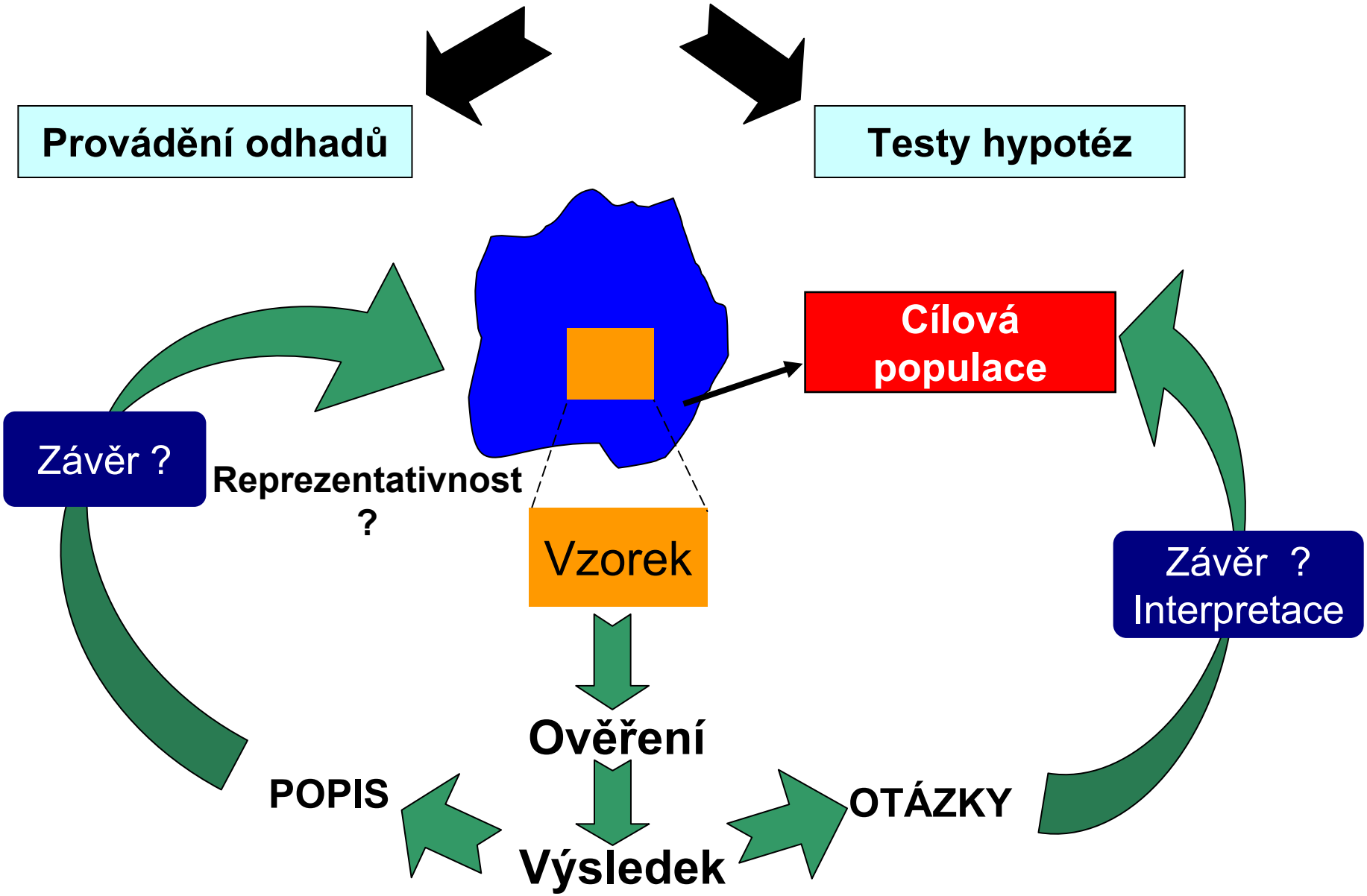
8. Provádění odhadů

Statistika v průzkumném studiu



INTERVAL SPOLEHLIVOSTI
- velmi užitečná míra věrohodnosti odhadů -

ODHADY

- Bodové
  - Číslo (chyba)
    (Odhad parametru)
- Intervalové
  - **Interval pravděpodobných hodnot**
    - Spolehlivost
      (Pravděpodobnostní interpretace)

