

# Bi8600: Vícerozměrné metody

## 1. cvičení



**Vyučující:** Mgr. Lucie Brožová

**Kontakt:** brozova@iba.muni.cz

# Průběh výuky



- Obsahem cvičení je praktická aplikace pokročilých statistických metod
  - Zopakování jednorozměrné analýzy dat
  - Investigativní vícerozměrná analýza dat
  - Diskriminační analýza
- Předpoklady úspěšného ukončení cvičení
  - Účast na cvičení (povolena jedna absence)
- Plán cvičení
  - 9. 10. Opakování jednorozměrné analýzy dat
  - 16. 10. Shluková analýza
  - 13. 11. Metoda hlavních komponent (PCA)
  - 20. 11. Ordinační metody (CA, NMDS) + diskriminační analýza

# Bi8600: Vícerozměrné metody

## 1. cvičení – 1. část



Základní popis a práce s daty v softwaru R

# Motivace



- Současná statistická analýza se neobejde bez zpracování dat pomocí statistických software. Předpokladem úspěchu je správné uložení dat ve formě „databázové“ tabulky umožňující jejich zpracování v libovolné aplikaci.
- Neméně důležité je věnovat pozornost čištění dat předcházející vlastní analýze. Každá chyba, která vznikne nebo není nalezena ve fázi přípravy dat se promítne do všech dalších kroků a může zapříčinit neplatnost výsledků a nutnost opakování analýzy.

# DATA – ukázka uspořádání datového souboru

## Parametry (znaky)

Základní jednotka dat

Pacient	Clovek	aLeu cell.10 <sup>6</sup> /	aTy% %	aSe% %	aNeu% %	aLy% %	aTy cell.10 <sup>6</sup> /	aSe cell.10 <sup>6</sup> /	aNeu cell.10 <sup>6</sup> /	aLy cell.10 <sup>6</sup> /	aHtc %	aCLsk mV.s.10 <sup>3</sup>	aCLNeus mV.s.10 <sup>3</sup>	aCLOZ mV.s.10 <sup>3</sup>	aCLNeuO mV.s.10 <sup>3</sup>
3	1	4									33	72		32	
4	2	7,6	8	58	66	24	0,6	4,4	5,0	1,8	33	95	19	48	10
8	3	4	3	52	55	40	0,1	2,1	2,2	1,6	22	77	35	33	15
11	4	6,1	5	59	64	35	0,3	3,6	3,9	2,1	33	103	26	49	13
12	5	6,9	3	85	88	9	0,2	5,9	6,1	0,6	37	81	13	45	7
14	6	5,9	15	55	70	19	0,9	3,3	4,1	1,1	32	137	33	61	15
16	7	8	18	75	93	7	1,4	6,0	7,4	0,6	34	151	20	59	8
20	8	9,6	3	72	75	23	0,3	6,9	7,2	2,2	40	77	11	38	5
21	9	6	10	67	77	19	0,6	4,0	4,6	1,1	32	120	26	52	11
22	10	3,3	4	55	59	39	0,1	1,8	2,0	1,3	28	81	42	24	12
37	11	3,8	10	60	70	30	0,4	2,3	2,7	1,1	32	111	42	29	11
38	12	6,4	2	76	78	17	0,1	4,9	5,0	1,1	25	366	73	115	23
39	13	6,8	1	57	58	39	0,1	3,9	3,9	2,7	20	234	59	71	18
49	14	8,5	7	67	74	26	0,6	5,7	6,3	2,2	30	156	25	108	17
51	15	9,3	7	57	64	35	0,7	5,3	6,0	3,3	35	129	21	23	4
52	16	2,2	10	56	66	34	0,2	1,2	1,5	0,7	33	46	30	12	8
55	17	9,9	3	78	81	10	0,3	7,7	8,0	0,1	30	189	24	140	18
56	18	5	2	80	82	13	0,1	4,0	4,1	0,7	26	101	25	54	13
6	1	8,8	11	72	83	12	1,0	6,3	7,3	1,1	44	268	36,6	145	19,9
9	2	9,2	2	66	68	28	0,2	6,1	6,3	2,6	42	168	26,9	76	12,2
13	3	10,0	7	83	90	8	0,7	8,3	9,0	0,8	54	181	20,1	81	9
15	4	9,6	1	75	76	23	0,1	7,2	7,3	2,2	45	343	47	124	16,9
17	5	6,0									45	40		21	
19	6	7,2	2	78	80	18	0,1	5,6	5,8	1,3	44	103	17,8	63	10,9
24	7	8,2	1	72	73	25	0,1	5,9	6,0	2,1	41	209	34,9	57	9,6
26	8	10,3	1	85	86	3	0,1	8,8	8,9	0,3	41	364	41,1	112	12,6
29	9	5,0	1	74	75	21	0,1	3,7	3,8	1,1	39	83	22,1	32	8,5
30	10	11,9	1	51	52	47	0,1	6,1	6,2	5,6	33	83	13,4	52	8,4
31	11	7,2	3	53	56	29	0,2	3,8	4,0	2,1	28	109	27,1	63	15,5
32	12	10,8	36	50	76	8	3,9	5,4	9,3	0,9	27	146	15,7	106	11,4
33	13	11,8	22	54	76	16	2,6	6,4	9,0	1,9	45	246	27,4	63	7
34	14	17,0	1	82	83	16	0,2	13,9	14,1	2,7	34	440	31,2	119	8,4
40	15	10,0	8	72	80	4	0,8	7,2	8,0	0,4	37	176	22,0	52	6,5

