

# VII. Popisná statistika dat



Popisné statistiky dat  
Vizualizace dat

# Anotace



- Popisná analýza dat je po vizualizaci dat dalším krokem v procesu statistického hodnocení. Poskytuje představu o rozsazích hodnocených dat a umožňuje vyhodnotit, srovnání s literárními údaji nebo dosavadní zkušeností, jejich realističnost.
- Již při výběru vhodné popisné statistiky se uplatňuje znalost rozložení dat. Některé popisné statistiky, odvozené od modelových rozložení, je možné využít pouze v případě, že data mají dané modelové rozložení. Typickým příkladem je průměr a směrodatná odchylka, jejichž předpokladem je přítomnost normálního rozložení.

# Typůrých



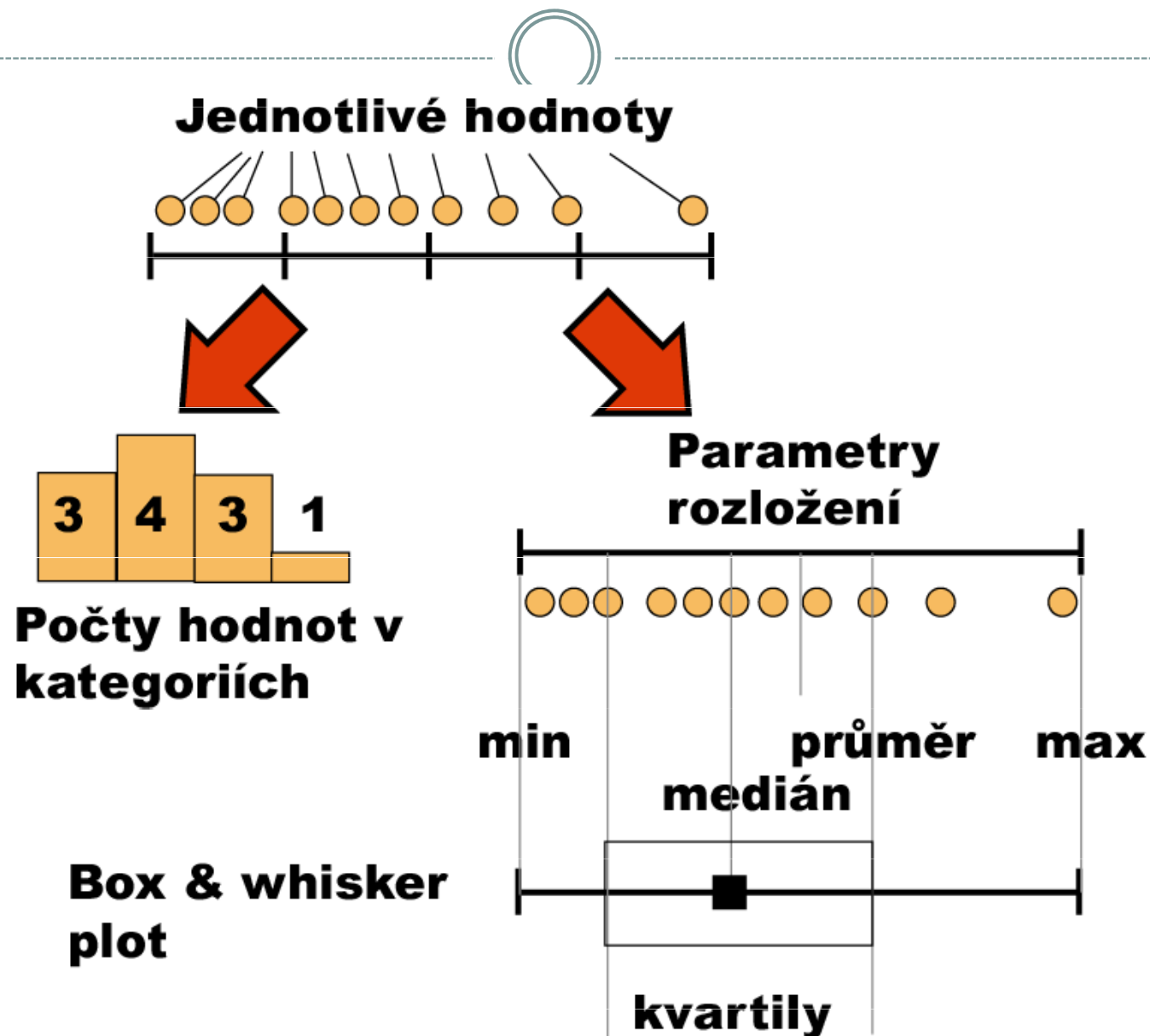
- **Kvalitativní/kategorická**

- binární - ano/ne
- nominální - A,B,C ... několik kategorií
- ordinální-  $1 < 2 < 3$  ...několik kategorií a můžeme se ptát, která je větší

- **Kvantitativní**

- nespojitá – čísla, která však nemohou nabývat všech hodnot (např. počet porodů)
- spojitá – teoreticky jsou možné všechny hodnoty (např. krevní tlak)

