Induktivní statistika

Odhady

Odhady

- bodové odhady
- intervalové odhady
- konstrukce intervalu spolehlivosti pro průměr
- odhady podílů (kategoriální proměnné)

Odhady

- v příkladech v předchozích přednáškách jsme znali hodnoty průměru a rozptylu populace
- obvykle tomu ale bývá přesně naopak: **známe hodnoty (statistiky) výběru a neznáme hodnoty (parametry) populace**
- ty chceme z výběru **odhadnout**

Odhady

- 2 typy odhadů: bodové a intervalové
- **bodový odhad**: použijeme průměr vzorku a odhadneme, že se rovná průměru populace

Bodový odhad

- bodový odhad je problematický v tom, že dva různé výběry nám mohou dát dva různé odhady
- bodový odhad **neobsahuje** žádnou **informaci** o jeho **přesnosti** či **spolehlivosti**
- na čem závisí přesnost odhadu?

Bodový odhad

přesnost odhadu závisí na dvou charakteristikách

- **velikost výběru** (čím větší n, tím menší výběrová chyba)
- **variabilita hodnot v populaci** (čím vyšší, tím vyšší i výběrová chyba)

Intervalový odhad

- poskytuje rozsah (interval) hodnot, který s určitou pravděpodobností obsahuje hledanou hodnotu parametru

Intervalový odhad

je založen na:

- bodovém odhadu
- velikosti výběru
- variabilitě znaku v populaci (známé nebo rovněž odhadované)