# Typy proměnných



## Kvalitativní (kategoriální) proměnná

- lze ji řadit do kategorií, ale nelze ji kvantifikovat

*Příklad: ??*

## Kvantitativní (numerická) proměnná

- můžeme ji přiřadit číselnou hodnotu

*Příklad: ??*

# Typy proměnných



## Kvalitativní (kategoriální) proměnná

- lze ji řadit do kategorií, ale nelze ji kvantifikovat

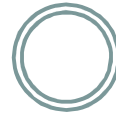
*Příklad: pohlaví, HIV status, barva vlasů ...*

## Kvantitativní (numerická) proměnná

- můžeme ji přiřadit číselnou hodnotu

*Příklad: výška, váha, teplota, počet hospitalizací ...*

# Kvalitativní znaky



- **Binární znaky**: dvě kategorie, obvykle se kódují pomocí číslic 1 (přítomnost sledovaného znaku) a 0 (nepřítomnost sledovaného znaku).  
*Příklad: ??*
- **Nominální znaky**: několik kategorií (A, B, C), které nelze uspořádat.  
*Příklad: ??*
- **Ordinální znaky**: několik kategorií, které lze vzájemně seřadit, tedy můžeme se ptát, která je větší/menší ( $1 < 2 < 3$ ).  
*Příklad: ??*



# Kvalitativní znaky



- **Binární znaky**: dvě kategorie, obvykle se kódují pomocí číslic 1 (přítomnost sledovaného znaku) a 0 (nepřítomnost sledovaného znaku).  
*Příklady: Diabetes (1-ano, 0-ne), Pohlaví (1-muž, 0-žena).*
- **Nominální znaky**: několik kategorií (A, B, C), které nelze uspořádat.  
*Příklad: krevní skupiny (A/B/AB/O).*
- **Ordinální znaky**: několik kategorií, které lze vzájemně seřadit, tedy můžeme se ptát, která je větší/menší ( $1 < 2 < 3$ ).  
*Příklady: stupeň bolesti (mírná/střední/velká), stadium maligního onemocnění (I/II/III/IV).*

# Kvantitativní znaky

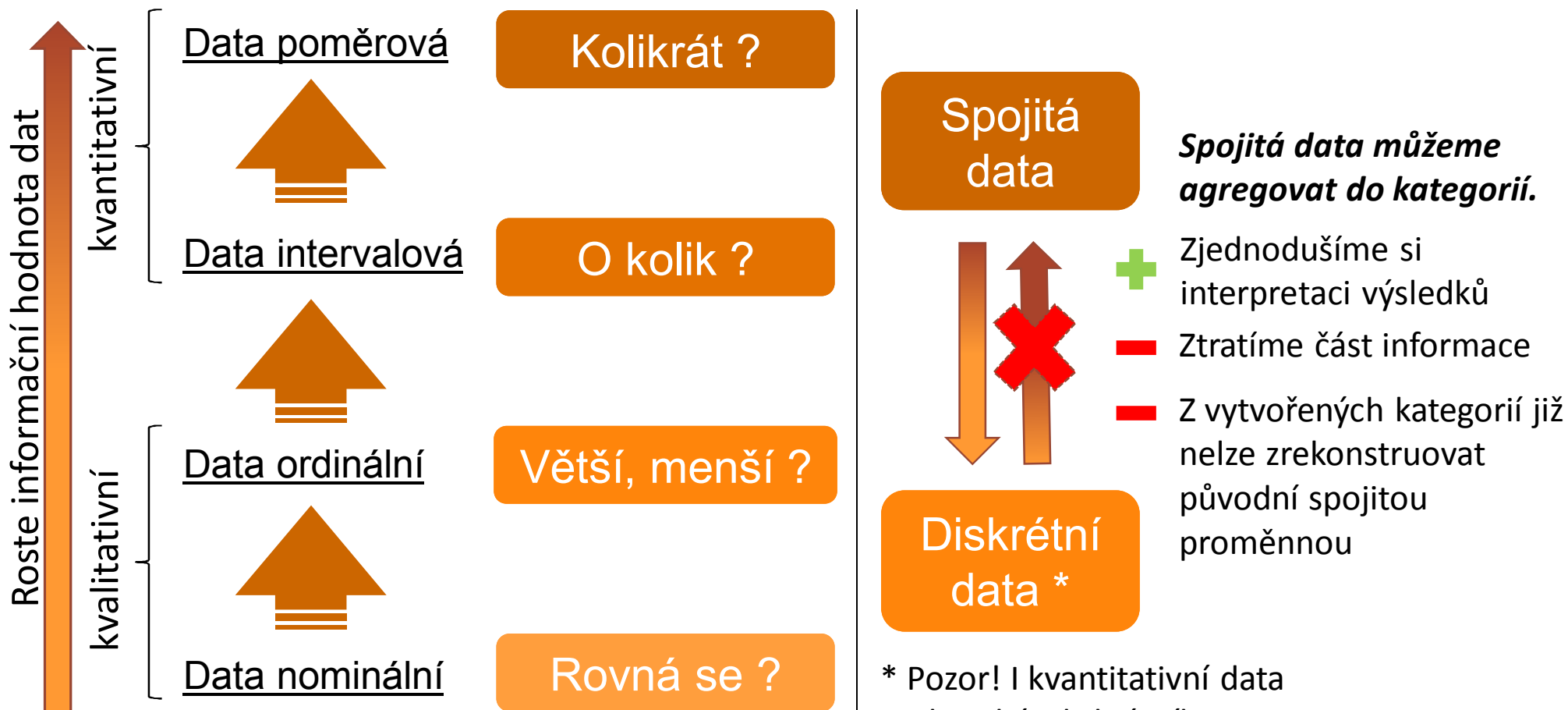


- **Intervalové znaky**: interpretace rozdílu dvou hodnot (stejný interval mezi jednou a druhou dvojicí hodnot vyjadřuje i stejný rozdíl v intenzitě zkoumané vlastnosti). Společný znak intervalových znaků: nula byla stanovena uměle, tedy pouhou konvencí. *Příklad: teplota měřená ve stupních Celsia, letopočet.*

Den	Teplota	Rozdíl	Podíl
1.	- 2 °C	-	-
3.	6 °C	+8	-3x ?

