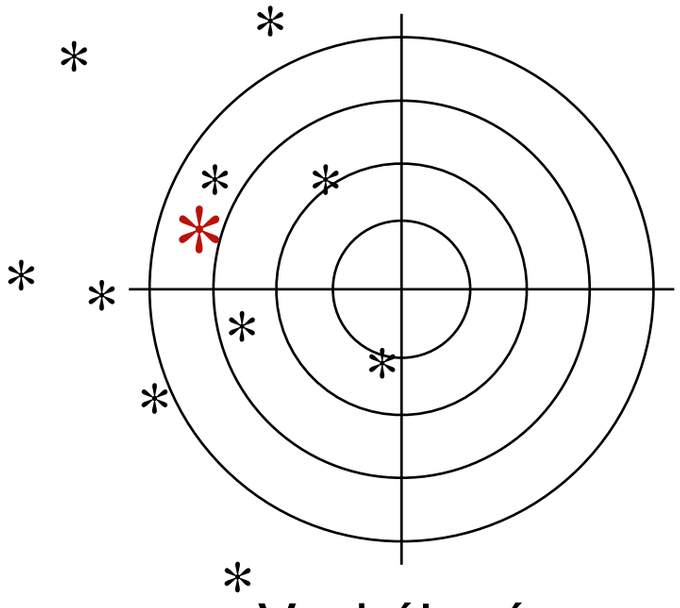
FAKULTA STROJNÍ ČVUT V PRAZE

prof. RNDr. Gejza Dohnal, CSc.