# Řada dat a její vlastnosti



# Frekvenční rozložení



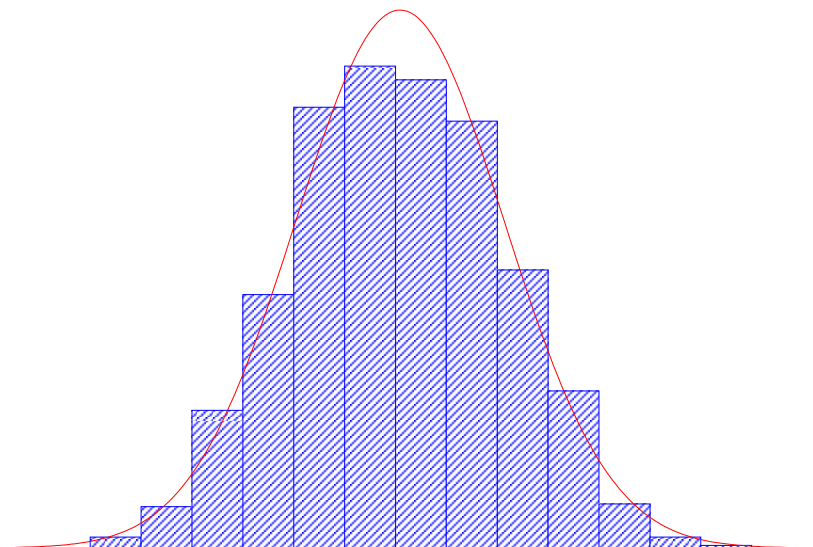
Kategorie	Četnost
B	5
C	8
D	1

## Kvalitativní data

Tabulka s četností jednotlivých kategorií.

## Kvantitativní data

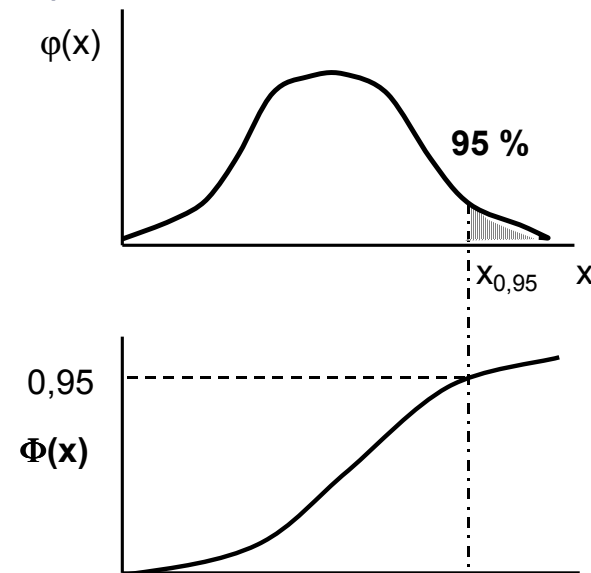
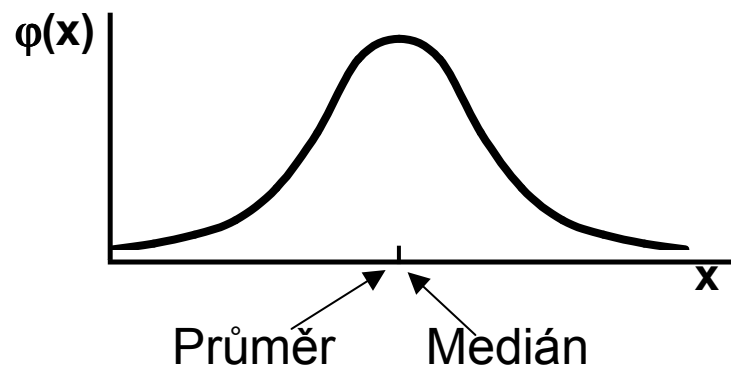
Četnost hodnot rozložení v jednotlivých intervalech.



# Parametrozložení



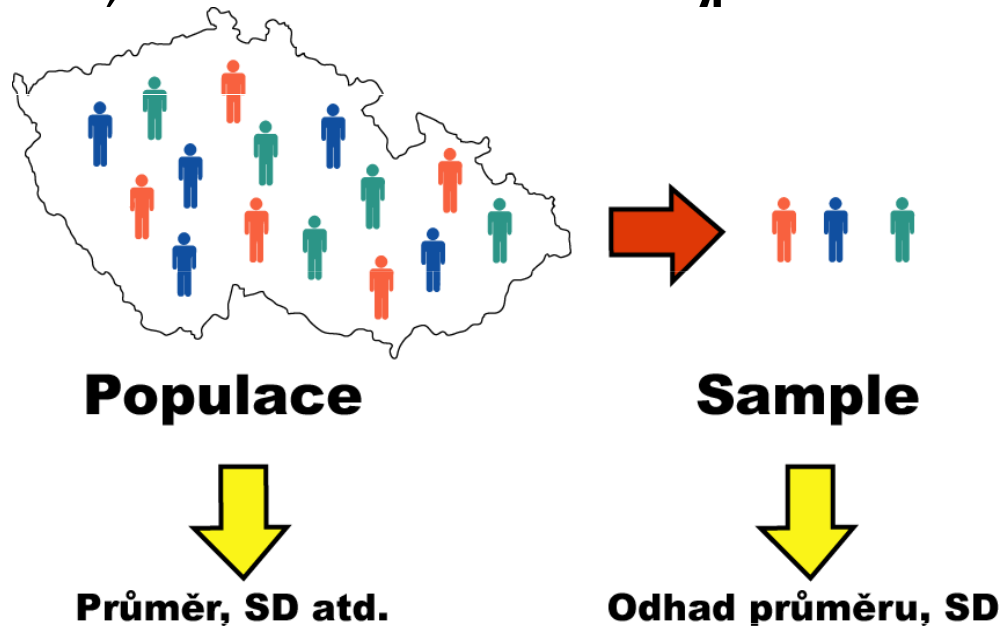
- Soubor dat (řada čísel) můžeme charakterizovat parametry jeho rozložení
- Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
  - Středu (medián, průměr, geometrický průměr)
  - Šířky rozložení (rozsah hodnot, rozptyl, směrodatná odchylka)
  - Tvaru rozložení (skewness, kurtosis)
  - Kvantily rozložení – kolik % řady dat leží nad a pod kvantilem