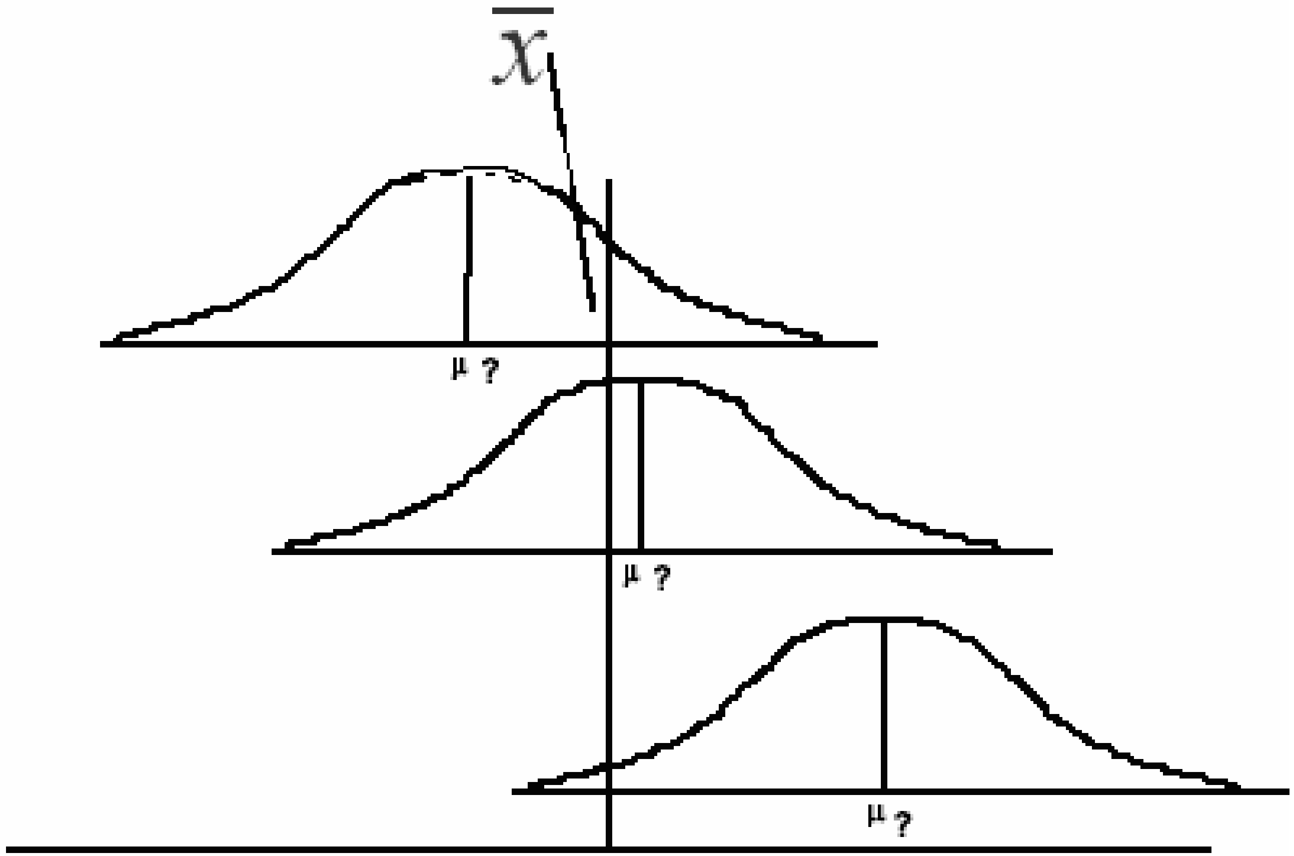
Intervalový odhad

- ptáme se: **jaká je hodnota μ ?**

Intervalový odhad

- ptáme se: **jaká je hodnota μ ?**
- výběrový průměr určité hodnoty může pocházet z populací o různých průměrech
- proto **nemůžeme jednoznačně určit hodnotu μ**