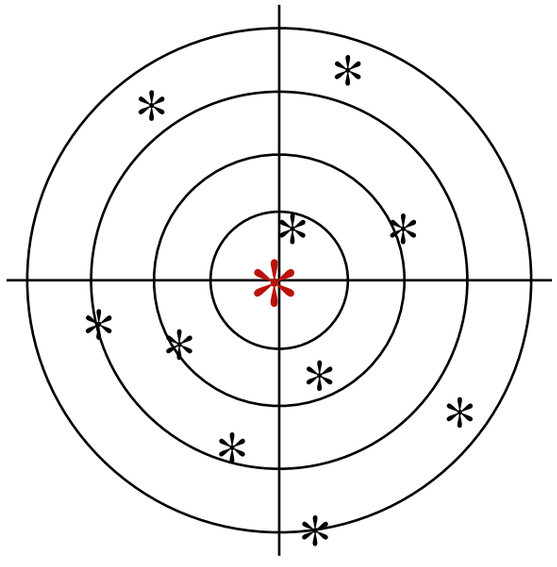
ak. rok 2023/2024

9. Odhady statistických charakteristik

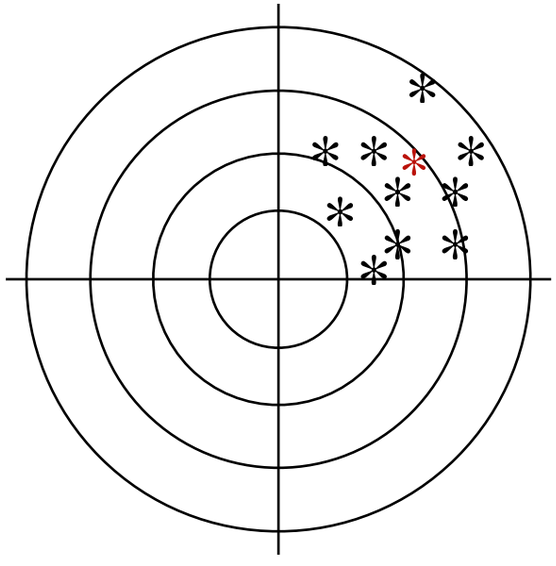
Bodové odhady



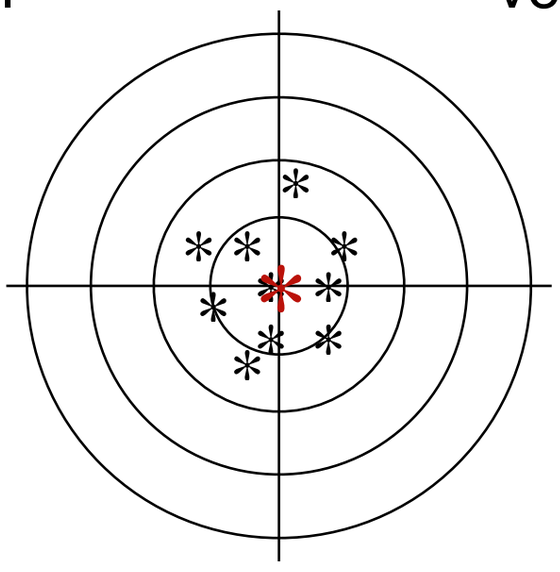
Vychýlený,
velký rozptyl



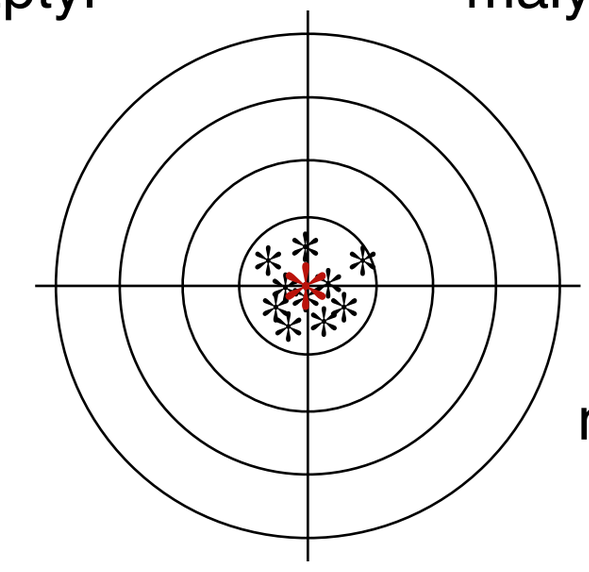
Nevychýlený,
velký rozptyl



Vychýlený,
malý rozptyl



Nevychýlený,
malý rozptyl



Nejlepší
nestranný

Jiří Likeš, Josef Machek: Matematická statistika, Kapitola II

https://sms.nipax.cz/_media/teorie_pravdepodobnosti.pdf

https://meloun.upce.cz/docs/lecture/chemometrics/slidy/32is.pdf

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)**

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

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

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

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

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)**

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

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

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

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

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)

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 $L(\theta; x_1, x_2, \ldots, x_n) = \ln(l(\theta; x_1, x_2, \ldots, x_n))$.

Výsledkem je tzv. ML odhad (MLE)

Metoda maximální věrohodnosti (Maximum likelihood)

Příklad 1: 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:

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

$$-n + \frac{1}{\hat{\lambda}}\sum_{i=1}^{n} k_i = 0 \quad \Rightarrow \quad \boxed{\hat{\lambda} = \frac{\sum_{i=1}^{n} k_i}{n}}$$ = ML odhad λ (MLE λ)

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.

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

$$\mu_1 = E(X) = \mu, \qquad m_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{odtud:} \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu_2 = E(X^2) = \sigma^2 + \mu^2 \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad \text{tedy:} \quad \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

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

Intervalové odhady - Intervaly spolehlivosti

Intervalový odhad je interval $(\hat{\Theta}_L(X), \hat{\Theta}_H(X))$ takový, že

$$P(\Theta \in (\hat{\Theta}_L(X), \hat{\Theta}_H(X))) = 1 - \alpha$$

α - hladina významnosti

1-α - koeficient spolehlivosti

Příklad 3: konstrukce intervalového odhadu střední hodnoty při výběru z normálního rozdělení.

$$\hat{\mu} = \bar{X}_n \qquad \bar{X}_n \sim N(\mu, \sigma^2/n) \qquad Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1)$$

Protože μ ani σ^2, musíme použít t-rozdělení: $$T = \frac{\bar{X} - \mu}{s}\sqrt{n} \sim t(n-1)$$

$$P(\hat{\mu}_L \le \mu \le \hat{\mu}_H) = P(-\hat{\mu}_L \ge -\mu \ge -\hat{\mu}_H) = P(\bar{X} - \hat{\mu}_L \ge \bar{X} - \mu \ge \bar{X} - \hat{\mu}_H)$$

$$= P\left(\frac{\bar{X} - \hat{\mu}_L}{s}\sqrt{n} \ge \frac{\bar{X} - \mu}{s}\sqrt{n} \ge \frac{\bar{X} - \hat{\mu}_H}{s}\sqrt{n}\right)$$

$$= P\left(\frac{\bar{X} - \hat{\mu}_L}{s}\sqrt{n} \ge T \ge \frac{\bar{X} - \hat{\mu}_H}{s}\sqrt{n}\right) = 1 - \alpha = P\big(t_{1-\alpha/2}(n-1) \ge T \ge t_{\alpha/2}(n-1)\big)$$

odtud: $$\frac{\bar{X} - \hat{\mu}_L}{s}\sqrt{n} = t_{1-\alpha/2}(n-1) \qquad \frac{\bar{X} - \hat{\mu}_H}{s}\sqrt{n} = t_{\alpha/2}(n-1)$$