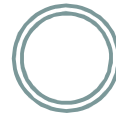
- **Poměrové znaky**: kromě rozdílu interpretujeme i podíl dvou hodnot.  
*Příklady: výška v cm, váha v kg, ...*

# Různé typy dat znamenají různou informaci



\* Pozor! I kvantitativní data mohou být diskrétního typu.  
Např.: počet dětí v rodině.

# Popisné statistiky



## Charakteristiky polohy (míry střední hodnoty, míry centrální tendence)

- Udávají, kolem jaké hodnoty se data centrují, resp. které hodnoty jsou nejčastější, popis „těžiště“ – míry polohy
- **Aritmetický průměr, medián, modus, geometrický průměr**

## Charakteristiky variability (proměnlivosti)

- Zachycují rozptýlení hodnot v souboru (proměnlivost dat)
- **Variační rozpětí, rozptyl, směrodatná odchylka, variační koeficient, střední chyba průměru**

# Popis kvalitativních dat



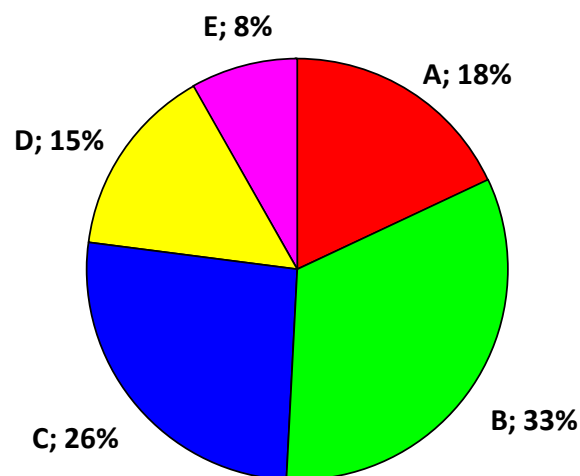
- **Popis kvalitativních dat:**
  - procentuální zastoupení jednotlivých kategorií
  - U ordinálních znaků lze využít  $\alpha$ -kvantil.
- **Vizualizace kvalitativních dat:** nejčastěji koláčový nebo sloupcový graf.

## Frekvenční tabulka

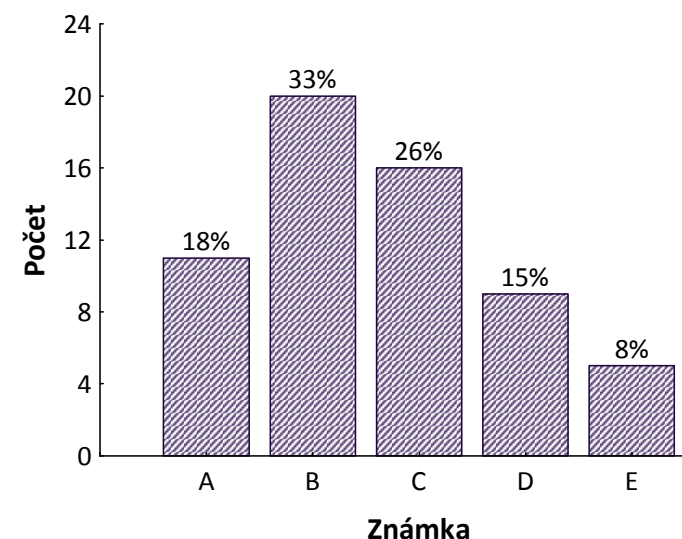
Známka	n	%
A	11	18,0
B	20	32,8
C	16	26,2
D	9	14,8
E	5	8,2
F	0	0,0
Celkem	61	100,0

modus

## Koláčový graf



## Sloupcový graf



# Popis kvantitativních dat

## – charakteristiky středu



- **Aritmetický průměr**: je definován jako součet všech naměřených údajů ( $x_i$ ) vydělený jejich počtem ( $n$ ):

$$\bar{x} = \sum_{i=1}^n x_i / n .$$

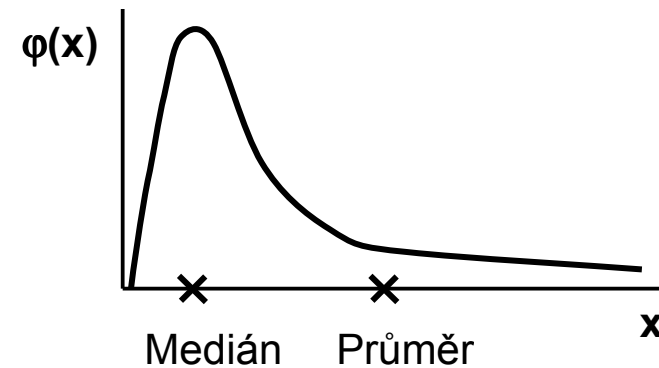
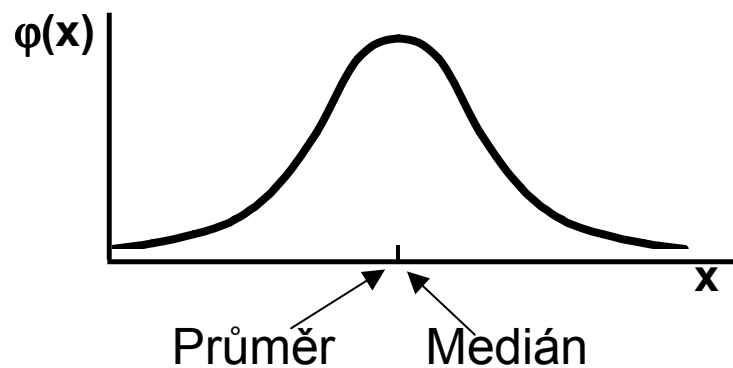
- **Geometrický průměr**: logaritmus geometrického průměru je roven aritmetickému průměru logaritmovaných hodnot souboru.
- **Medián**: znamená hodnotu, jež dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny. Jestliže  $n$  je sudé číslo, pak  $\bar{x} = (x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) / 2$ ,  
jestliže  $n$  je liché číslo, pak  $\bar{x} = x_{(n+1)/2}$ .