Intervalový odhad



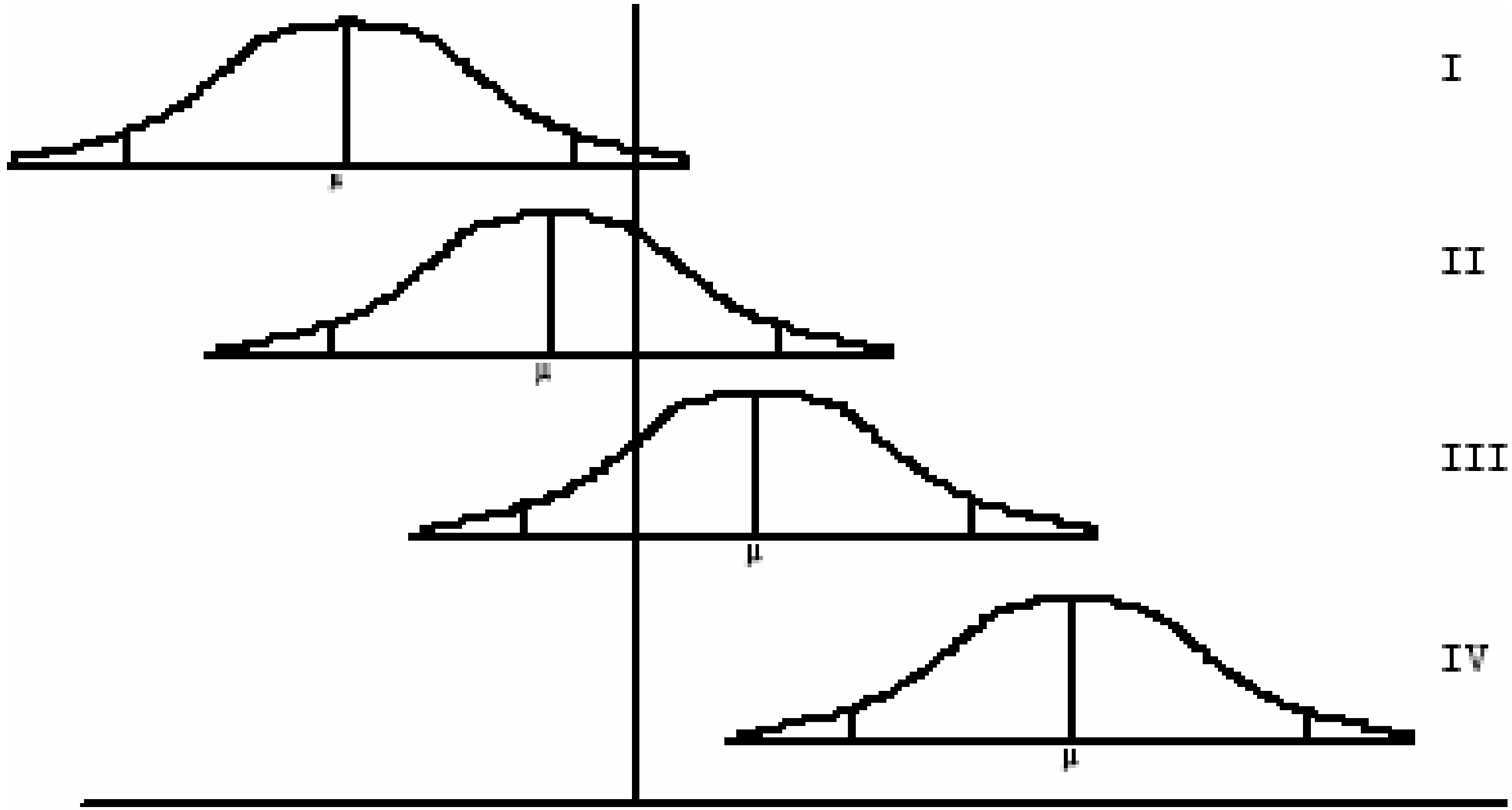
Intervalový odhad

- takže se místo toho snažíme určit, jaký je **možný rozsah hodnot μ**
- jaké populace (tj. s jakou hodnotou průměru) by mohly být pravděpodobným zdrojem našeho vzorku?

Intervalové odhady

- ze které populace nejpravděpodobněji pochází výběr, jehož průměr je v následujícím grafu naznačen svislou čarou?

RVP pro populace I-IV



Intervalové odhady

- výběr pochází
  - nejpravděpodobněji z populace II nebo III
  - méně pravděpodobně z populace I
  - a velmi málo pravděpodobně z populace IV

Intervalové odhady

- intervalový odhad spočívá v konstrukci tzv. **intervalu spolehlivosti** (confidence interval) = rozsahu hodnot, ve kterém s určitou pravděpodobností leží průměr populace

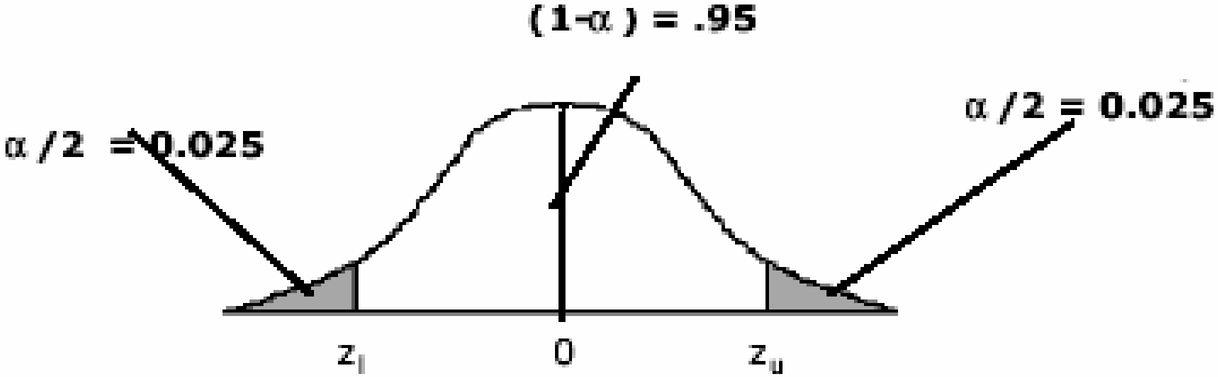
Interval spolehlivosti

- nejprve je třeba si **stanovit tuto pravděpodobnost** – tj. úroveň přesnosti(spolehlivosti);
- obvyklá je např. **95%** - snažíme se najít interval hodnot, ve kterém s 95% pravděpodobností leží průměr populace
- pak jde o tzv. **95% interval spolehlivosti**

Interval spolehlivosti

- poté **najít hodnotu *z* pro tuto pravděpodobnost** – tj. rozsah, ve kterém bude ležet středních 95% hodnot (výběrových průměrů)
- 2,5% na každé straně rozdělení