# Populace a vzorek



- Populace představuje veškeré možné objekty vzorkování, např. veškeré obyvatelstvo ČR při sledování na úrovni ČR, z populace získáme reálné parametry rozložení
- Z populace je prováděno vzorkování za účelem získání reprezentativního vzorku (**sample**) populace, toto vzorkování by mělo být náhodné, důležitá je také velikost vzorku, ze vzorku získáme **odhad parametrů rozložení**



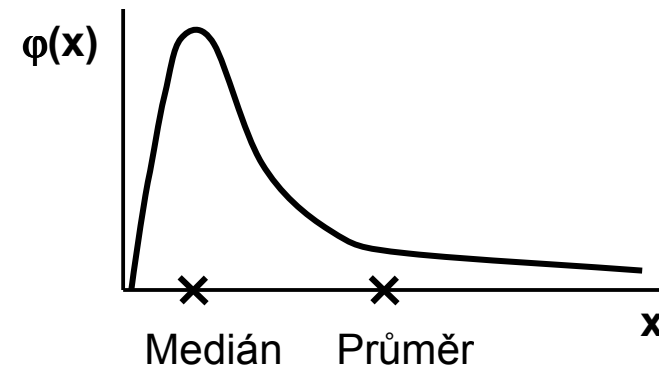
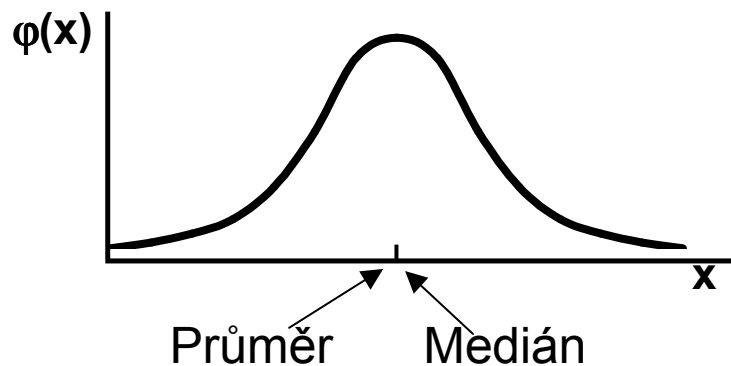
# Ukazatele středu rozložení I



- **Průměr** – vhodný ukazatel středu u normálního/symetrického rozložení, kde  $x_i$  jsou jednotlivé hodnoty a  $n$  jejich počet

$$E(x) = \bar{x} = \sum_{i=1}^n \frac{x_i}{n}$$

- **Medián** – jde vlastně o 50% kvantil, tj. polovina hodnot leží nad a polovina pod mediánem
- V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné

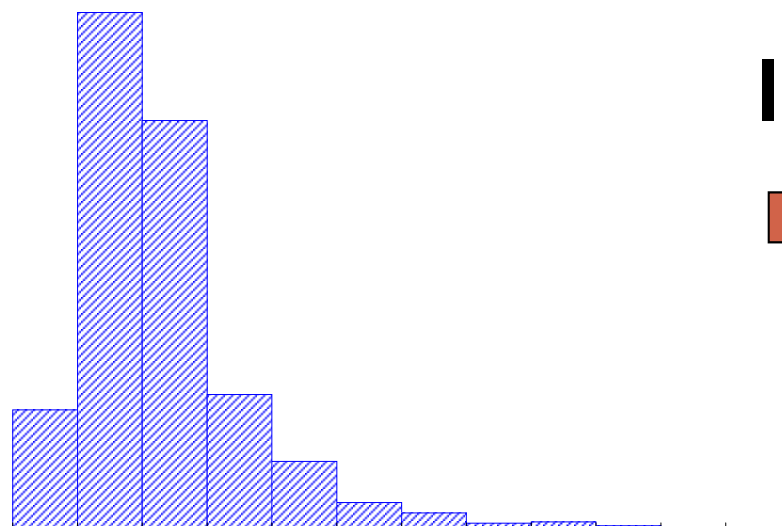




# Ukazatele středu rozložení II.

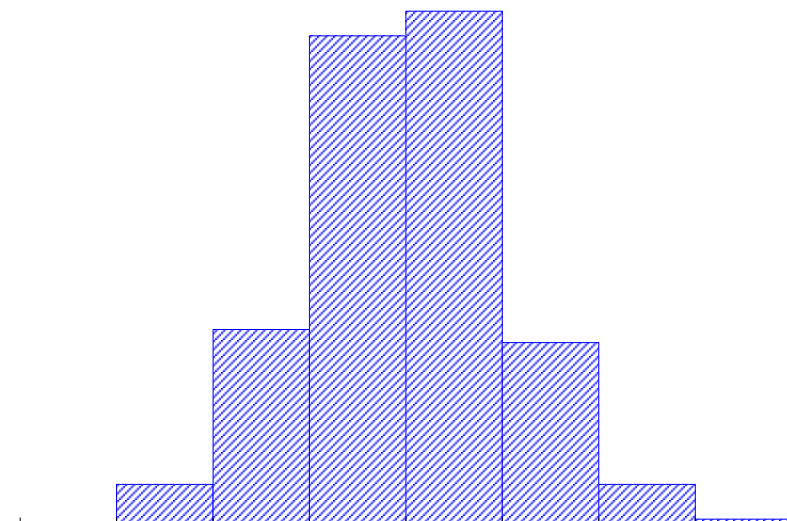
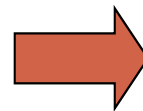


- Geometrický průměr – antilogaritmus průměru logaritmovaných dat, je vhodný pro doleva asymetrická data (lognormální rozložení), která jsou v biologii velmi častá, jeho hodnota v podstatě odpovídá mediánu
- Takto asymetrická data je možné převést logaritmickou transformací na normální rozložení



Průměr  
Medián, geometrický průměr

log



Průměr (logaritmovaných dat)

# Ukazatele šířky rozložení

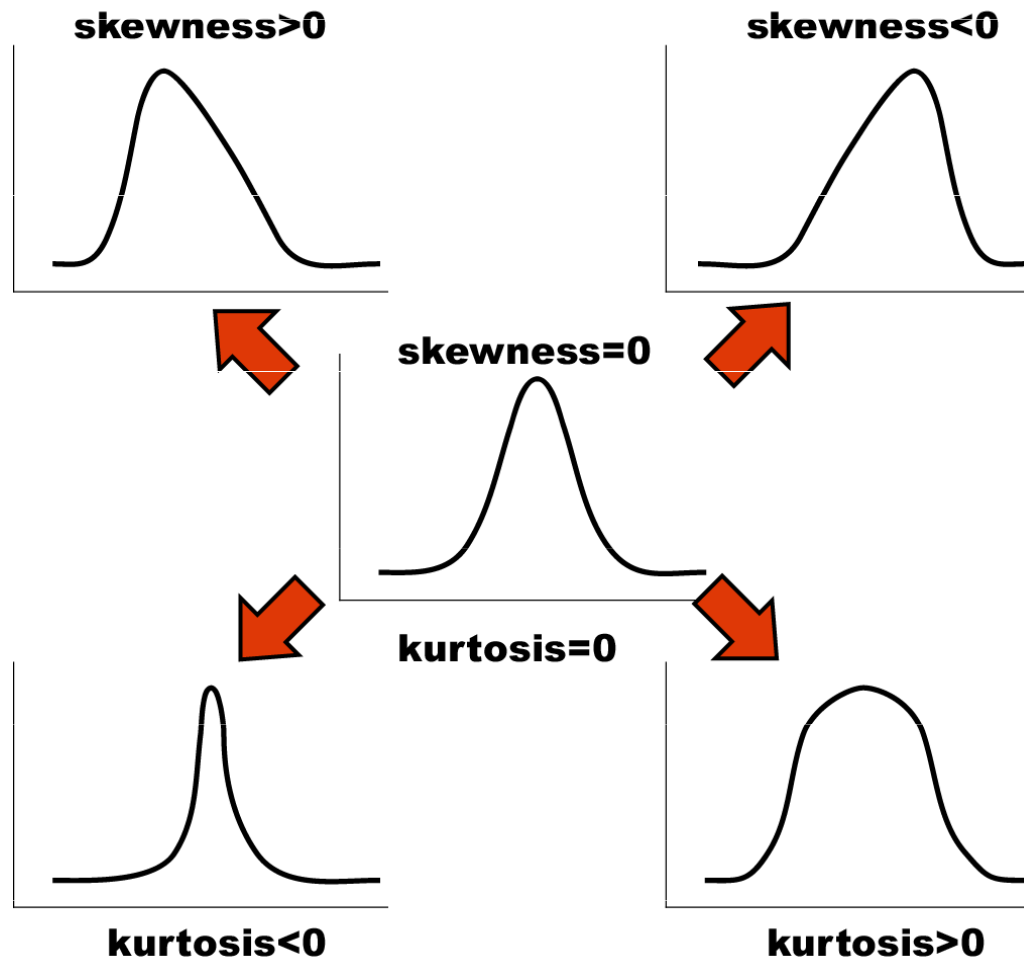


- **Rozptyl** je ukazatelem šířky rozložení získaný na základě odchylky jednotlivých hodnot od průměru. 
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$
- Obdobně jako u průměru je jeho vypovídací schopnost nejvyšší v případě symetrického/normálního rozložení
- **Střední odchylka** je druhá odmocnina z rozptylu
- **Koeficient variance** - podíl SD ku průměru (u normálního rozložení by se 95% hodnot mělo vejít do průměr  $\pm 3$  SD), pokud je SD větší než 1/3 průměru jsou teoreticky pravděpodobné záporné hodnoty v rozložení – ukazatel problémů s normalitou dat