# Průměr vs. medián



## PAMATUJ:

- Průměr je silně ovlivněn extrémními hodnotami (tzv. odlehlá pozorování), medián není ovlivněn vybočujícími pozorováními.
- Průměr je vhodný ukazatel středu u normálního/symetrického rozložení, medián je vhodnou charakteristikou středu souboru i v případě veličin s neznámým rozdělením.
- V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné, v případě asymetrického rozložení však nikoliv!



# Popis kvantitativních dat – charakteristiky variability



- **Rozptyl (variance)** 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.

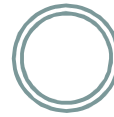
- **Směrodatná odchylka (SD – standard deviation)** je druhá odmocnina z rozptylu.
- **Koeficient variance** = podíl SD ku průměru, umožňuje porovnat variabilitu několika znaků (často se vyjadřuje v procentech – potom udává, z kolika procent se podílí směrodatná odchylka na aritmetickém průměru).
- **Kvartilové rozpětí (odchylka):**

$q = x_{0,75} - x_{0,25}$ , kde  $x_{0,25}$  = dolní kvartil,  $x_{0,75}$  = horní kvartil.

( $x_\alpha$  je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl  $\alpha$  všech dat a na horní úsek obsahující aspoň podíl  $1-\alpha$  všech dat.)



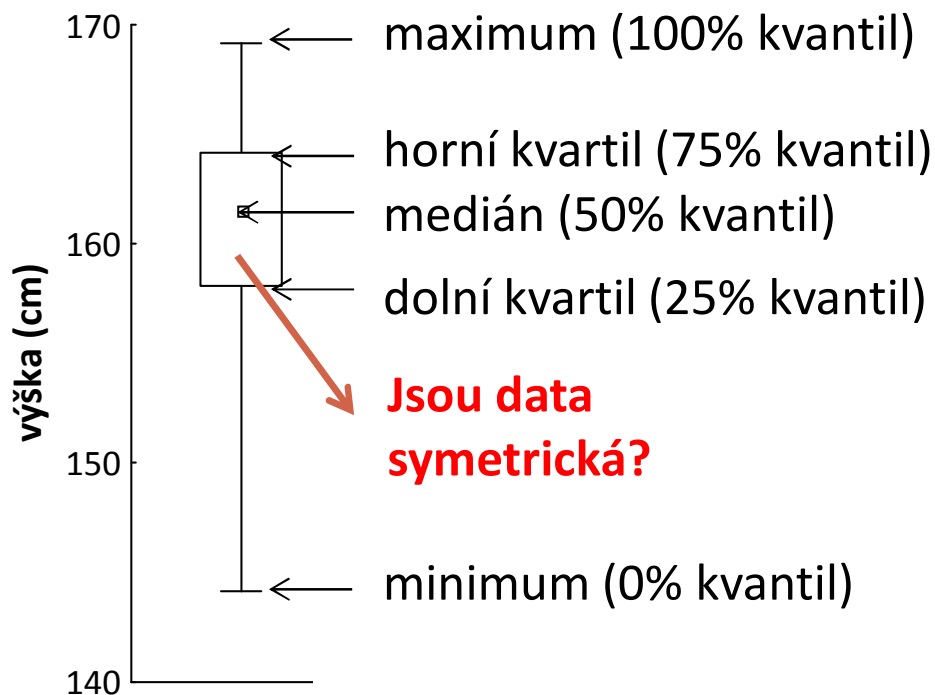
# Ukázka vizualizace kvantitativních dat



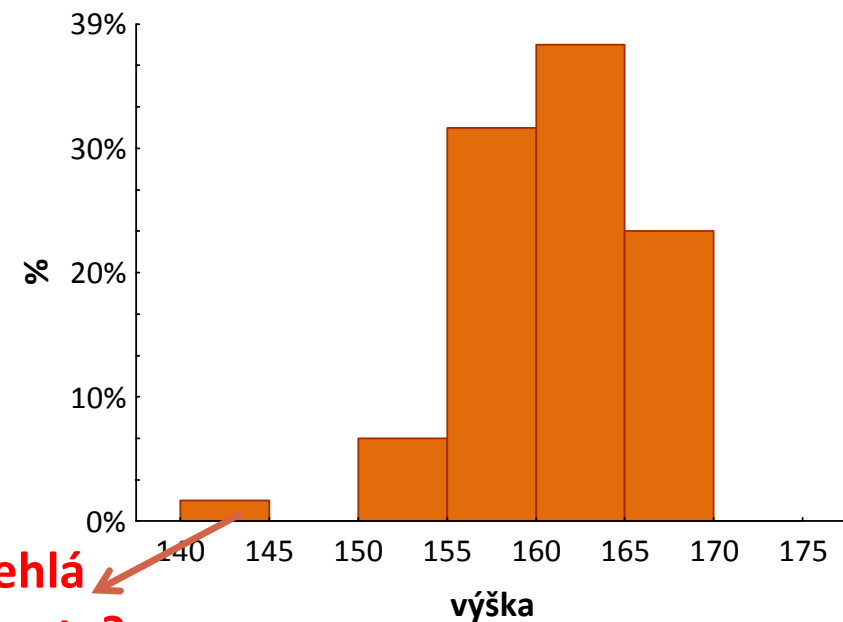
- **Vizualizace kvantitativních dat:** nejčastěji pomocí krabicového grafu nebo histogramu.