Intervalové odhady - Intervaly spolehlivosti

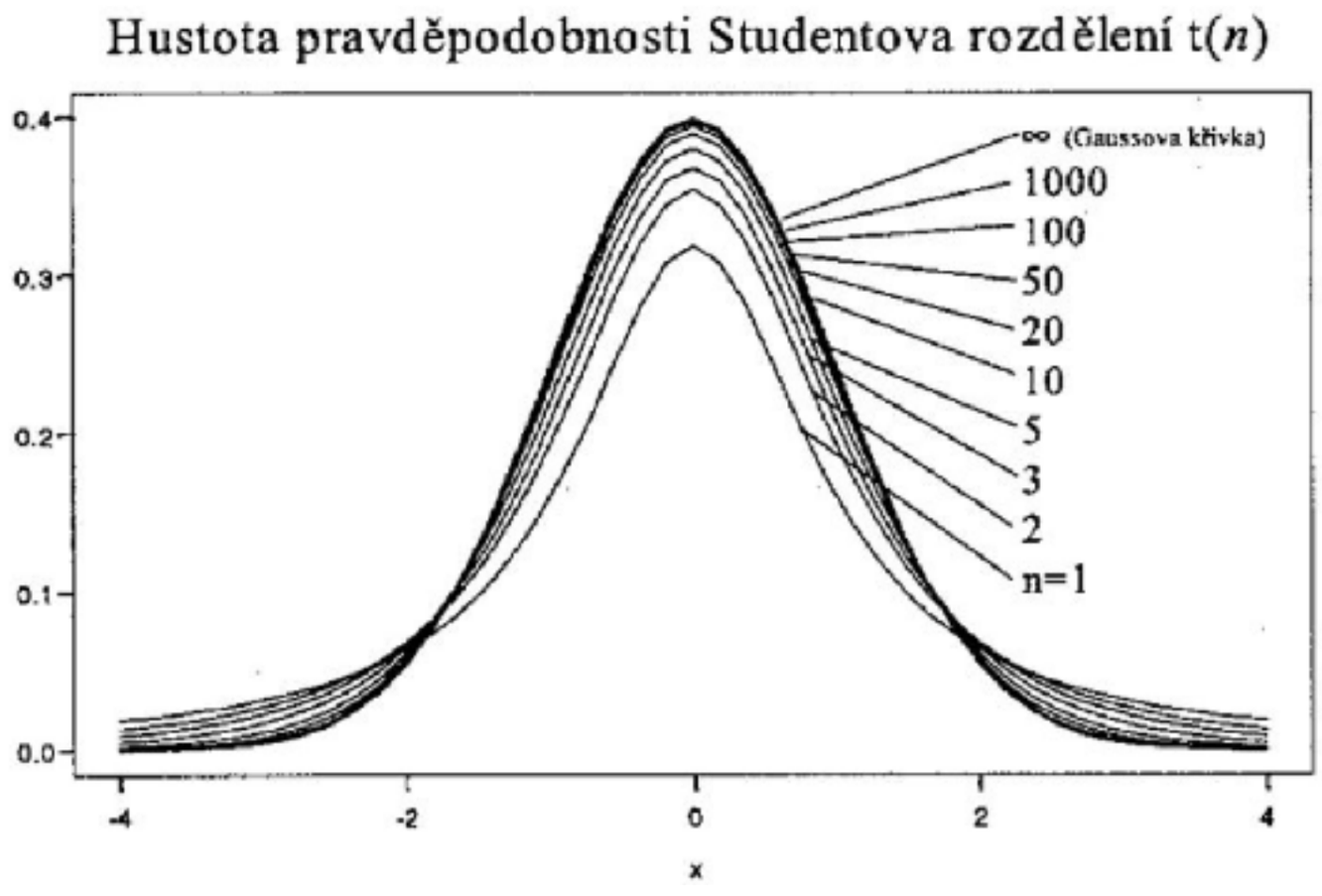
Příklad 3: konstrukce intervalového odhadu střední hodnoty při výběru z normálního rozdělení.

$$\frac{\bar{X} - \hat{\mu}_L}{s}\sqrt{n} = t_{1-\alpha/2}(n-1) \qquad \frac{\bar{X} - \hat{\mu}_H}{s}\sqrt{n} = t_{\alpha/2}(n-1)$$

a tedy: $\hat{\mu}_L = \bar{X} - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)$ $\hat{\mu}_H = \bar{X} - \frac{s}{\sqrt{n}} t_{\alpha/2}(n-1)$

ze symetrie t-rozdělení víme, že $t_{\alpha/2}(n-1) = -t_{1-\alpha/2}(n-1)$, a tedy

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



Intervalové odhady - Intervaly spolehlivosti

Příklad 3: konstrukce intervalového odhadu střední hodnoty při výběru z normálního rozdělení.

$$\frac{\bar{X} - \hat{\mu}_L}{s}\sqrt{n} = t_{1-\alpha/2}(n-1) \qquad \frac{\bar{X} - \hat{\mu}_H}{s}\sqrt{n} = t_{\alpha/2}(n-1)$$

a tedy: $\hat{\mu}_L = \bar{X} - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)$ $\hat{\mu}_H = \bar{X} - \frac{s}{\sqrt{n}} t_{\alpha/2}(n-1)$

ze symetrie t-rozdělení víme, že $t_{\alpha/2}(n-1) = -t_{1-\alpha/2}(n-1)$, a tedy

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

Označíme-li $\frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = SE$, potom lze intervalový odhad psát ve tvaru

$$(\hat{\mu}_L, \hat{\mu}_U) = (\bar{X} - SE, \bar{X} + SE)$$

$$P(\bar{X} - SE \leq \mu \leq \bar{X} + SE) = 1 - \alpha$$

SE = standardní chyba

Poznámka: Pokud rozptyl σ^2 známe, potom můžeme použít kvantily standardního normálního rozdělení a standardní chyba bude mít tvar $SE = \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$.