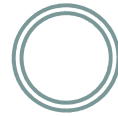
# Ukazatele tvaru rozložení



- **Skewness** – ukazatel „šikmosti“ rozložení, asymetrie rozložení
- **Kurtosis** – ukazatel „špičatosti/plochosti“ rozložení



# Další parametry rozložení



- **Počet hodnot** – důležitý ukazatel, znamená jak moc lze na data spoléhat
- **Střední chyba odhadu průměru** - je založena na směrodatné odchylce rozložení a **počet hodnot**, vlastně jde o směrodatnou odchylku rozložení průměru. Říká jak přesný je náš výpočet průměru. Čím větší počet hodnot rozložení, tím je náš odhad skutečného průměru přesnější.
- **Suma hodnot**
- **Modus** – nejčastější hodnota, vhodný např. při kategoriálních datech
- **Minimum, maximum**
- **Rozsah hodnot**
- **Harmonický průměr** - převrácená hodnota průměru převrácených hodnot (vždy platí harmonický průměr < geometrický průměr < aritmetický průměr)

# VIII. Provádění odhadů



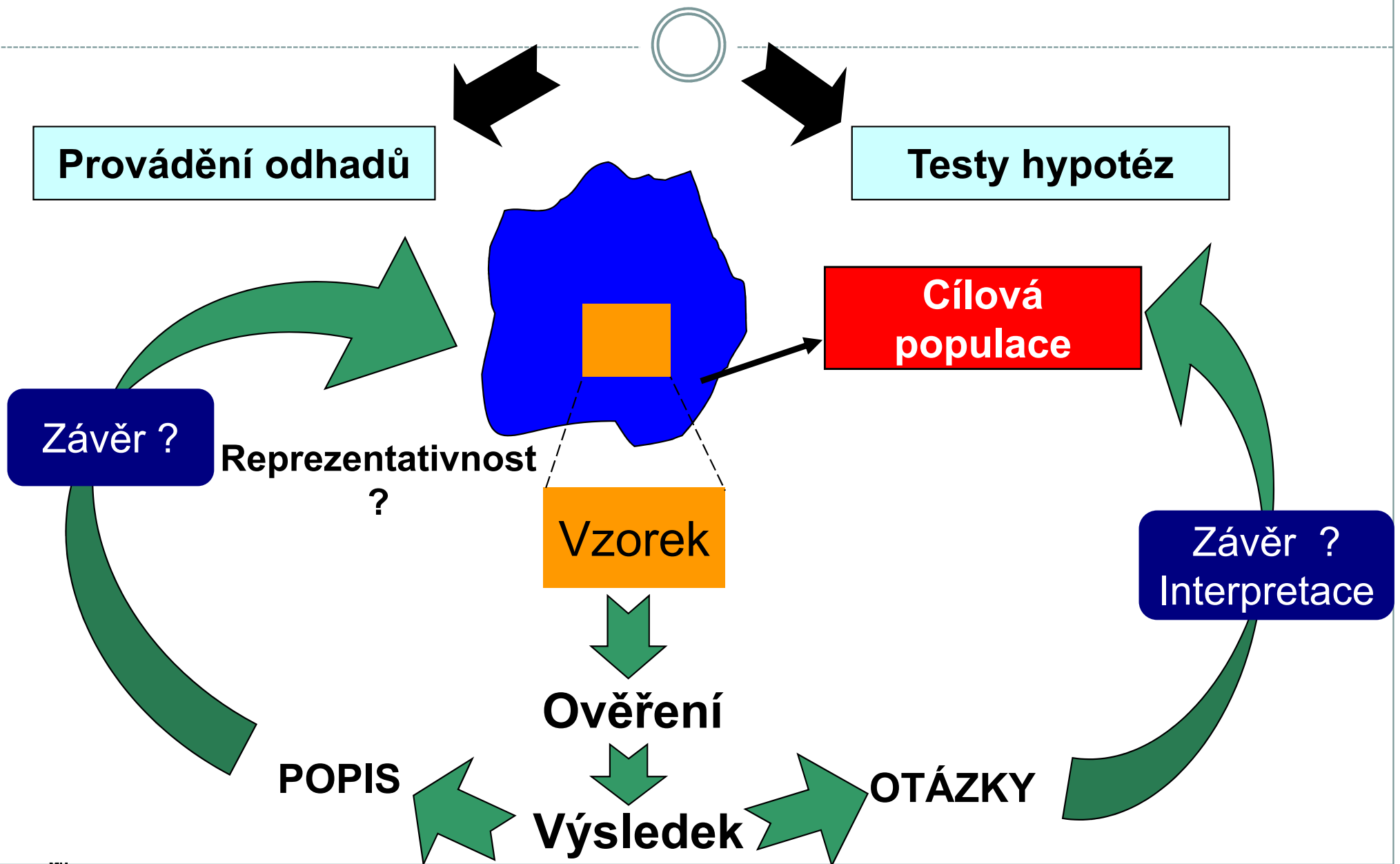
**Bodové a intervalové odhady**  
**Význam intervalu spolehlivosti**