## Příklad: Popis výšky (cm)

### Krabicový graf



### Histogram



**Odlehlá  
hodnota?**

# Ukázka popisu kvantitativních dat



- **Popis kvantitativních dat:** charakteristika středu (průměr, medián aj.), charakteristika variability (rozptyl, rozsah hodnot, kvartilové rozpětí aj.).

## Příklad: Popis výšky (cm) pacientů

### Popisné statistiky

Charakteristika	
N	61
Průměr (cm)	161,0
Medián (cm)	161,5
sm. odchylka (cm)	4,7
Rozptyl (cm <sup>2</sup> )	22,2
min-max (cm)	144,1-169,2
dolní-horní kvartil (cm)	158,1-164,2

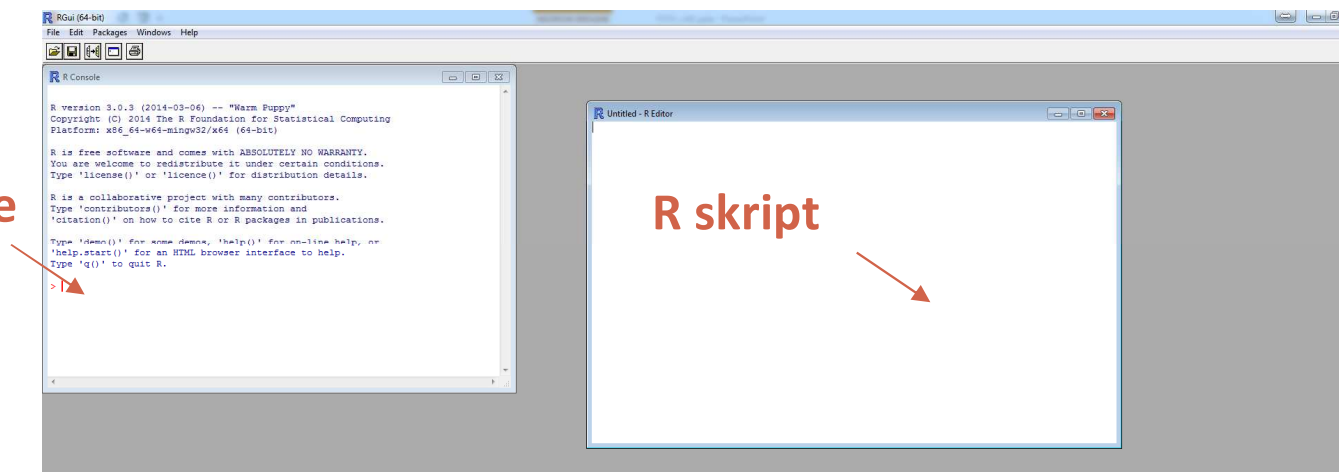
**Průměr a medián se téměř shodují. Co nám to říká?**

# Software R / RStudio



- Volně dostupný software (<https://www.r-project.org/>).
- Pro pokročilé analýzy je nutné načíst balíček, kde jsou naprogramovány funkce.
- Každý má možnost implementovat svůj balíček – R nezaručuje správnost kódu.
- Nevidíme datovou tabulku – nutné kontrolovat provedení výpočtu.
- R console – zápis skriptu + enter spustí skript (alternativou je vytvořit si R script, který umožní kompletní uchování syntaxu, který je spouštěn pomocí Ctrl+R).

R console



R skript

- Nápověda: `help(funkce)`, `?funkce`, <http://rseek.org/>, [www.google.cz](http://www.google.cz).

# Bi8600: Vícerozměrné metody

## 1. cvičení – 2. část

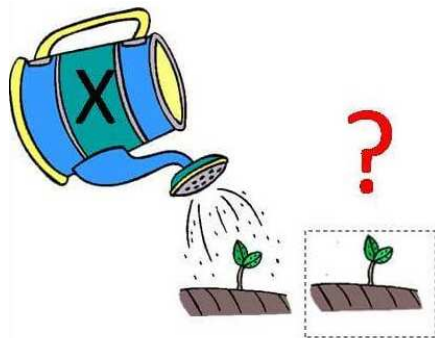


**Základy testování hypotéz**  
**Přehled a aplikace statistických testů**

# Hypotéza

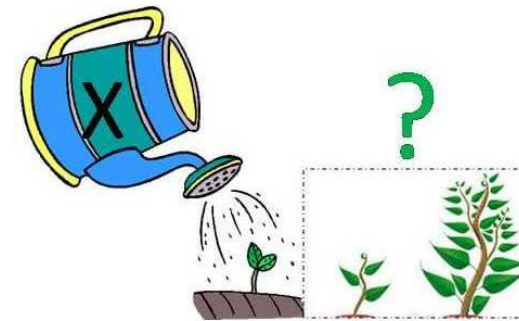
## Nulová hypotéza $H_0$

- Pozorovaný efekt je nulový.
- Cílem je zamítnout platnost  $H_0$  a přijmout platnost  $H_A$ .



## Alternativní hypotéza $H_A$

- Pozorovaný efekt není nulový (existuje rozdíl mezi skupinami).
- Vyvrací platnost  $H_0$ .
- Tvrzení, které chceme dokázat.