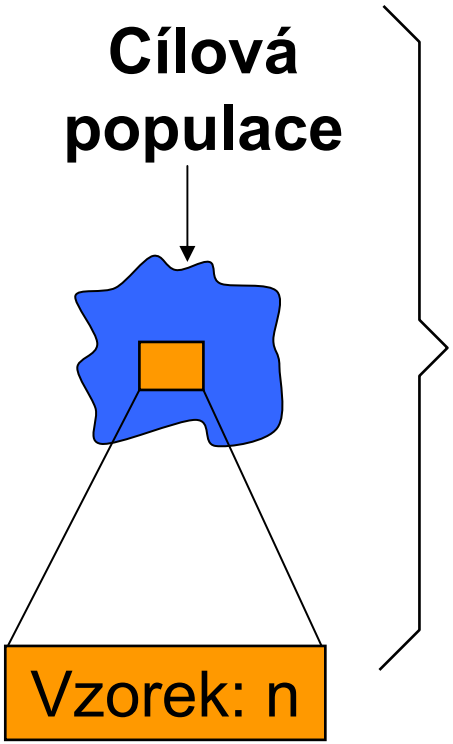
Obecný tvar:

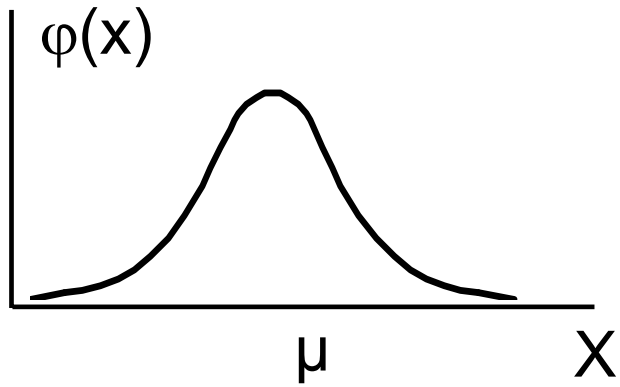
$$P\ (L_1 < \text{Odhad} < L_2) \geq 1 - \alpha/2$$

Odhadovaný parametr ± Kvantil modelového rozložení . SE (odhadu)