Intervalové odhady - Intervaly spolehlivosti

Příklad 4: Test spotřeby automobilu

Deklarovaná průměrná spotřeba je 13,5 l/100 km

Otázka: **Odpovídá deklarovaná průměrná spotřeba naměřeným datům?**

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

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

Zvolíme α=0,05 => $t_{0,975}(13) = 216036866$

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

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

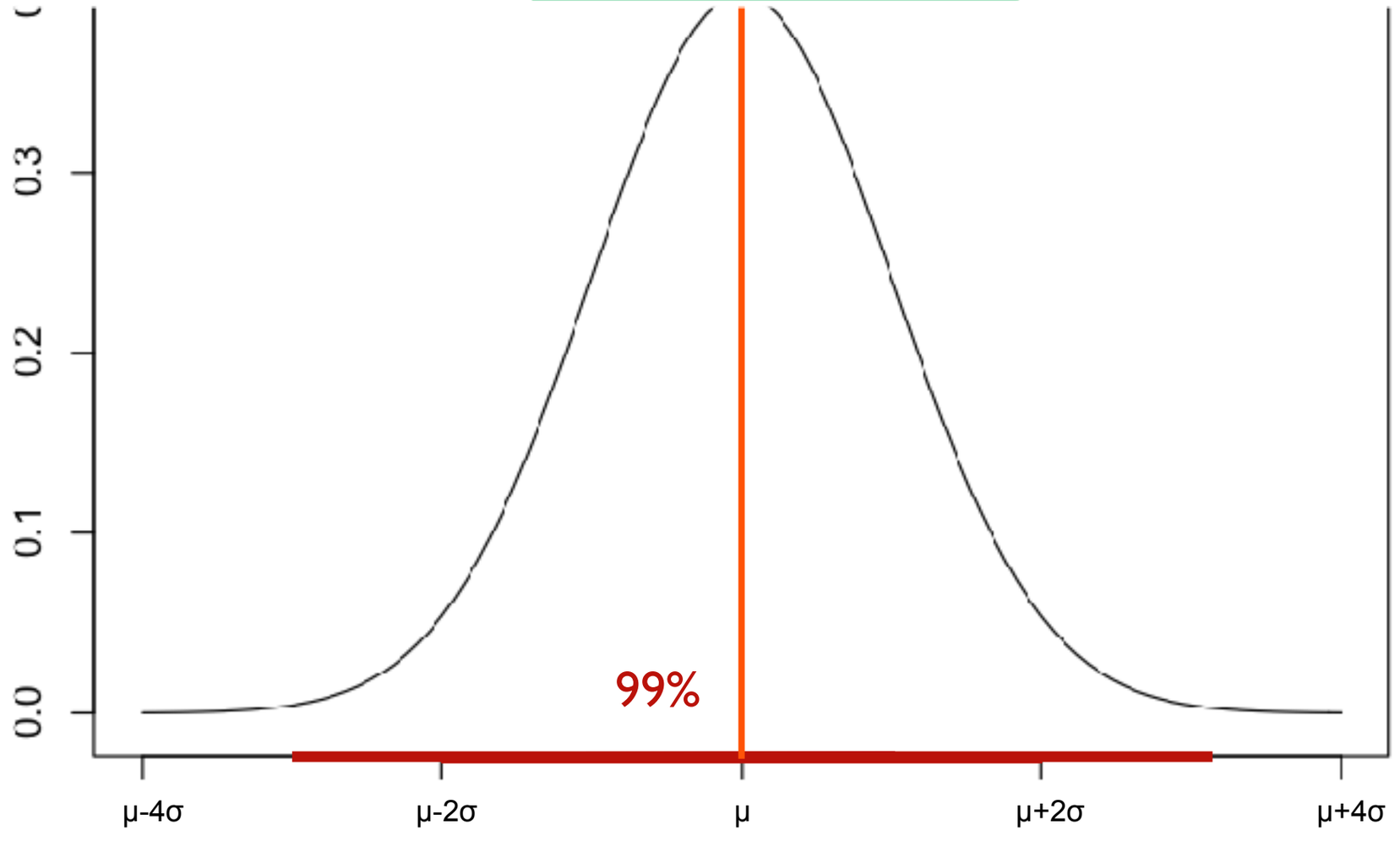
T.INV(0,975;13)=
=2,16036866

Intervalové odhady - Intervaly spolehlivosti

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

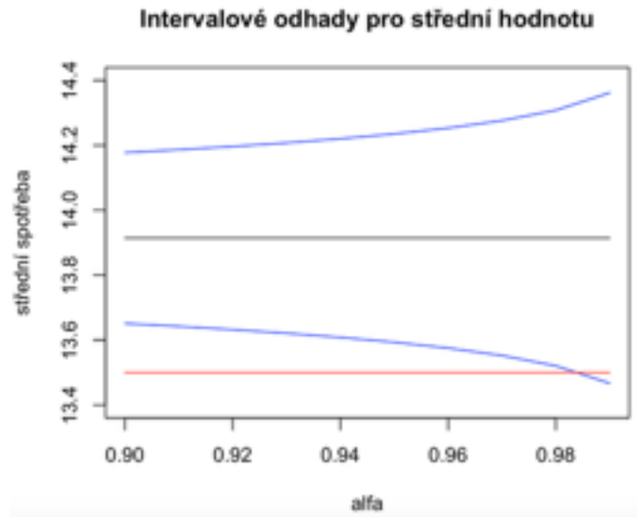
Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Int. spolehl. -95,000%	Int. spolehl. +95,000%	Referenční konstanta	t	SV	p
spotřeba l/100	13,91429	0,555888	14	0,148567	13,59333	14,23525	12,50000	9,519502	13	0,000000

