

# Přednáška XI.

## Asociace ve čtyřpolní tabulce a základy korelační analýzy

- Relativní riziko a poměr šancí
- Princip korelace dvou náhodných veličin
- Korelační koeficienty – Pearsonův a Spearmanův
- Korelace a kauzalita



evropský  
sociální  
fond v ČR



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání  
pro konkurenceschopnost



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

# Opakování – Testování hypotéz o podílech

- V čem se liší konstrukce intervalů spolehlivosti a testování hypotéz při rozhodování o podílech (zastoupení „úspěchů“ v náhodném výběru)?

# Opakování – Fisherův exaktní test

→ Jak funguje Fisherův exaktní test?

Veličina $X$	Veličina $Y$		Celkem
	$Y = 1$	$Y = 2$	
$X = 1$	$a$	$b$	$a + b$
$X = 2$	$c$	$d$	$c + d$
Celkem	$a + c$	$b + d$	$n$

# Opakování – Chí-kvadrát test dobré shody

- Lze použít chí-kvadrát test dobré shody na testování normality dat?
- Pokud ano, jak?

# 1. Vyjádření rizik ve čtyřpolní tabulce

# Motivace

- Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS).

Výsledky dány v tabulce:

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

- Pomocí Pearsonova chí-kvadrát nebo Fisherova exaktního testu můžeme rozhodovat o závislosti/nezávislosti dvou sledovaných veličin. Testy ale neumožňují tento vztah kvantifikovat.
- **Má-li to smysl a chceme-li kvantifikovat (rozhodovat o těsnosti této závislosti) můžeme použít tzv. **relativní riziko (RR)** a **poměr šancí (OR)**.**

# Srovnávané skupiny

→ Pomocí *RR* i *OR* můžeme srovnat pravděpodobnosti výskytu sledovaného jevu ve dvou různých skupinách:

→ **1. skupina s pravděpodobností výskytu události  $P_1$ :**

- experimentální skupina – např. léčená novou léčbou
- riziková skupina – např. hypertonici
- skupina s expozicí určitému faktoru – např. horníci

→ **2. skupina s pravděpodobností výskytu události  $P_0$ :**

- kontrolní skupina
- skupina bez expozice

# Relativní riziko = Relative risk

- Výpočet relativního rizika (RR) umožňuje srovnat pravděpodobnosti výskytu sledovaného jevu ve dvou různých skupinách.
- 1. skupina – **experimentální nebo skupina s expozicí určitému faktoru**
- 2. skupina – **kontrolní nebo skupina bez expozice**

$$RR = \frac{\text{Pravděpodobnost výskytu jevu v 1. skupině (experimentální)}}{\text{Pravděpodobnost výskytu jevu ve 2. skupině (kontrolní)}} = \frac{P_1}{P_0}$$

Sledovaný jev	Skupina		Celkem
	Experimentální	Kontrolní	
Ano	$a$	$b$	$a + b$
Ne	$c$	$d$	$c + d$
Celkem	$a + c$	$b + d$	$n$



$$RR = \frac{P_1}{P_0} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$$



# Příklad – relativní riziko

→ Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS).

Výsledky dány v tabulce:

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

$$RR = \frac{P_1}{P_0} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{\frac{29}{29+7301}}{\frac{15}{15+11241}} = 2,97$$



**Riziko výskytu SIDS u dětí matek ve věku do 25 je téměř třikrát vyšší než u dětí matek rodičích ve vyšším věku.**

# Riziko vs. „šance“ (odds)

- **Riziko a pravděpodobnost** – odhad pravděpodobnosti vzniku onemocnění
- **Relativní riziko** – poměr dvou pravděpodobností
- **Šance** – poměr pravděpodobnosti výskytu jevu a výskytu opačného jevu

$$odds = \frac{P_1}{1 - P_1}$$

- nabývá hodnot mezi 0 a nekonečnem
- pokud kůň vyhraje s pravděpodobností 10%, jaká je jeho **šance** na výhru?