K_V pro $(1 - \alpha/2)$

NORMÁLNÍ ROZLOŽENÍ
- model pro odhad průměru -





Prezentace

n; $\bar{x}$; s

n; $\bar{x}$; $\frac{s}{\sqrt{n}}$

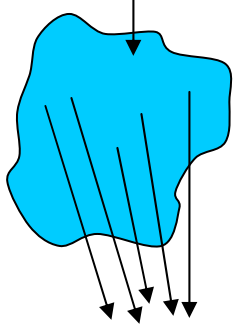
n; $\bar{x}$; c

n; $\bar{x}$; Interval spolehlivost i pro odhad průměru

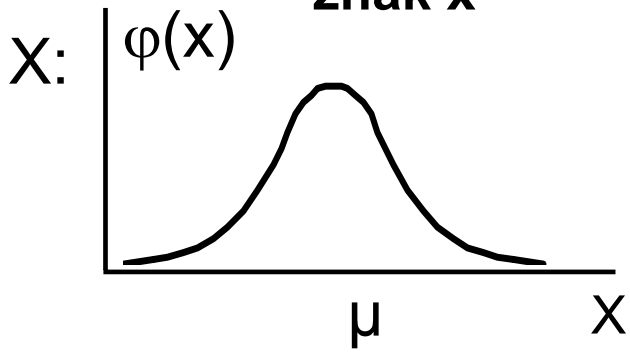
$\bar{x}$ ..... odhad průměru

NORMÁLNÍ ROZLOŽENÍ
- odhad průměru je rovněž normálně rozložen -

Cílová populace



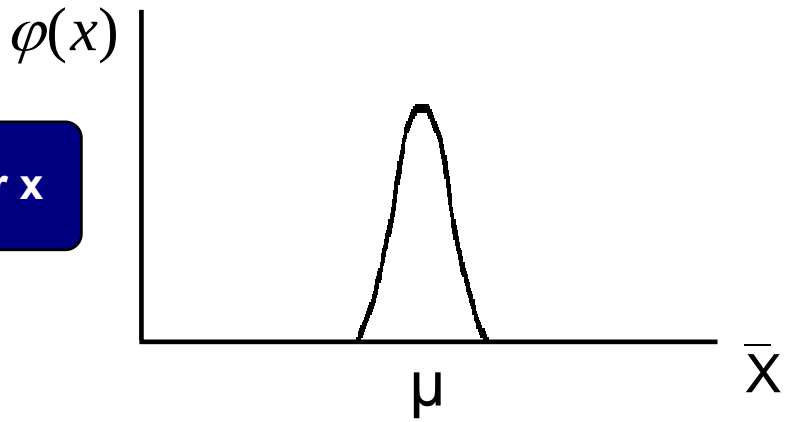
znak x



x: μ ± 3s

Náhodné výběry o n = 100

$\bar{X}1 \quad \bar{X}2 \quad \bar{X}3 \quad \bar{X}4 \ldots. \quad \bar{X}i$



průměr x

$$\mu \pm 3 . \frac{s}{\sqrt{n}}$$

$\frac{s}{\sqrt{n}}$ ~ Standardní chyba odhadu průměru

ODHAD PRŮMĚRU
I. Vztahy

Bodový

$$\bar{x}; \left(\frac{s}{\sqrt{n}} \right)$$

Intervalový

$$\bar{x} - t_{1-\alpha/2}^{(\nu=n-1)} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}^{(\nu=n-1)} \cdot \frac{s}{\sqrt{n}}$$

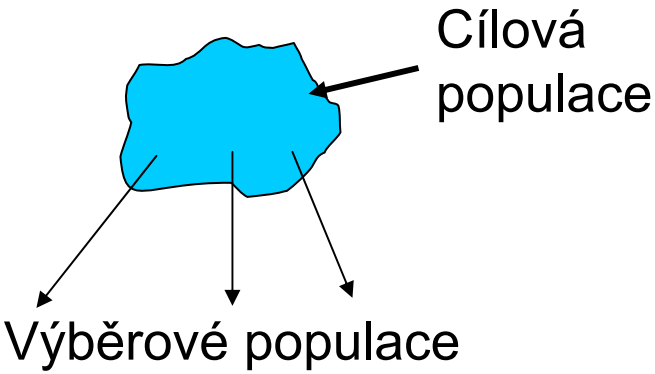
$$\mu : \bar{x} \pm t_{1-\alpha/2}^{(\nu=n-1)} \cdot \frac{s}{\sqrt{n}}$$

$$\mu : \bar{x} \pm t_{1-\alpha/2}^{(\nu=n-1)} \cdot s_{\bar{x}}$$

t ... příslušný kvantil Studentova rozložení
1 - α ... spolehlivost hodnoceného intervalu

Interval spolehlivosti odhadu průměru je pouze informací o přesnosti tohoto odhadu

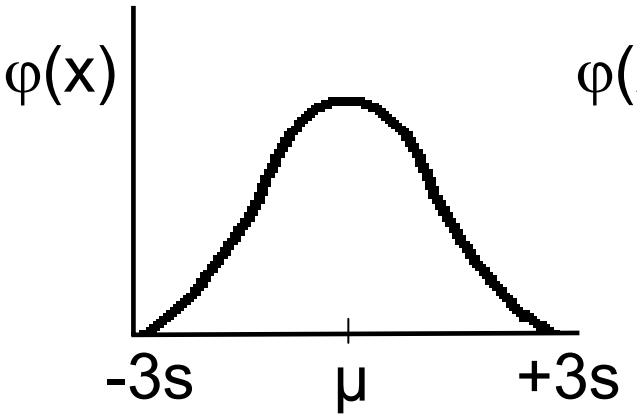
Interval spolehlivosti je hodnocen pro (1 - α) procentní spolehlivost



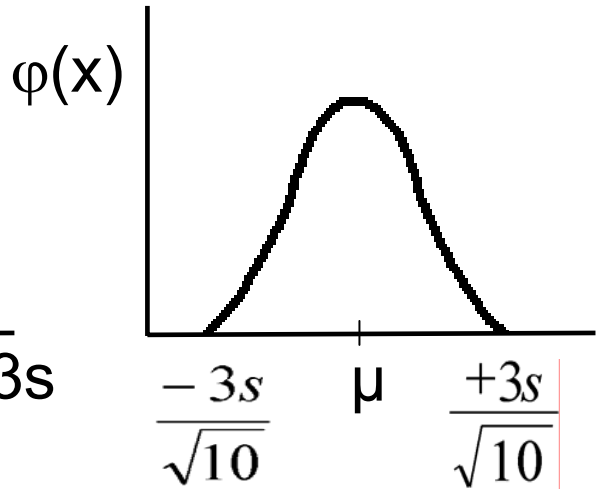
Šířku intervalu určuje:

a) velikost vzorku
b) rozptyl (variabilita) vzorku
c) požadovaná spolehlivost

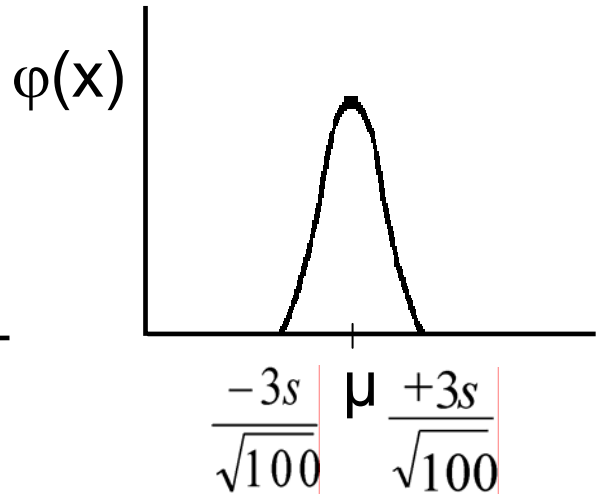
Původní proměnná x



Výběr n=10 pro odhad průměru



Výběr n=100 pro odhad průměru



ODHAD PRŮMĚRU
II. Příklad

X: Cena výrobku v n = 21 obchodech

Data:

$$n = 21; \bar{x} = 3{,}58; s^2 = 0{,}12$$

$$s_{\bar{x}} = \sqrt{0{,}12/21} = 0{,}075$$

95% Interval spolehlivosti:

$$t_{1-\alpha/2}^{(u = n-1)} = t_{0{,}975}^{(20)} = 2{,}086$$

$$\mu : \bar{x} \pm 2{,}086 \; .s_{\bar{x}}$$

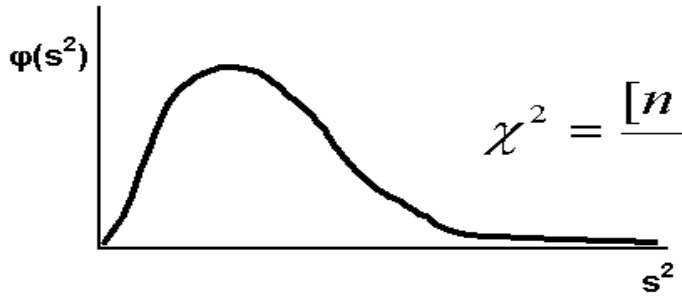
$$3{,}58 - 2{,}086.0{,}075 \le \mu \le 3{,}58 + 2{,}086.0{,}075$$

$$3{,}423 \le \mu \le 3{,}737$$

$$P\ (3{,}423 \le \mu \le 3{,}737) \ge 0{,}95$$

Interval spolehlivosti pro odhad rozptylu

$s^2 \sim \sigma^2$ pro velká n



$$\chi^2 = \frac{[n-1]s^2}{\sigma^2}$$

Interval spolehlivosti

a) pro σ^2 : $\frac{(n-1)s^2}{\chi^2_{\alpha/2}{}^{(n-1)}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}{}^{(n-1)}}$

b) pro σ : $\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}{}^{(n-1)}}} \le \sigma \le \sqrt{\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}{}^{(n-1)}}}$

c) pro $\sigma/\sqrt{\text{n}}$: $\sqrt{\frac{(n-1)s^2}{n\chi^2_{\alpha/2}{}^{(n-1)}}} \le \frac{\sigma}{\sqrt{n}} \le \sqrt{\frac{(n-1)s^2}{n\chi^2_{(1-\alpha/2)}{}^{(n-1)}}}$

$\sigma/\sqrt{n}$

-směrodatná odchylka odhadu průměru (S.E.)