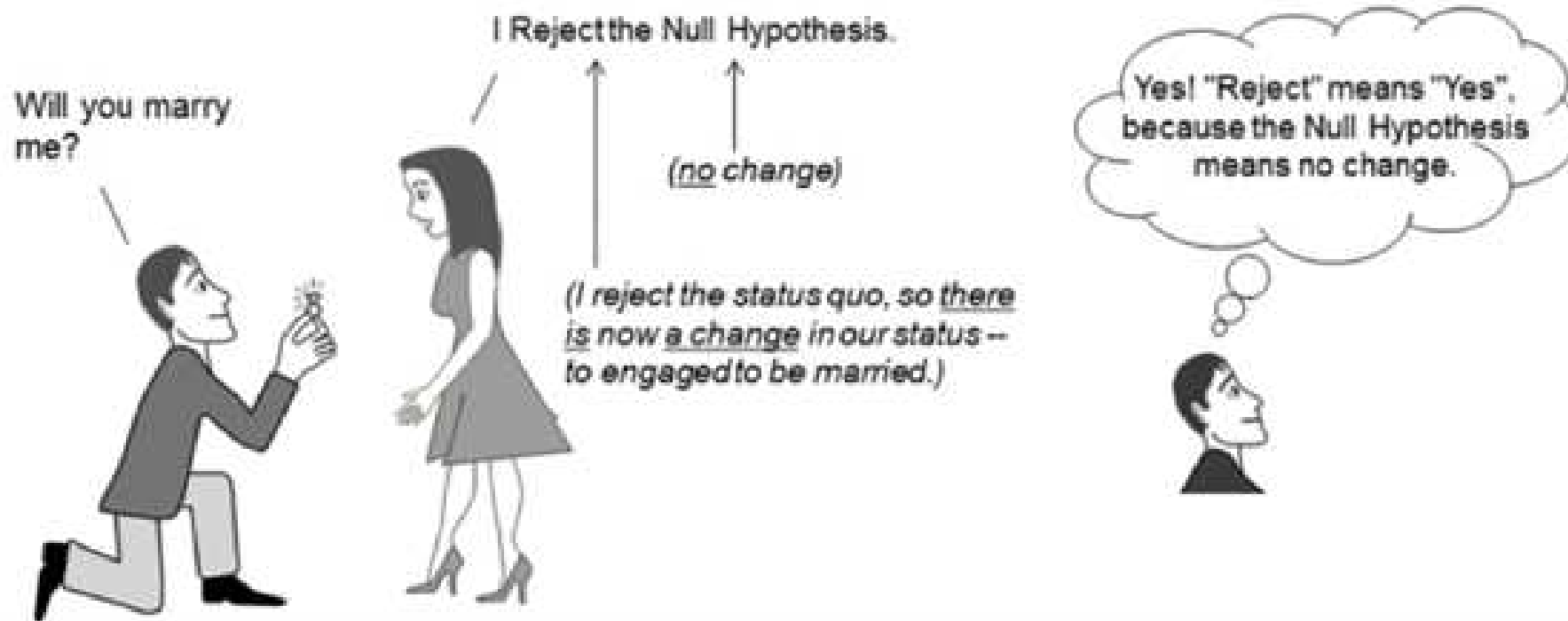


Příklad	$H_0$	$H_A$
Změní aplikace hnojiva proces růstu rostlin?	nezmění	změní
Je průměrný plat populace 20 000?	$\mu = 20\,000,-$	$\mu \neq 20\,000,-$
Liší se úspěšnost u zkoušky mezi muži a ženami?	úspěšnost je stejná	úspěšnost se liší
Došlo po absolvování diety ke změně tělesné váhy?	váha se po dietě nezměnila	váha se po dietě změnila

# Hypotéza



# Hypotéza



Reproduced by permission of John Wiley and Sons, Inc  
from the book *Statistics from A to Z — Confusing Concepts Clarified*.

# Statistická významnost



- Informace z dat je kvantifikována pomocí **testové statistiky** (variabilita dat, efekt = rozdíl mezi skupinami, velikost souboru).
- **Otázka:** Je hodnota testové statistiky dostatečně extrémní, abychom mohli usoudit, že pozorovaný rozdíl není pouze důsledkem náhody?
- **P-hodnota** vyjadřuje pravděpodobnost, že testová statistika nabyde stejné nebo extrémnější hodnoty za **předpokladu, že nulová hypotéza platí** = cílem je dosáhnout co nejnižší p-hodnoty, jelikož tak minimalizujeme pravděpodobnost, že jsme chybně zamítli  $H_0$ , která ve skutečnosti platí.
- P-hodnotu porovnáme s  **$\alpha$  (hladina významnosti)**. Nejčastěji volíme  $\alpha = 0,05$ , tzn., že připouštíme 5% chybu testu, tedy, že zamítneme  $H_0$ , i když ve skutečnosti platí).

- Je-li p-hodnota  $\leq \alpha$ , zamítáme  $H_0$  na hladině významnosti  $\alpha$  a přijímáme platnost  $H_A$ .
- Je-li p-hodnota  $> \alpha$ , pak  $H_0$  nezamítáme na hladině významnosti  $\alpha$  (nepřipouštíme platnost  $H_0$ ).