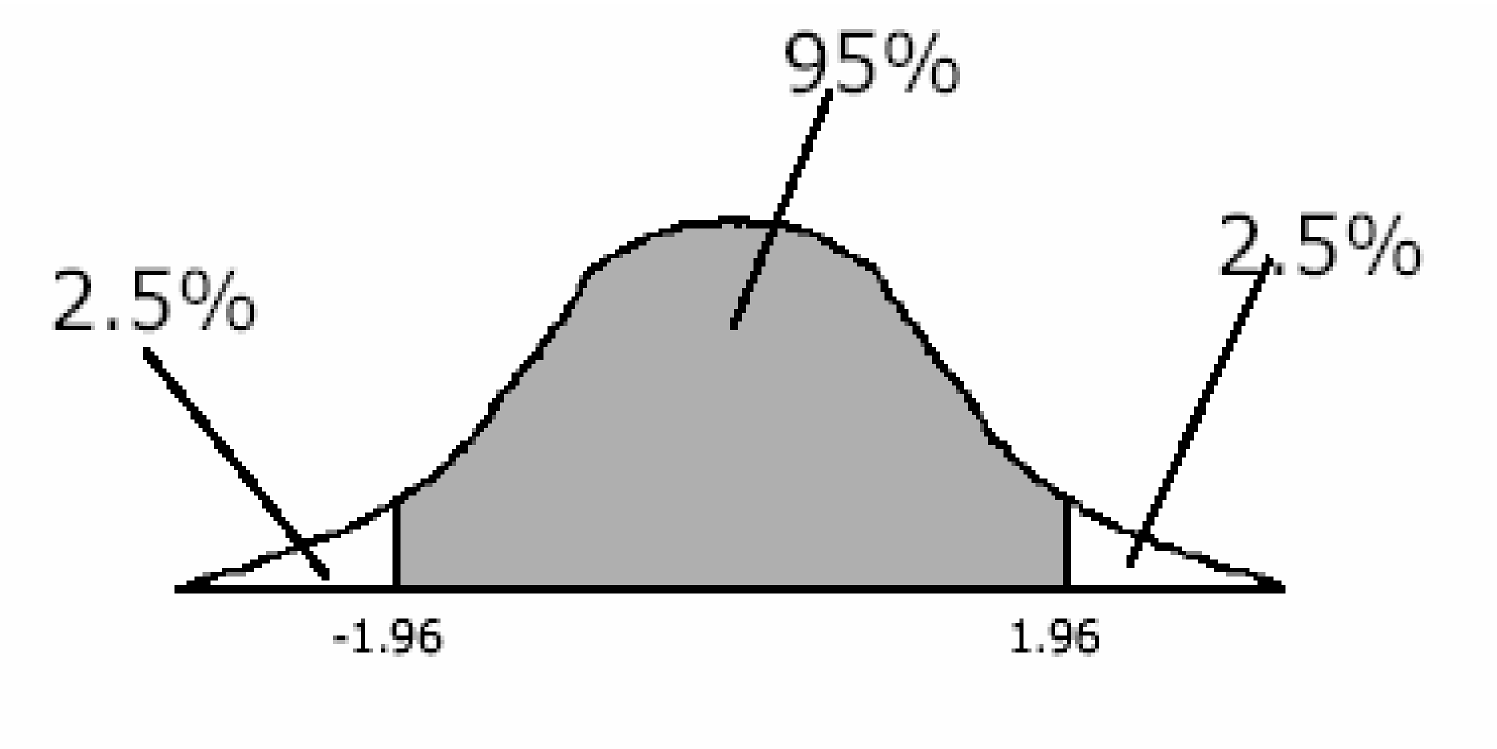
Interval spolehlivosti



Interval spolehlivosti

- tomu odpovídají hodnoty
  z=-1,96
  z=1,96

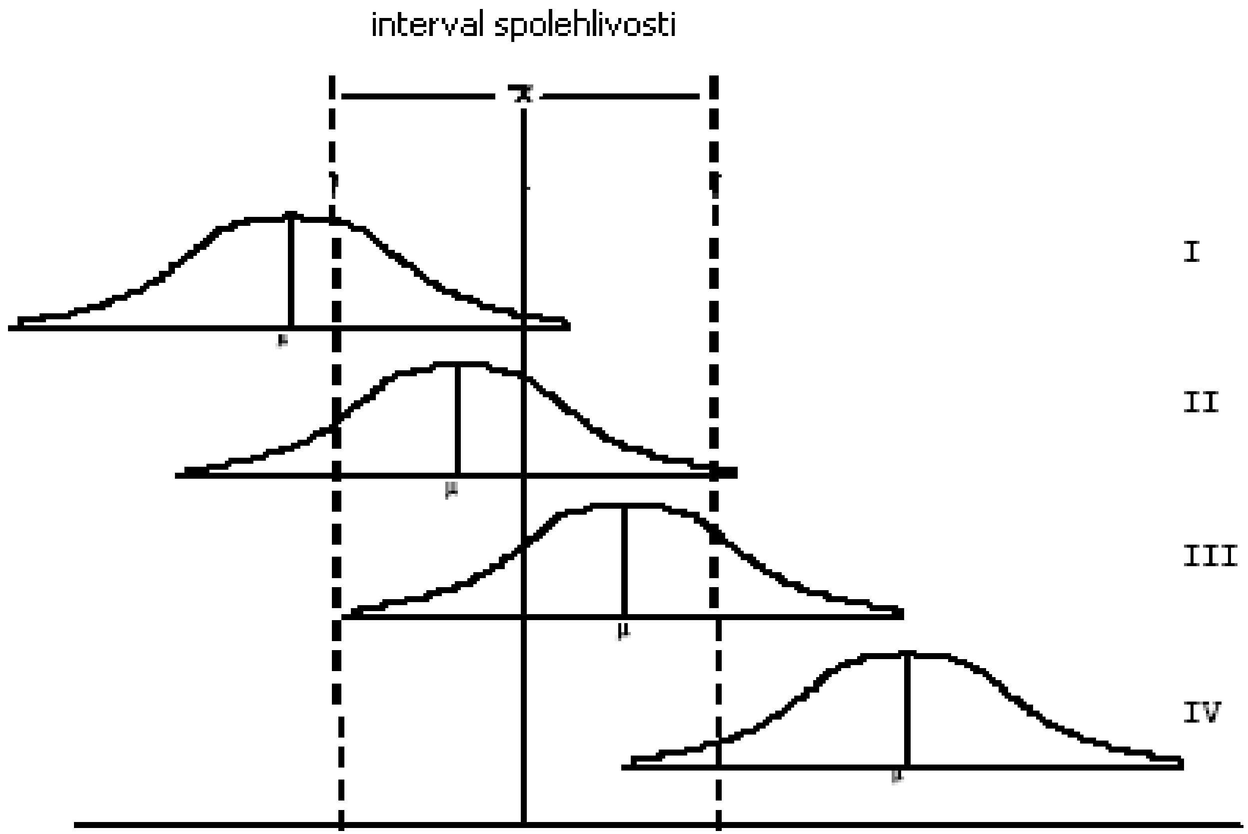
Interval spolehlivosti



Interval spolehlivosti - výpočet

$$\overline{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Interval spolehlivosti



Interval spolehlivosti

- interpretace intervalu spolehlivosti:

pokud bychom z populace vybrali 100 náhodných výběrů o velikosti n a pro každý z nich sestrojili tento interval, 95 intervalů by obsahovalo průměr populace a 5 nikoliv

- opatrně můžeme říct : máme 95% pravděpodobnost, že se v tomto intervalu nachází průměr populace