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

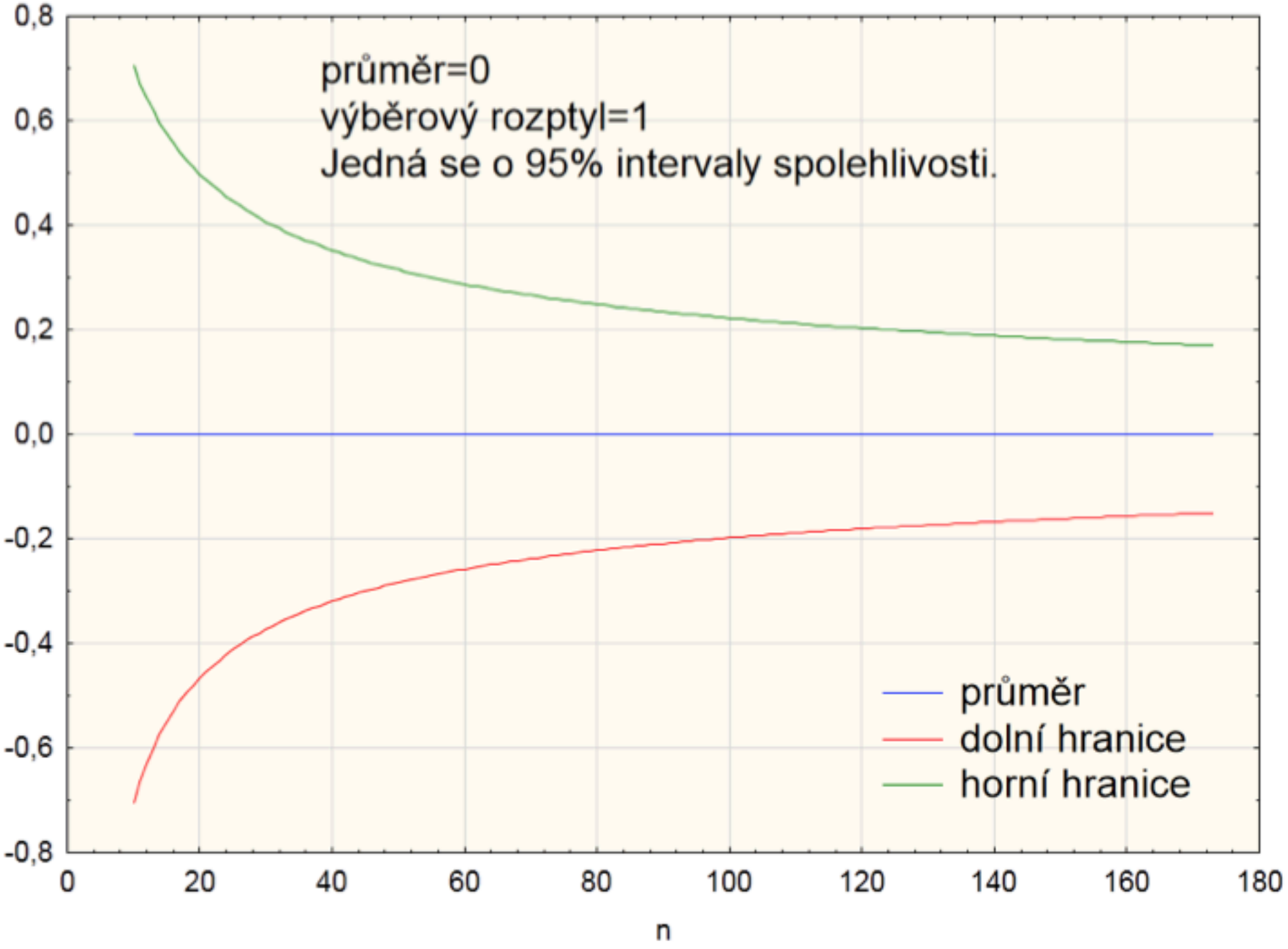


Intervalové odhady - Intervaly spolehlivosti

```
spotreba <- data.matrix(read.table(„spotreba.txt"))     #načteme data
for (i in 1:10) a[i]=1-i/100                            #hladina významnosti α
clm=matrix(0, nrow=10, ncol=2)                          # sem uložíme meze int. spolehlivosti
for(i in 1:10) {
  a[i]=1-i/100                                          #měníme hladinu významnosti od  0,99 do 0,9
  t=t.test(spotreba, conf.level=a[i])                   #funkce t.test spočte intervalové odhady,
  clm[i,]=c(t$conf.int)                                 #které ukládáme do matice clm
  }
plot(a,clm[,1], ylim=c(13.4, 14.4), type="l",           # vykreslíme graf s dolní mezí
     main="Intervalové odhady pro střední hodnotu",
     xlab="alfa", ylab="střední spotřeba", col="blue")
lines(a,clm[,2], col=„blue")                            # doplníme horní mez
for (i in 1:10) d[i]=13.5                               # deklarovaná spotřeba, kterou
lines(a,d, col=„red")                                   # zde vykreslíme jako červenou linku
for (i in 1:10) m[i]=mean(spotreba)                     # průměrnou naměřenou spotřebu
lines(a,m)                                              # zde vykreslíme jako černou linku
```