# Anotace



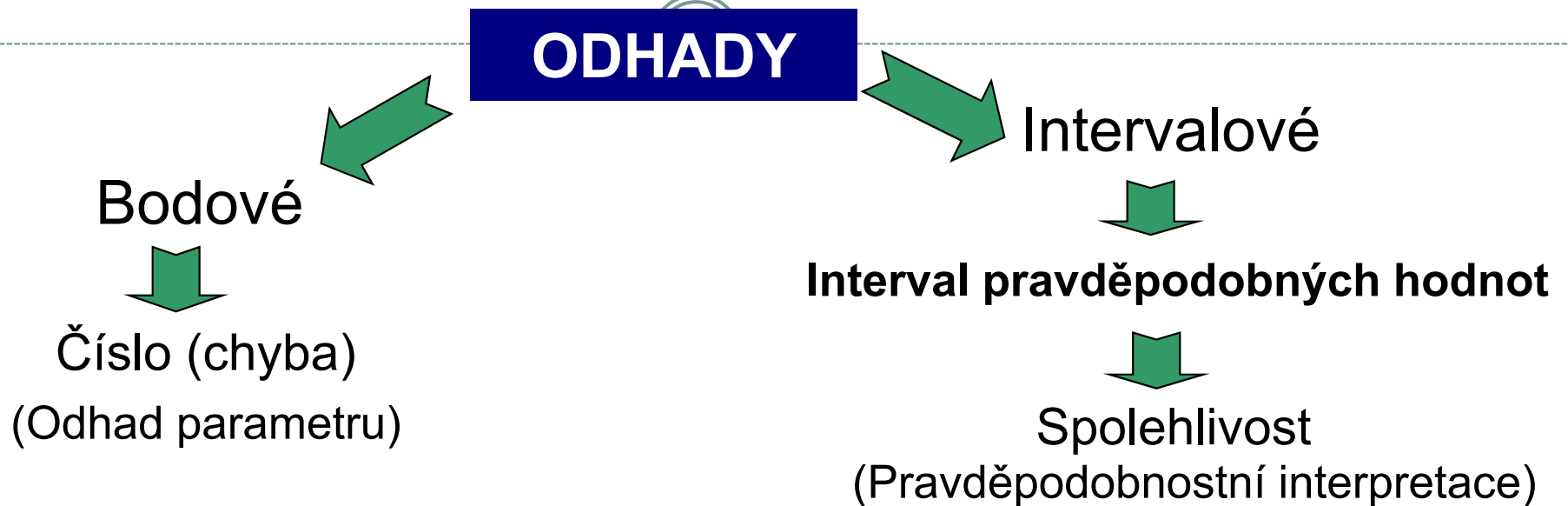
- Dva základní přístupy statistického hodnocení jsou popis dat a testování hypotéz. Při popisu dat je třeba si uvědomit, že popisné statistiky získané ze vzorku nejsou skutečnou hodnotou v cílové populaci, ale pouze jejím odhadem. Přesnost odhadu závisí jednak na variabilitě dat, jednak na velikosti vzorku, při navzorkování celé cílové populace by výsledná popisná statistika již byla přesnou hodnotou, nikoliv odhadem.
- Odhady a s nimi související intervaly spolehlivosti jsou univerzálním statistickým postupem a je možné je dopočítat k libovolné popisné statistice.

# Statistika v průzkumném studiu



# INTERVAL SPOLEHLIVOSTI

## ve hmotnosti užitečného vědomosti odhadů



**Obecný tvar:**

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

**Odhadovaný  
parametr**

$\pm$

Kvantil  
modelového  $\times$  SE (odhadu)  
rozložení

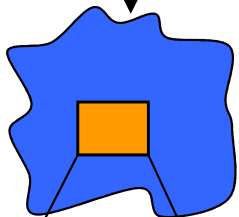
$K_V$  pro  $(1 - \alpha/2)$



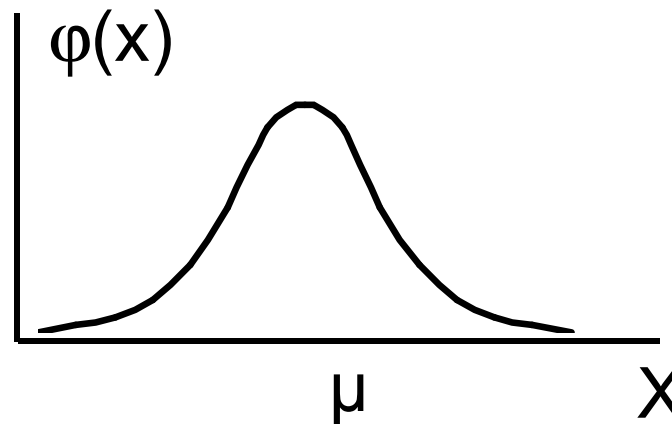
# NORMÁLNÍ ROZLOŽENÍ: rozděl po odhad průměru



Cílová populace



Vzorek:  $n$



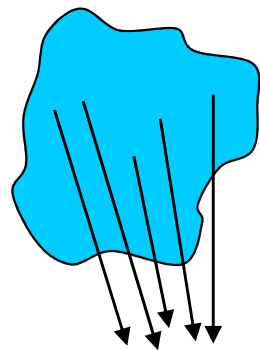
## Prezentace

- $n; \bar{x}; s$
- $n; \bar{x}; \frac{s}{\sqrt{n}}$
- $n; \bar{x}; c$
- $n; \bar{x};$  Interval spolehlivost i pro odhad průměru