Interval spolehlivosti

- oblíbený omyl:
  - v 95% intervalu spolehlivosti leží 95% hodnot populace (NEPLATÍ!)
- kromě 95% intervalu spolehlivosti se používá také např. 99% a 90% pravděpodobnost

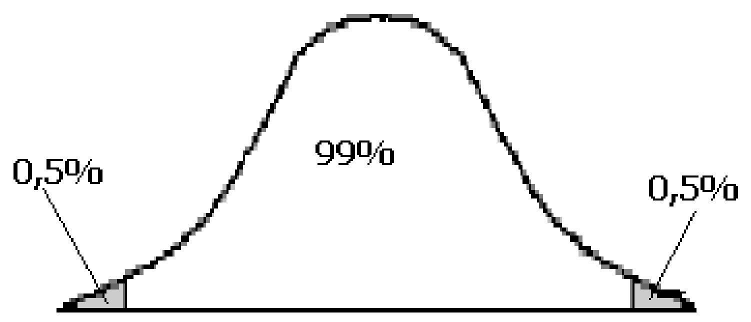
Příklad

- náhodný výběr 36 dětí romského původu, průměrné IQ vzorku = 96
- na základě tohoto zjištění odhadněte průměrné IQ populace romských dětí (sestavte 99% interval spolehlivosti)

Příklad

- Postup:
  - bodový odhad: μ=96
  - výpočet výběrové chyby (směrodatné odchylky RVP):
    $\sigma/\sqrt{n} = 15/\sqrt{36} = 15/6 = 2{,}5$
  - stanovení úrovně spolehlivosti: 99%
  - najít hodnotu *z* pro 99% pravděpodobnost