# Poměr šancí = Odds ratio

- Poměr šancí (OR) je další charakteristikou, která umožňuje srovnat výskyt sledovaného jevu ve dvou různých skupinách.
- 1. skupina – **experimentální nebo skupina s expozicí určitému faktoru**
- 2. skupina – **kontrolní nebo skupina bez expozice**

$$OR = \frac{\frac{\text{Pravděpodobnost výskytu jevu v 1. skupině (experimentální)}}{1 - \text{Pravděpodobnost výskytu jevu v 1. skupině (experimentální)}}}{\frac{\text{Pravděpodobnost výskytu jevu ve 2. skupině (kontrolní)}}{1 - \text{Pravděpodobnost výskytu jevu ve 2. skupině (kontrolní)}}} = \frac{O_1}{O_0} = \frac{\frac{P_1}{1 - P_1}}{\frac{P_0}{1 - P_0}}$$

Sledovaný jev	Skupina		Celkem
	Experimentální	Kontrolní	
Ano	$a$	$b$	$a + b$
Ne	$c$	$d$	$c + d$
Celkem	$a + c$	$b + d$	$n$




$$OR = \frac{\frac{P_1}{1 - P_1}}{\frac{P_0}{1 - P_0}} = \frac{\frac{a}{c}}{\frac{b}{d}}$$

# Příklad – odds ratio

→ Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS).

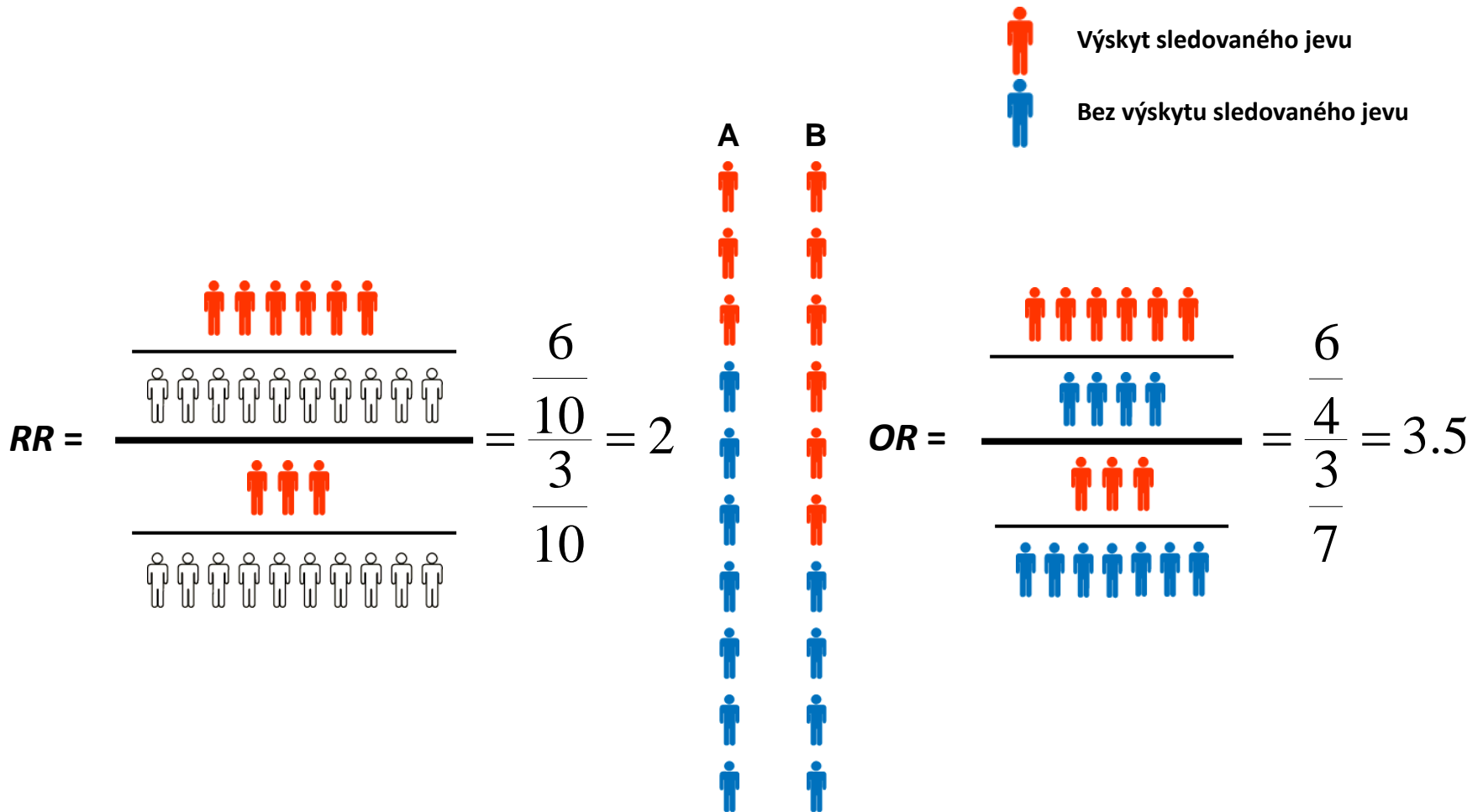
Výsledky dány v tabulce:

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

$$OR = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{\frac{29}{7301}}{\frac{15}{11241}} = 2,98$$


**„Šance“ na výskyt SIDS u dětí matek ve věku do 25 je téměř třikrát vyšší než u dětí matek rodičích ve vyšším věku.**

# Grafické srovnání *RR* a *OR*



# Umělý příklad – pití slazených nápojů

→ Sledujeme vliv pití slazených nápojů na výskyt zubního kazu. Výsledky dány v tabulce:

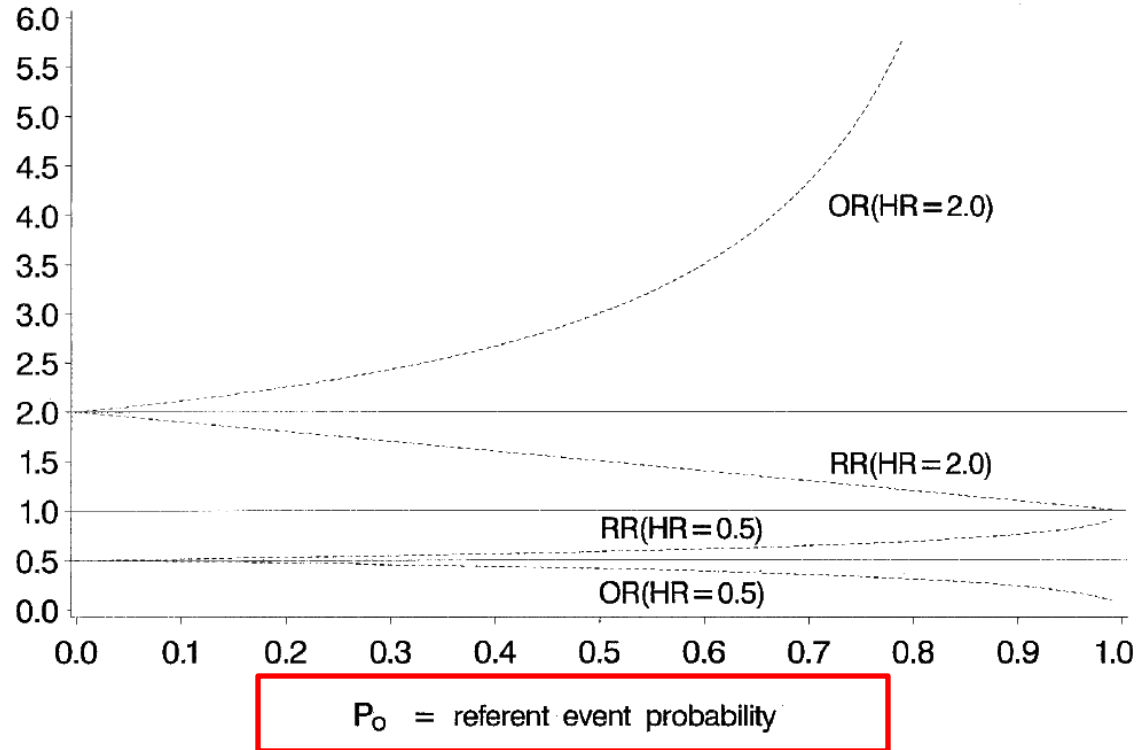
Zubní kaz	Pití slazených nápojů		Celkem
	Ano	Ne	
Ano	34	19	53
Ne	16	31	47
Celkem	50	50	100

$$RR = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{\frac{34}{34+16}}{\frac{19}{19+31}} = 1,79$$

$$OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{\frac{34}{16}}{\frac{19}{31}} = 3,47$$