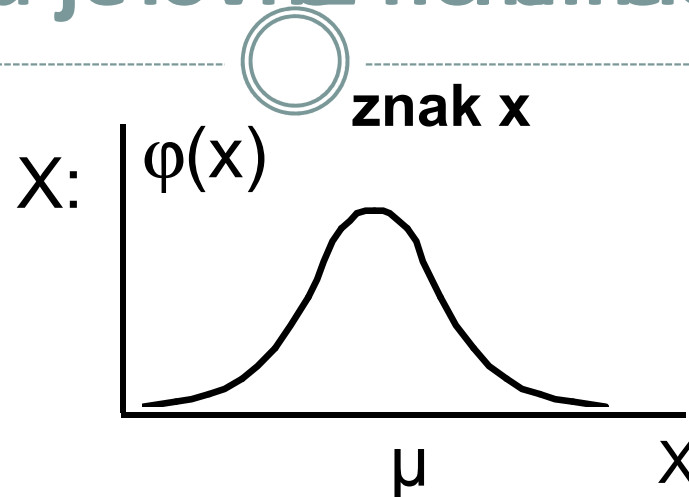


$\bar{X}$  ..... odhad průměru

# NORMÁLNÍ ROZLOŽENÍ: odhad průměru je rovněž normálně rozložen



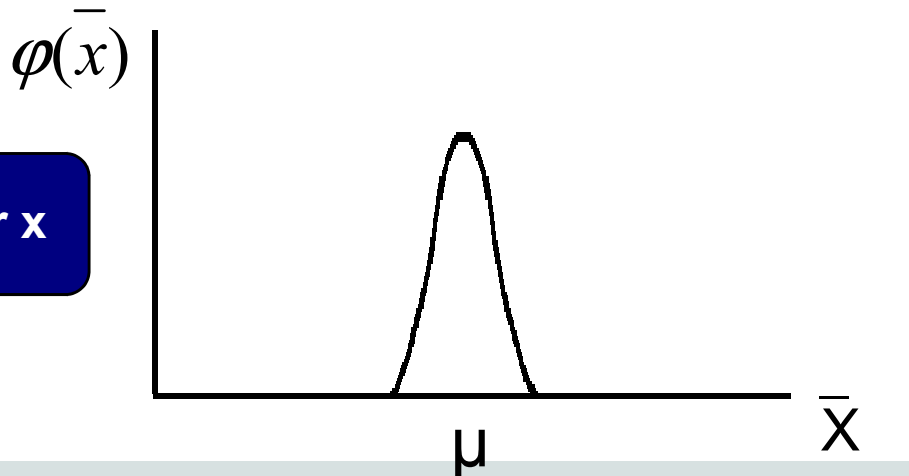
Cílová populace



$$x: \mu \pm 3s$$

Náhodné výběry o  $n = 100$

$\bar{x}_1$   $\bar{x}_2$   $\bar{x}_3$   $\bar{x}_4$  ....  $\bar{x}_i$



průměr  $x$

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

$$\frac{s}{\sqrt{n}} \sim$$

Standardní chyba odhadu průměru

# ODHAD PRŮMĚRU: Vztahy



## Bodový

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



## Intervalový

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

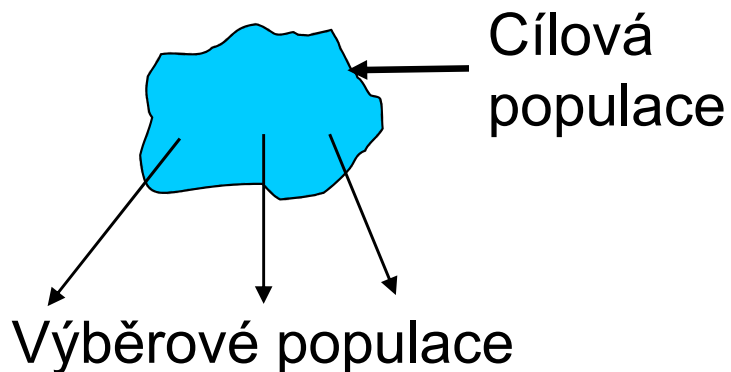
$$\mu : \bar{x} \pm t_{1-\alpha/2}^{(v=n-1)} \cdot \frac{s}{\sqrt{n}}$$

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

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

# Interval spolehlivosti odhadu průměru je pouze informací o přesnosti tohoto odhadu

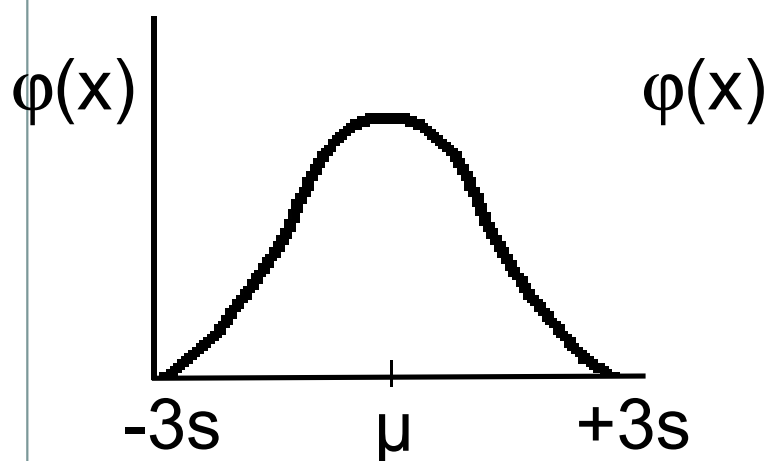
**Interval spolehlivosti je hodnocen pro  $(1 - \alpha)$  procentní spolehlivost**