Interval spolehlivosti pro odhad rozptylu - příklad

Příklad: měření produkce metabolitu (x) u buněk dvou nádorových linií

Linie 1
n = 50
$s^2(x) = 10\ (mg/ml)^2$
s(x) = 3,16 mg/ml
$\bar{x}$ = 2 mg/ml
$\bar{s}_x$ = 0,447 mg/ml

95% IS

$$\frac{49 * 10}{77{,}22} \leq \sigma^2 \leq \frac{49 * 10}{31{,}56}$$

$$6{,}98 \leq \sigma^2 \leq 15{,}53$$

c = 1,58

Linie 1
n = 100
$s^2(x) = 16\ (mg/ml)^2$
s(x) = 4 mg/ml
$\bar{x}$ = 2,8 mg/ml
$\bar{s}_x$ = 0,4 mg/ml

95% IS

$$\frac{99 * 16}{128{,}42} \leq \sigma^2 \leq \frac{99 * 16}{73{,}36}$$

$$12{,}33 \leq \sigma^2 \leq 13{,}49$$

c = 1,43

Výpočet mediánu z frekvenčních dat a jeho odhady

a) Určete medián tohoto souboru dat: 1,3,4,5,7,8 [4,5]

b) Určete medián tohoto souboru dat: 5,1,8,3,4 [4]

c) Tento příklad je ukázkou výpočtu mediánu u velkého souboru dat. V následující tabulce je uveden rozbor rozložení souboru dat od 179 krav, kde sledovanou veličinou byl počet dní od narození telete do znovuobnovení menstruačního cyklu. Uvedená data jsou velmi zjednodušena a jsou zde uvedena pouze pro ilustraci:

Class limits (days)	**0,5-20,5**	**20,5-40,5**	**40,5-60,5**	**60,5-80,5**	**80,5-100,5**	**100,5-120,5**	**120,5-140,5**	**140,5-160,5**	**160,5-180,5**	**180,5-200,5**	**200,5-220,5**
Frequency	8	33	50	32	15	20	11	6	2	1	1
Cumulative frequency	8	41	91	123	138	158	169	175	177	178	179

Frekvence zastoupení dosahuje nejvyšší hodnoty u třídy od 40,5 – 60,5 dnů. Druhý (menší) frekvenční pík lze pozorovat u intervalu od 100,5 do 120,5 dní. Existence dvou maxim (bimodální data) je důkazem nenormality tohoto konkrétního souboru.

Výpočet mediánu z frekvenčních dat a jeho odhady

Jelikož n =179, pak je medián devadesátá hodnota od počátku souboru, a dále je zřejmé, že bude velmi blízko horní hranici třídy 40,5 – 60,5 dní. Za předpokladu, že 50 hodnot této třídy je v ní rovnoměrně rozmístěno lze použít následující vzorec:

$$M = X_L + \frac{gl}{f}, \quad kde$$

X_L = hodnota X (sledované veličiny) na spodní hranici třídy obsahující medián: zde 40,5 dní

g = pořadová hodnota mediánu minus kumulativní frekvence do horní hranice předchozí třídy, tj. 90 - 41= 49

l = třídní interval: 20 dní

f = frekvence ve třídě obsahující medián

- Dosadíme-li do uvedeného vzorce, získáme odhad mediánu jako 60 dní. Průměr tohoto datového souboru je 69,9, což je významně odlišná hodnota, a potvrzuje znovu nenormální charakter dat.
- U velkých vzorků z normálních populací je výběrový odhad mediánu normálně rozložen kolem populační hodnoty se směrodatnou odchylkou $1{,}253\ \sigma / \sqrt{n}$. U normálního rozložení, kde medián i průměr představují odhad stejné hodnoty, je medián méně přesný než průměr. Proto hlavní význam mediánu spočívá u nesymetrických distribucí.
- Existuje velmi jednoduchá metoda pro výpočet intervalu spolehlivosti pro odhad mediánu a jako horní a spodní hranice slouží pořadová čísla vypočítaná podle následujícího vztahu:

$$\frac{(n+1)}{2} \pm \frac{z\sqrt{n}}{2}, \quad kde$$

n představuje velikost datového souboru, **z** je kvantil standardizovaného normálního rozložení pro příslušnou pravděpodobnost. U našeho příkladu je n = 179 a pro 95% interval spolehlivosti je z přibližně rovno 2. Horní a spodní limit pro odhad mediánu tedy je $90 \pm \sqrt{179} = 77$ a 103. 95% interval spolehlivosti je tedy tvořen počty dní, které mají pořadí 77 a 103:

77: Počet dní = 40,5+(36)(20)/50 = 55 dní

103: Počet dní = 60,5+(12)(20)/32 = 68 dní

Medián cílové populace byl tedy odhadnut 95% intervalem spolehlivosti jako hodnota ležící mezi 55 a 68 dny. Interpretujte tento výsledek.