# Srovnání $RR$ a $OR$

- Hodnoty, jakých může nabývat  $RR$  i  $OR$ , souvisí s četností výskytu sledované události v kontrolní (referenční) skupině.



# Komentáře k RR, OR

- hodnota relativního rizika leží mezi 0 a  $1/P_0$
- pro běžné jevy nelze pozorovat vysoké hodnoty relativního rizika  
pokud je riziko v kontrolní skupině 66%, maximální RR je 1,5
- OR je obtížnější interpretovat
- může být vhodné konvertovat na RR, musíme ale znát riziko v kontrolní skupině

$$RR = \frac{OR}{1 - P_0(1 - OR)} \quad OR = \frac{RR(1 - P_0)}{1 - P_0RR}$$

- nevychází stejně, ale oba jsou validní ukazatele účinku
- **ALE POKUD SE NEJEDNÁ O VZÁCNÝ JEV, OR NELZE INTERPRETOVAT JAKO RR!!!**



# Výhody a nevýhody *RR* a *OR*

→ Nevýhoda *OR*:

→ obtížná interpretace.

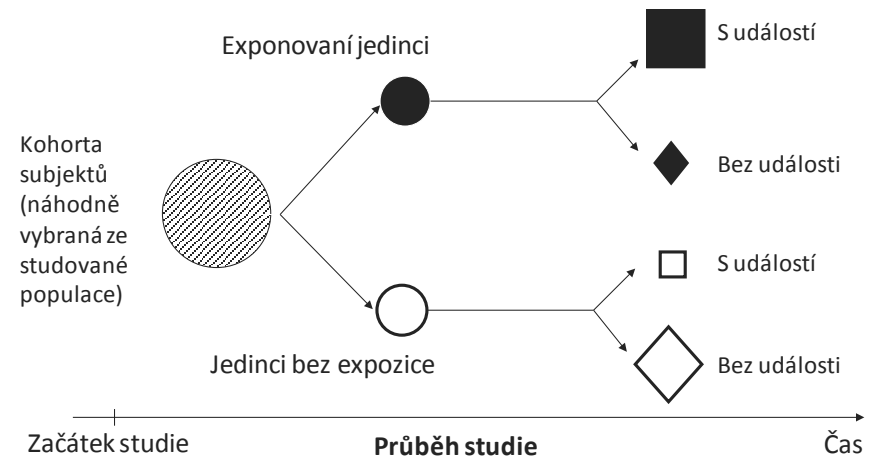
→ Výhoda i nevýhoda *RR*:

→ nezajímá ho samotná pravděpodobnost výskytu jevu, ale pouze jejich podíl → korektní použití *RR* je však pouze v případě, že pravděpodobnost výskytu jevu v kontrolní skupině je reprezentativní (není ovlivněna výběrem sledovaných subjektů).