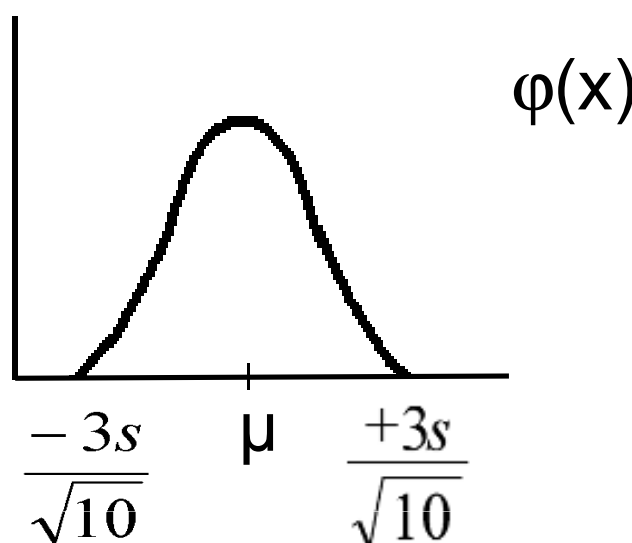
Šířku intervalu určuje:

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

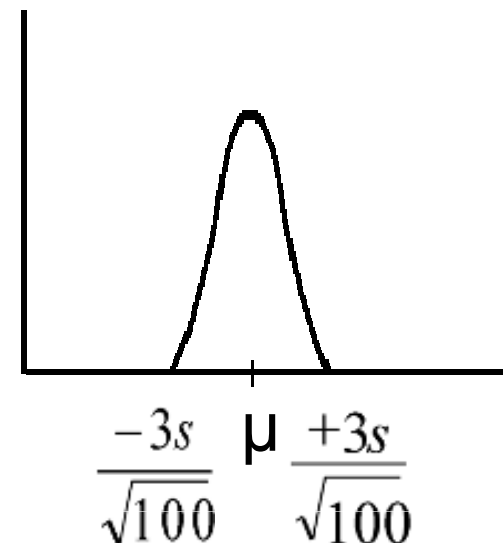
**Původní proměnná  $x$**



**Výběr  $n=10$  pro odhad průměru**



**Výběr  $n=100$  pro odhad průměru**



# ODHAD PRŮMĚRU: 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:

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

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

$$3,58 - 2,086 \cdot 0,075 \leq \mu \leq 3,58 + 2,086 \cdot 0,075$$

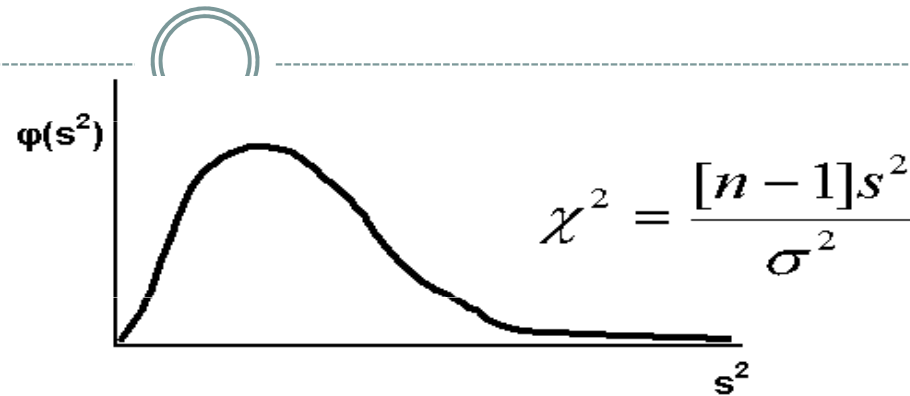
$$3,423 \leq \mu \leq 3,737$$



$$P(3,423 \leq \mu \leq 3,737) \geq 0,95$$

# Interval spolehlivosti odhad rozptylu

$s^2 \sim \sigma^2$  pro velká  $n$



## Interval spolehlivosti

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

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

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

$$\sigma/\sqrt{n}$$

-směrodatná odchylka odhadu průměru (S.E.)

# Interval spolehlivosti pro odchad rozptylu: příklad

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

Linie 1

$n = 50$

$s^2(x) = 10 \text{ (mg/ml)}^2$

$s(x) = 3,16 \text{ mg/ml}$

$\bar{x} = 2 \text{ mg/ml}$

$\bar{s}_x = 0,447 \text{ mg/ml}$

**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

$n = 100$

$s^2(x) = 16 \text{ (mg/ml)}^2$

$s(x) = 4 \text{ mg/ml}$

$\bar{x} = 2,8 \text{ mg/ml}$

$\bar{s}_x = 0,4 \text{ mg/ml}$

**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 zjednodušená 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 odhad



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:

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

$X_L$  = 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:

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

$n$  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.**