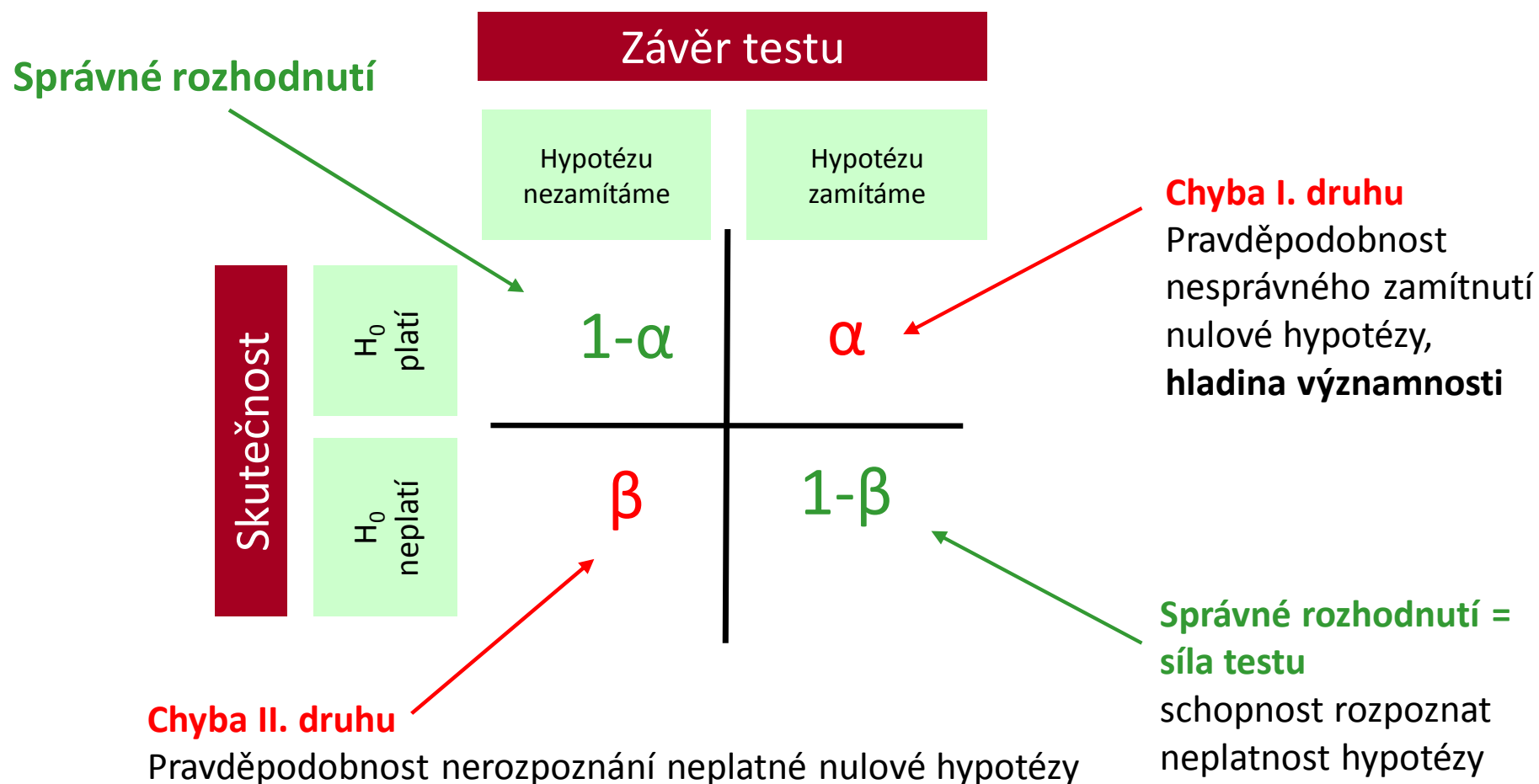
- **Malá p-hodnota nemusí znamenat velký efekt.** Hodnota testové statistiky a p-hodnota mohou být ovlivněny velkou velikostí vzorku a malou variabilitou pozorovaných dat.
- **Statistická významnost** indikuje, že pozorovaný rozdíl není náhodný, ale nemusí znamenat, že je významný i ve skutečnosti. Důležitá je i **praktická (klinická) významnost**.



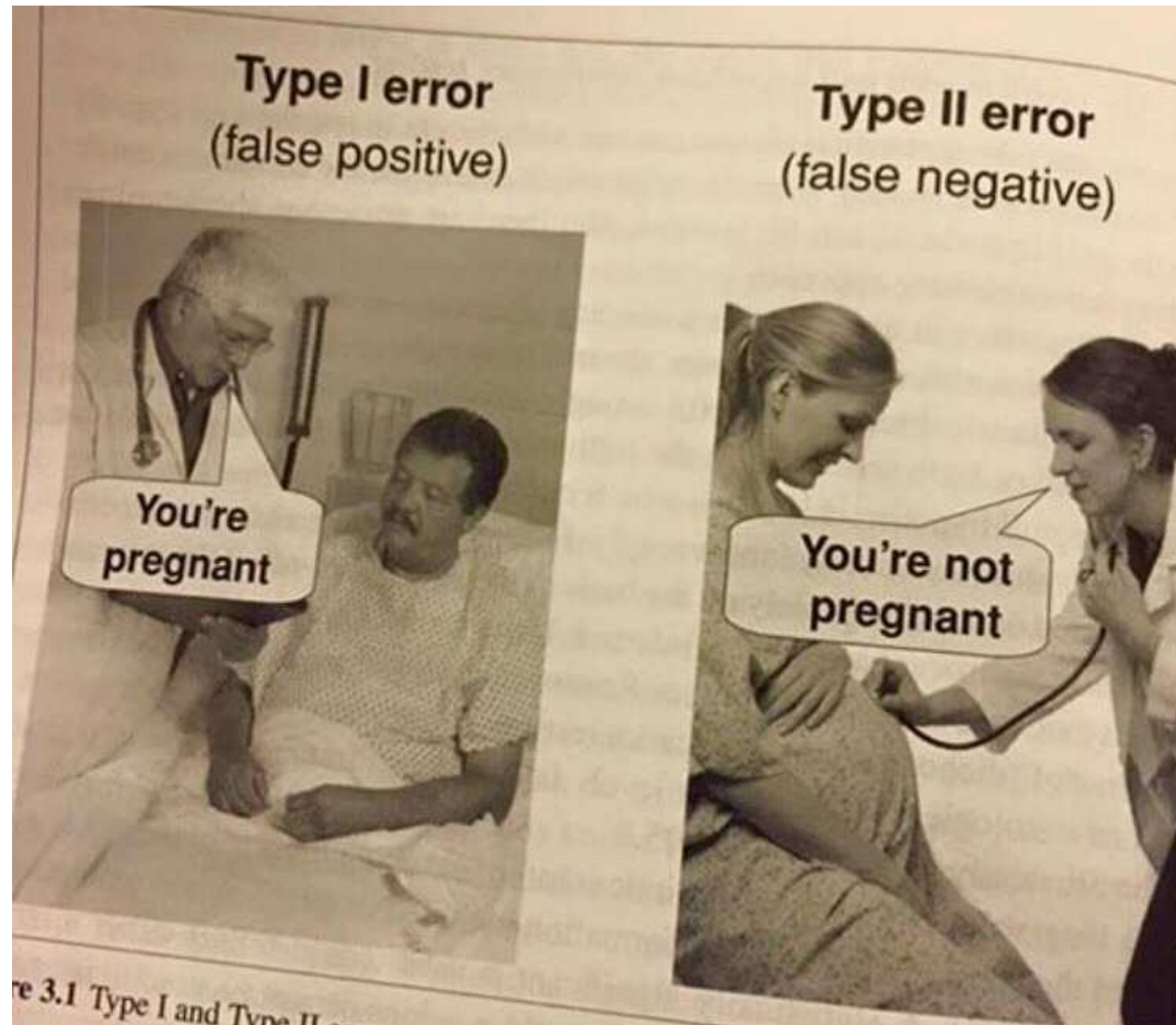
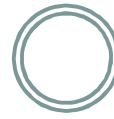
# Možné chyby při testování hypotéz