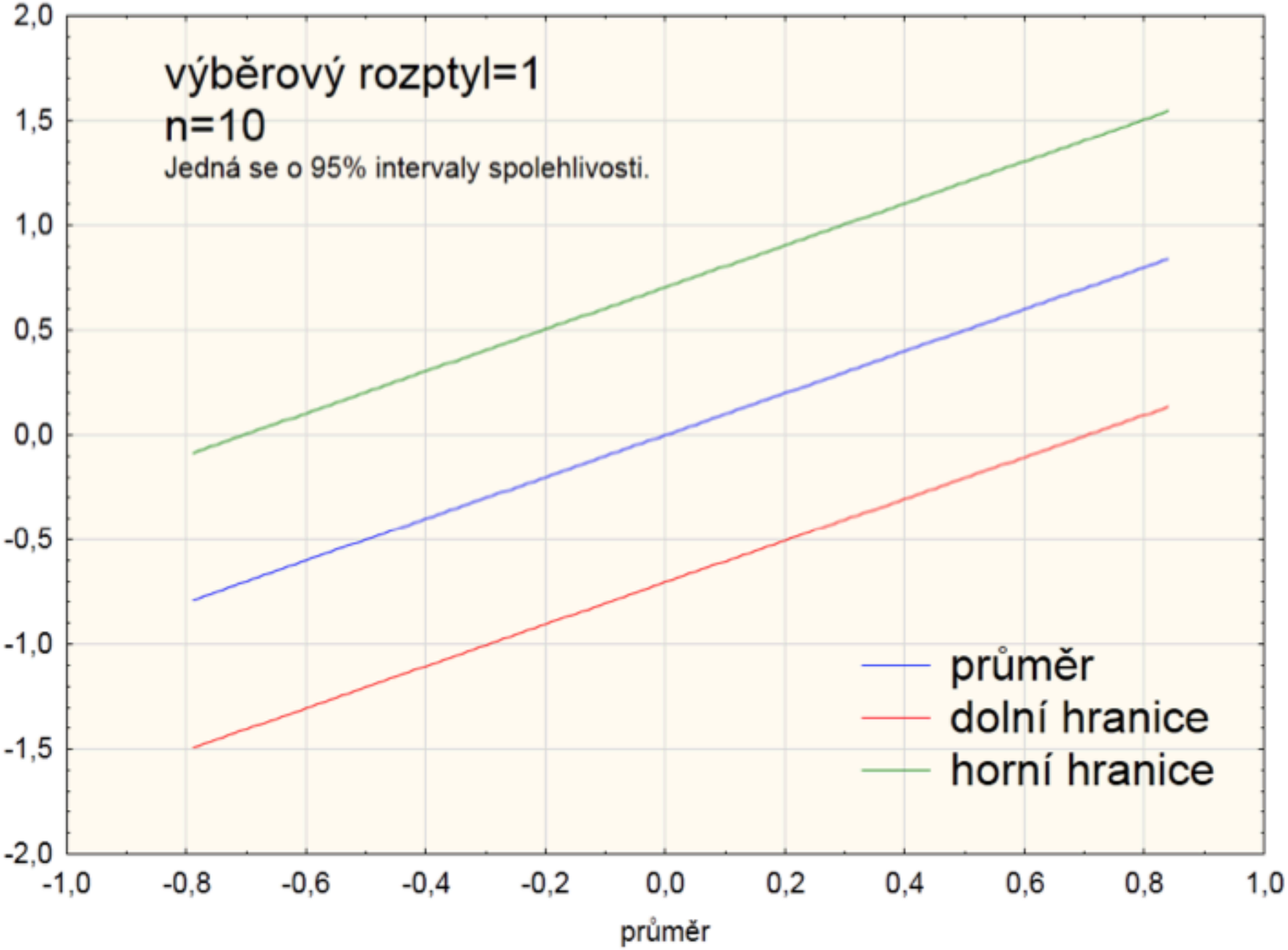


Intervalové odhady - Intervaly spolehlivosti



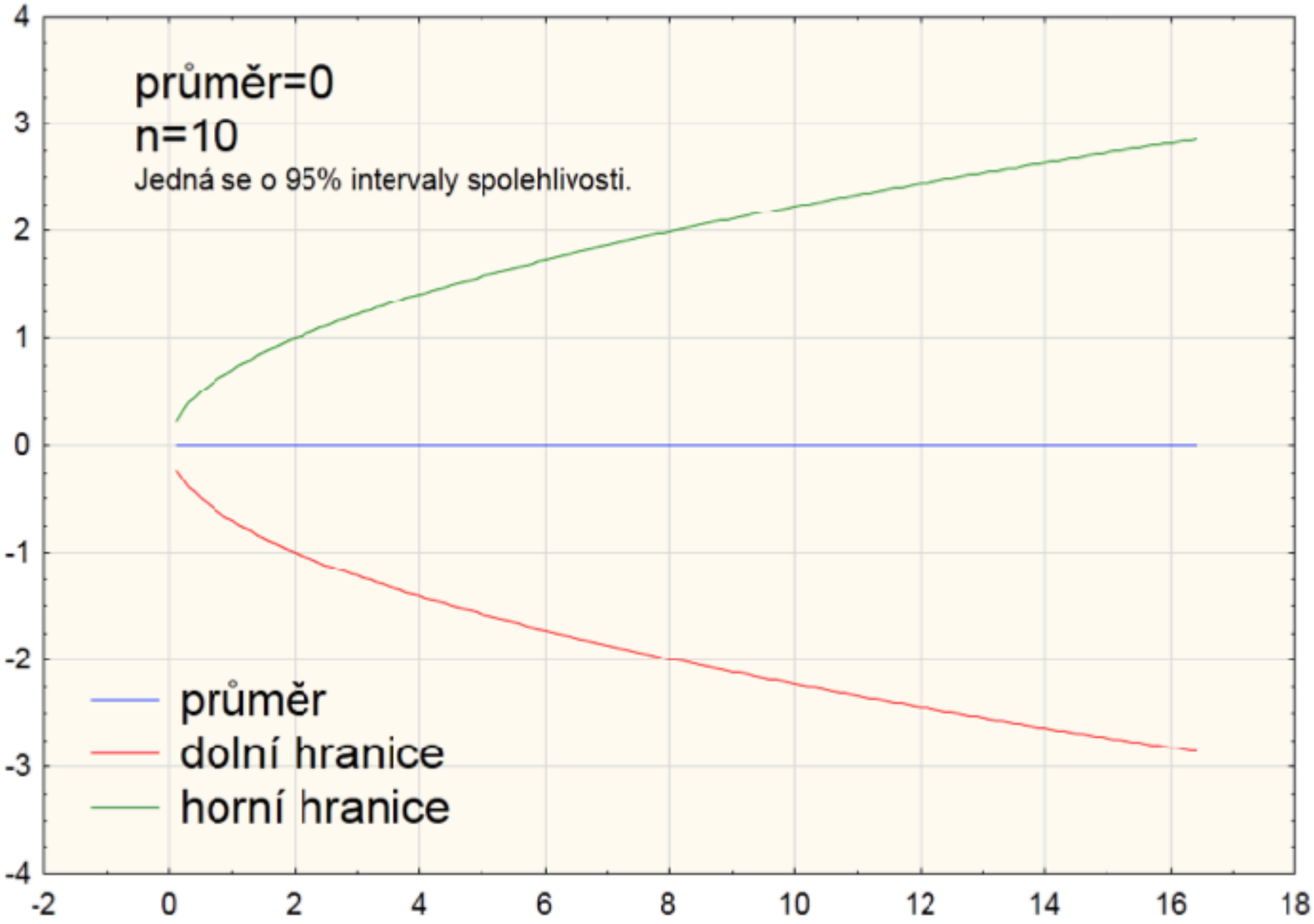
Intervalové odhady - Intervaly spolehlivosti