Příklad



Příklad

- v tabulce normálního rozdělení najdeme hodnoty *z*
- hodnoty *z* pro 99% : 2,57 a -2,57

Příklad

- k výběrovému průměru přičteme (pro horní hranici intervalu) a odečteme (pro spodní hranici) výběrovou chybu, vynásobenou hodnotou *z*

Příklad

-

$$CI(\mu) = \bar{x} \pm z * \sigma/\sqrt{n}$$

CI (μ) = 96 + 2,57 * 2,5 = **102,43**

CI (μ) = 96 - 2,57 * 2,5 = **89,58**

99% interval spolehlivosti je 89,6 – 102,4

Interval spolehlivosti

- **hodnoty *z*** pro nejčastěji užívané pravděpodobnosti:
  - 90% (zbývá 5% + 5%) z= +/- 1,645
  - 95% (zbývá 2,5% + 2,5%) z= +/- 1,96
  - 99% (zbývá 0,5% + 0,5%) z= +/- 2,57

Odhady podílů

- u kategoriálních proměnných nemůžeme počítat průměry
- odhadujeme proto **podíly** jednotlivých kategorií proměnné

Odhady podílů

- např. podíl kuřáků v populaci českých adolescentů
- podíl pacientů s rakovinou plic, kteří přežijí 5 let od diagnózy
- podíl chlapců mezi dětmi s poruchou pozornosti

Odhady podílů

- pokud zkoumáme místo celé populace pouze výběr z ní, nezajímá nás tolik, jaký je podíl kategorií proměnné ve výběru (četnost **p**)
- ale spíše jaký je skutečný podíl v populaci – četnost $\boldsymbol{\pi}$

Odhady podílů

- při dostatečně velkém *n* platí i pro rozdělení podílů centrální limitní věta
- rozdělení výběrových podílů je normální rozdělení, s **průměrnou četností π** a směrodatnou odchylkou (výběrovou chybou)

$$SE = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Příklad 4

- chceme zjistit, jaká je podpora zachování hlavního nádraží v Brně na stávajícím místě
- náhodný výběr z populace brněnských voličů (n=1000 osob)
- 585 osob se vyjádřilo <u>pro</u> (p=0,585)
- odhadněte s 95% spolehlivostí podporu zachování nádraží v populaci brněnských voličů

Odhady podílů

- interval spolehlivosti pro podíly se spočítá podobně jako pro průměry:

$$p \pm z_{1-\alpha/2} \sigma_p$$

Odhady podílů

- nemůžeme však spočítat výběrovou chybu, protože neznáme π
- v tomto případě je však možné dosadit místo toho *p* a přitom použít normální rozdělení (pokud je $n>30$)
- pokud je $n<30$, pak dosadíme místo π hodnotu 0,5

Příklad 4

- p=0,585
- z=1,96
- SE(p)=$\sqrt{[0,585(1-0,585)/1000]}$

=0,156

interval spolehlivosti

0.585 ± 1.96(0.0156)

0.585 ± 0,0305

--- **přesnost odhadu je ± 3%**

Příklad 4

- s 95% pravděpodobností je podíl osob souhlasících se zachováním hlavního nádraží na stávajícím místě **mezi 55.4% a 61.6%**
- tj. máme 95% pravděpodobnost, že kdyby se v době průzkumu hlasovalo, bude většina pro

Odhady podílů

vztah mezi velikostí vzorku a přesností odhadu

n=100	± 10%
n=200	± 7%
n=400	± 5%
n=1000	± 3%
n=2400	± 2%
n=9600	± 1%

Odhady podílů

- požadovaná velikost vzorku roste mnohem rychleji než spolehlivost odhadu (pro zdvojnásobení spolehlivosti je nutné asi čtyřnásobně zvětšit vzorek)
- důležité při plánování výzkumu – jakou přesnost potřebujeme? jaké budou náklady?
- podobný vztah platí pro odhad průměrů

Kontrolní otázky

- 2 typy odhadů
- na čem závisí šířka intervalu spolehlivosti? (*není nutno znát zpaměti vzorce, ale je třeba chápat princip výpočtu)*
- vztah velikosti výběru a spolehlivosti odhadu

Literatura

- Hendl: kapitoly 4 a 5