# Prospektivní a retrospektivní studie

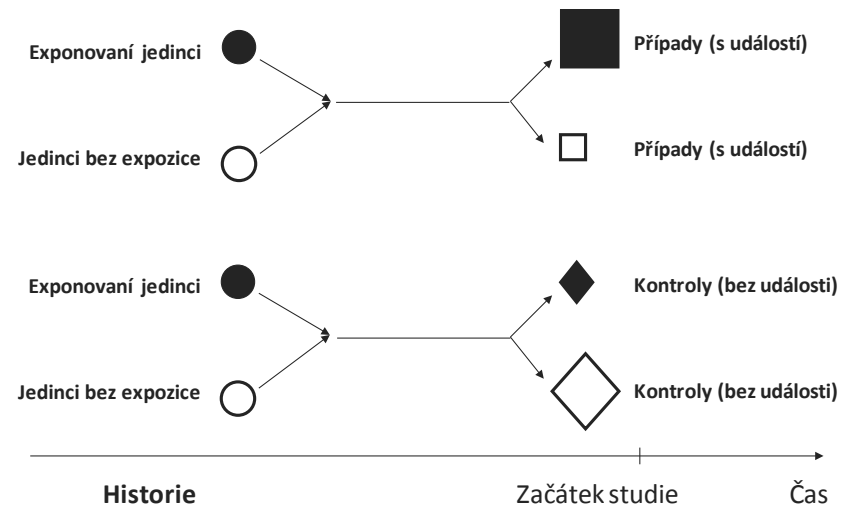
## ➔ **Prospektivní studie**

➔ U některých subjektů je rizikový faktor přítomen a u jiných ne → sledujeme v čase, zda se vyskytne událost.



## ➔ **Retrospektivní studie**

➔ U některých subjektů se událost vyskytla a u jiných ne → zpětně hodnotíme, zda se liší s ohledem na nějaký rizikový faktor.



# Použití *RR* a *OR*

- **Prospektivní studie** – u některých subjektů je rizikový faktor přítomen a u jiných ne → sledujeme, zda se vyskytne událost.
- Zjištěná pravděpodobnost výskytu události v kontrolní skupině je reprezentativní, neboť prospektivně zařazujeme všechny pacienty  
→ **korektní použití *RR*.**
  
- **Retrospektivní studie** – u některých subjektů se událost vyskytla a u jiných ne → zpětně hodnotíme, zda se liší s ohledem na nějaký rizikový faktor.
- Zjištěná pravděpodobnost výskytu události v kontrolní skupině není reprezentativní, neboť ji ovlivňujeme zpětným výběrem skupin subjektů.  
→ **nekorektní použití *RR*.**  
→ **korektní použití *OR*.**

# Intervalové odhady

- $RR$  i  $OR$  jsou variabilní stejně jako četnosti v kontingenční tabulce – bodový odhad je tak vhodné doplnit  $100(1-\alpha)\%$  intervalem spolehlivosti.
- Lze ukázat, že pro nepříliš malé hodnoty  $a, b, c, d$  má přirozený logaritmus  $RR$  ( $\ln RR$ ) i přirozený logaritmus  $OR$  ( $\ln OR$ ) normální rozdělení.

→ Pak platí:

$$SE(\ln RR) = \sqrt{\frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d}}$$

$$SE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

→  $100(1-\alpha)\%$  IS pro přirozené logaritmy:

$$(d^*, h^*) = \ln RR \pm z_{1-\alpha/2} SE(\ln RR)$$

$$(d^*, h^*) = \ln OR \pm z_{1-\alpha/2} SE(\ln OR)$$

→  $100(1-\alpha)\%$  IS pro  $RR$  a  $OR$ :

$$(d^{RR}, h^{RR}) = (\exp(d^*), \exp(h^*))$$

$$(d^{OR}, h^{OR}) = (\exp(d^*), \exp(h^*))$$

# Příklad – intervalové odhady

→ Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS):

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

$$RR = \frac{29/(29+7301)}{15/(15+11241)} = 2,97$$

$$OR = \frac{29/7301}{15/11241} = 2,98$$

→ Logaritmická transformace:

$$SE(\ln RR) = \sqrt{\frac{1}{29} - \frac{1}{29+7301} + \frac{1}{15} - \frac{1}{15+11241}} = 0,317$$

$$SE(\ln OR) = \sqrt{\frac{1}{29} + \frac{1}{15} + \frac{1}{7301} + \frac{1}{11241}} = 0,318$$



$$(d^*, h^*) = 1,089 \pm 1,96 * 0,317 = (0,47; 1,71)$$

$$(d^*, h^*) = 1,092 \pm 1,96 * 0,318 = (0,47; 1,72)$$

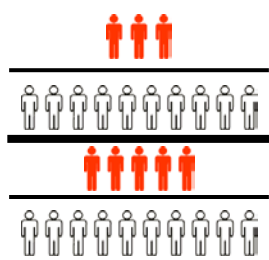
→ Zpětná transformace:

$$(d^{RR}, h^{RR}) = (\exp(d^*), \exp(h^*)) = (1,60; 5,53)$$

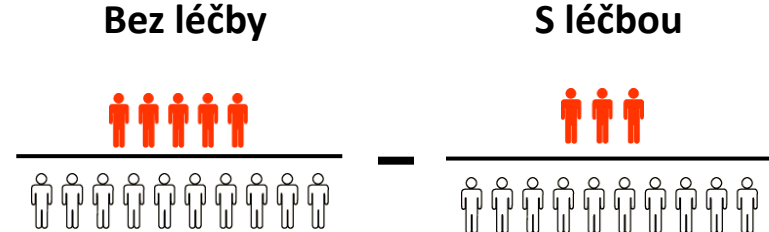
$$(d^{OR}, h^{OR}) = (\exp(d^*), \exp(h^*)) = (1,60; 5,58)$$

# Další způsoby vyjádření rozdílu rizika

## → Relativní redukce rizika (RRR)

$$\text{RRR} = 1 - \text{RR} = 1 - \frac{\frac{\text{3}}{\text{10}}}{\frac{\text{5}}{\text{10}}} = 1 - \frac{\text{10}}{\text{5}} = 1 - 0.6 = 40\%$$


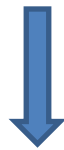
## → Absolutní redukce rizika (ARR)

$$\text{ARR} = \frac{\text{5}}{\text{10}} - \frac{\text{3}}{\text{10}} = 0.2 = 20\%$$


# Další způsoby vyjádření rozdílu rizika

- Počet pacientů, které je potřeba léčit, abychom zabránili výskytu jedné události – „**number needed to treat**“ (NNT).

ARR = 20%    →    Pro snížení počtu událostí o 20 je třeba léčit 100 pacientů.



$$\text{NNT} = \frac{1}{0,2} = \frac{100}{20} = 5$$

NNT = Pro snížení počtu událostí o 1 je třeba léčit 5 pacientů.

# Zvláštní případ RRR – účinnost vakcíny (vaccine efficacy)

- Hodnotíme dvojitě zaslepenou placebem kontrolovanou studii zaměřenou na účinnost bivalentní vakcíny proti incidentní HPV infekci (Harper a kol., 2004)
- *According to protocol group*, 18 měsíců

HPV infekce	Skupina		Celkem
	Vakcinace	Placebo	
Ano	2	23	25
Ne	364	332	696
Celkem	366	355	721

$$VE = 1 - \frac{P_1}{P_0} = 1 - \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = 1 - \frac{\frac{2}{2+364}}{\frac{23}{23+332}} = 1 - 0,084 = 91,6$$

relativní riziko



**Riziko infekce u vakcinovaných je pouhých 8,4% ve srovnání s kontrolní skupinou – vakcína předejde 91,6% infekcí**



# Absolutní vs. relativní četnost

- Vyjádření výsledků v relativní formě (procento) má často příjemnou interpretaci, ale může být zavádějící.
- Relativní vyjádření účinnosti by mělo být vždy doprovázeno absolutním vyjádřením účinnosti.
- **Příklad:** Srovnání účinnosti léčiva ve smyslu prevence CMP u kardiaků.
  - Studie 1: Výskyt CMP ve skupině A je 12 %, ve skupině B je 20 %.  
Relativní změna v účinnosti = **40 %**; absolutní změna = **8 %**.
  - Studie 2: Výskyt CMP ve skupině A je 0,9 %, ve skupině B je 1,5 %.  
Relativní změna v účinnosti = **40 %**; absolutní změna = **0,6 %**.
- Výsledkem je rozdílný přínos léčby při stejné relativní účinnosti.

# NNT a absolutní vs. relativní četnost

→ **Příklad:** Srovnání účinnosti léčiva ve smyslu prevence CMP u kardiaků.