Vliv změny odhadovaného parametru na šířku intervalu spolehlivosti



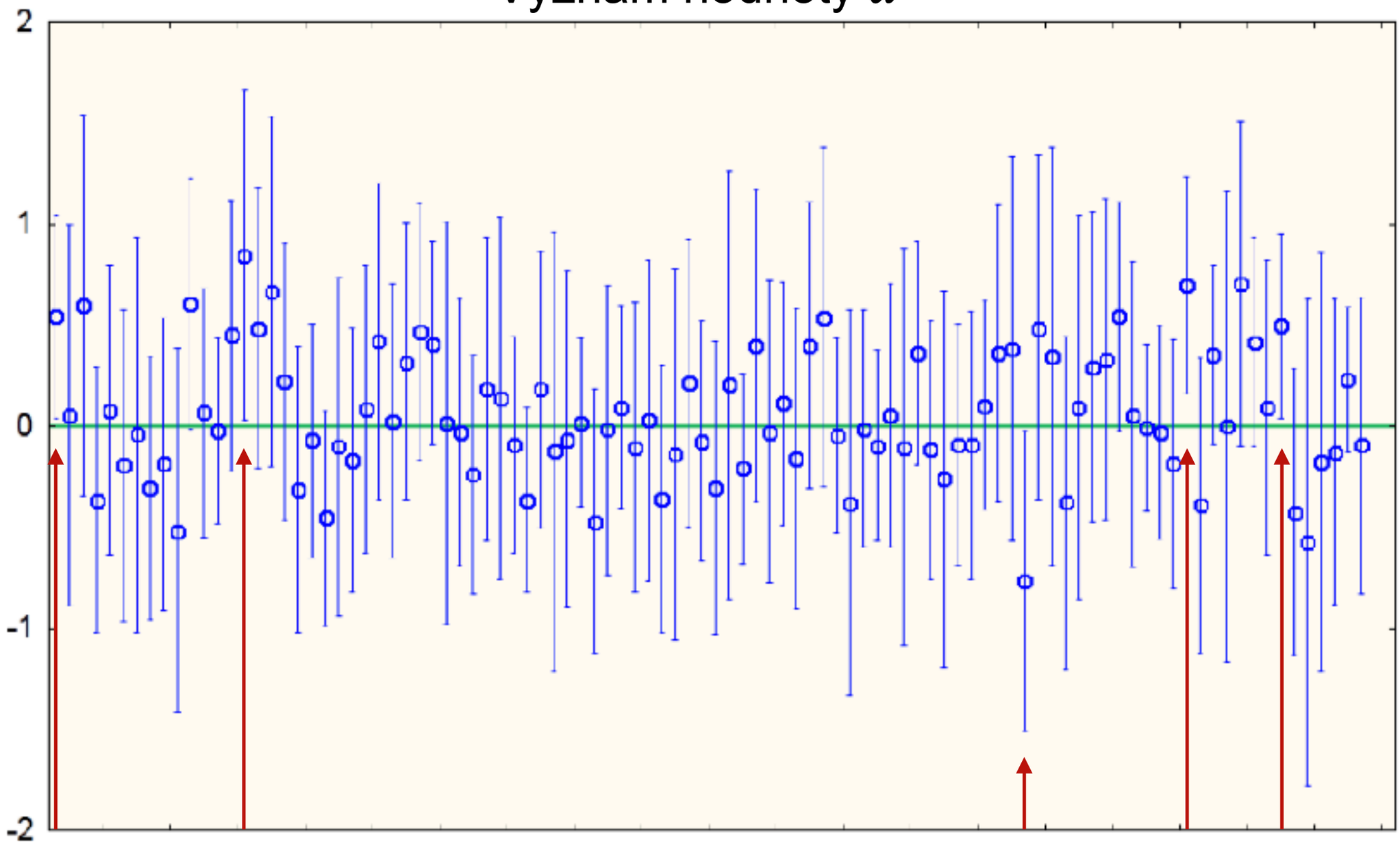
Intervalové odhady - Intervaly spolehlivosti