- I přes dostatečnou velikost vzorku a kvalitní design experimentu se můžeme při rozhodnutí o zamítnutí/nezamítnutí nulové hypotézy dopustit chyby.



# Možné chyby při testování hypotéz



re 3.1 Type I and Type II

# One-tailed vs. two-tailed tests



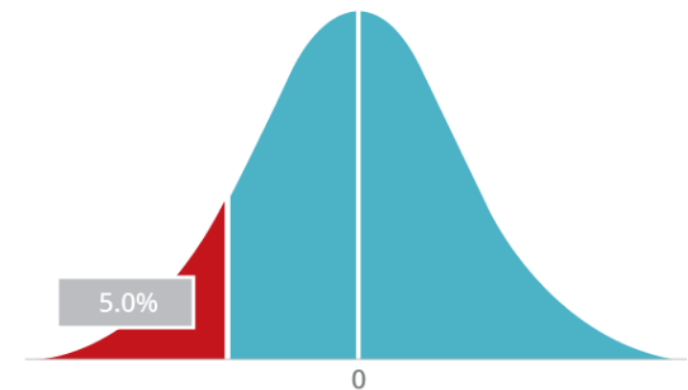
## Jednostranné testy (one-tailed)

- Hypotéza testu je postavena asymetricky, tedy ptáme se na větší než / menší.

$$H_0: \tilde{x} \geq c \quad H_A: \tilde{x} < c$$

$$H_0: \tilde{x} \leq c \quad H_A: \tilde{x} > c$$

- $H_A$  testuje pouze jeden směr asociace.
- Jednodušší zamítnout  $H_0$ .
- Např.:  $H_0$ : váha se po dietě nezměnila/vzrostla  
 $H_1$ : váha se po dietě snížila

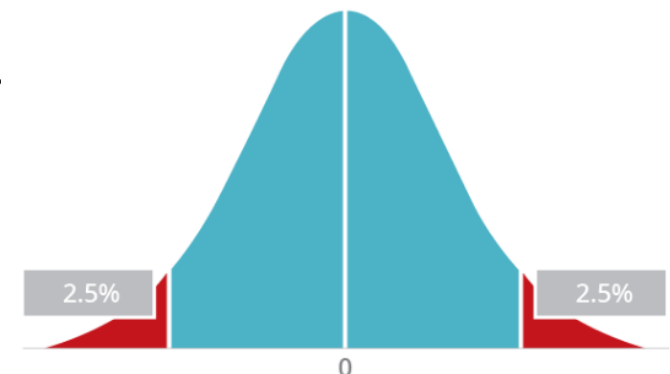


## Oboustranné testy (two-tailed)

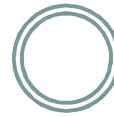
- Hypotéza testu se ptá na otázku rovná se / nerovná se.

$$H_0: \tilde{x} = c \quad H_A: \tilde{x} \neq c$$

- $H_A$  testuje oba směry asociace.
- Obtížněji dosáhneme statistické významnosti.



# Parametrické vs. neparametrické testy



## Parametrické testy

- Mají předpoklady o rozložení vstupujících dat (např. normální rozložení).
- Při stejném  $N$  a dodržení předpokladů mají vyšší sílu testu než testy neparametrické.
- Pokud nejsou dodrženy předpoklady parametrických testů, potom jejich síla testu prudce klesá a výsledek testu může být zcela chybný a nesmyslný.

## Neparametrické testy

- Vyžadují méně předpokladů o rozložení vstupujících dat, lze je tedy použít i při asymetrickém rozložení, odlehlých hodnotách.
- Snížená síla těchto testů je způsobena redukcí informační hodnoty původních dat, kdy neparametrické testy nevyužívají původní hodnoty, ale nejčastěji pouze jejich pořadí.
- Souvisí s malou velikostí souboru (nejsme schopni normalitu dat ověřit).

# Základní rozhodování o výběru statistických testů