Vliv změny směrodatné odchylky na šířku intervalu spolehlivosti



Intervalové odhady - Intervaly spolehlivosti

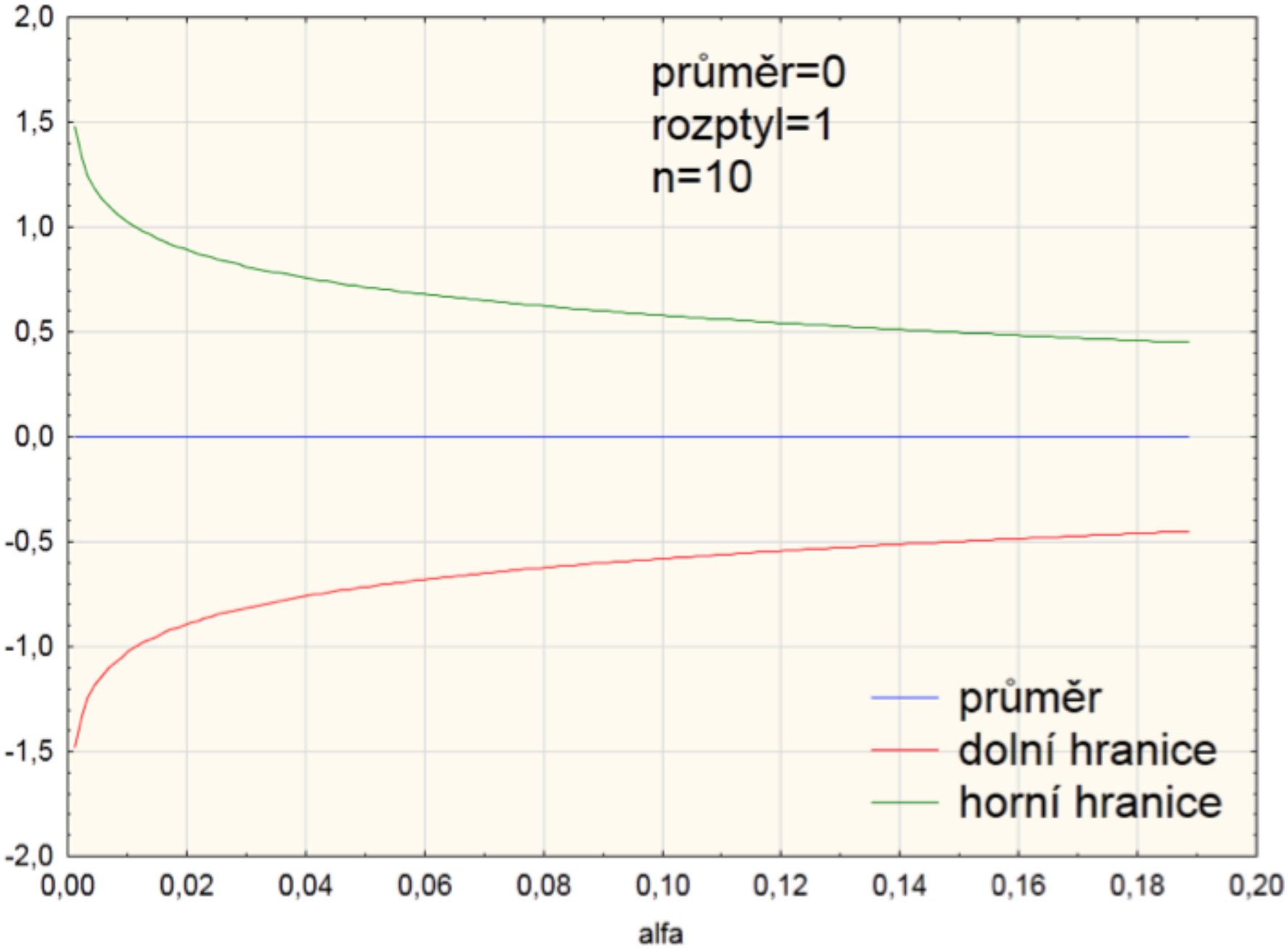
Význam hodnoty α



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

Intervalové odhady - Intervaly spolehlivosti

Vliv změny hodnoty α na šířku intervalu spolehlivosti



Intervalové odhady - Intervaly spolehlivosti

Příklad 5: konstrukce intervalového odhadu rozptylu při výběru z normálního rozdělení.

Zde využijeme znalost toho, že $$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

Podobnými úpravami jako v případě střední hodnoty se dostaneme k intervalovému odhadu ve tvaru

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

Intervalové odhady - Intervaly spolehlivosti

Příklad 6: 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?

X_i : i-tý dotázaný bude volit SMPvMZ = 1, nebude volit SMPvMZ = 0,

Y = počet těch, kteří budou volit SMPvMZ = $\sum_{i=1}^{100} X_i \sim Bin(100, p)$

Y lze aproximovat rozdělením $N(np, np(1-p))$

$\bar{X}$ má potom také přibližně normální rozdělení $N\left(p, \frac{p(1-p)}{n}\right)$

Odhad pravděpodobnosti p lze tedy chápat jako odhad střední hodnoty $\bar{X}$.

Bodový odhad pravděpodobnosti p je $\bar{X} = 0,43$. Intervalový odhad je potom

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

$u_{0,975} = 1,96$ Tedy intervalový odhad je (0,33; 0,53) => SMPvMZ má statisticky významnou šanci na hladině významnosti 5 % získat nadpoloviční většinu.