Studie 1: Výskyt CMP ve skupině A je 12 %, ve skupině B je 20 %.  
Relativní změna v účinnosti = **40 %**; absolutní změna = **8 %**.

→ 
$$\text{NNT} = \frac{1}{0,08} = \frac{100}{8} = 12,5$$

NNT = Pro snížení počtu událostí o 1 je třeba léčit 13 pacientů.

Studie 2: výskyt CMP ve skupině A je 0,9 %, ve skupině B je 1,5 %.  
Relativní změna v účinnosti = **40 %**; absolutní změna = **0,6 %**.

→ 
$$\text{NNT} = \frac{1}{0,006} = \frac{100}{0,6} = 166,7$$

NNT = Pro snížení počtu událostí o 1 je třeba léčit 167 pacientů.

## 2. Hodnocení vztahu dvou spojených veličin – základy korelace

# Proč hodnotit vztah dvou spojitých veličin?

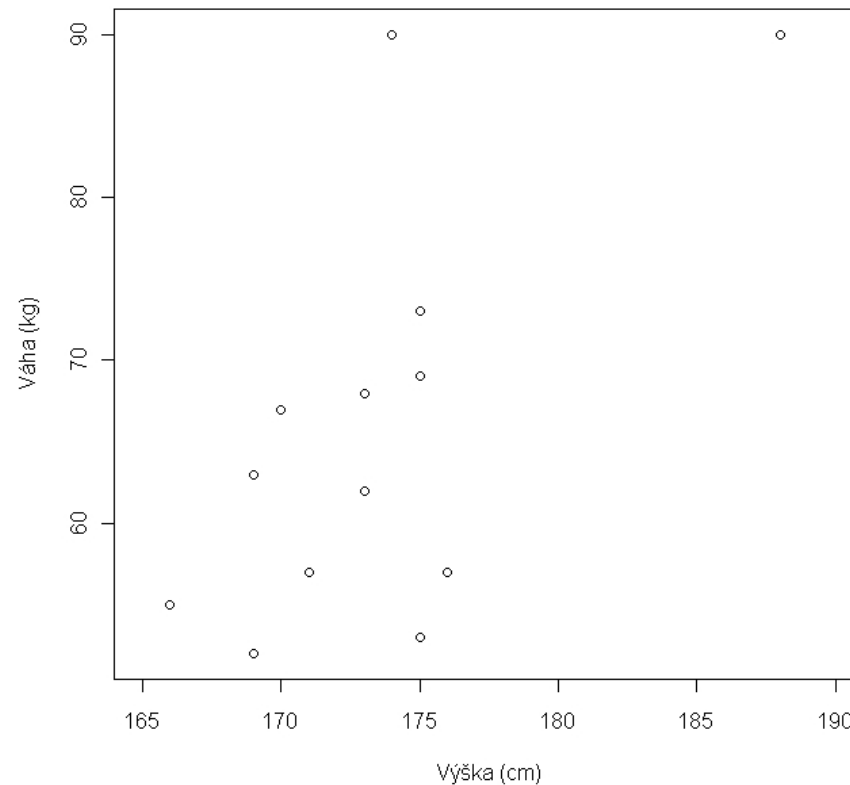
➔ Zatím jsme se zabývali spojitou veličinou v jedné skupině, spojitou veličinou ve více skupinách, diskrétní veličinou v jedné skupině, diskrétní veličinou ve více skupinách, dvěma diskrétními veličinami v jedné skupině.

➔ Teď se chceme zabývat dvěma spojitými veličinami v jedné skupině:

- 1. Chceme zjistit, jestli mezi nimi existuje vztah** – např. jestli vyšší hodnoty jedné veličiny znamenají nižší hodnoty jiné veličiny.
- 2. Chceme predikovat hodnoty jedné veličiny na základě znalosti hodnot jiných veličin.**
- 3. Chceme kvantifikovat vztah mezi dvěma spojitými veličinami** – např. pro použití jedné veličiny na místo druhé veličiny.

# Jak hodnotit vztah dvou spojitých veličin?

- ➔ Nejjednodušší formou je bodový graf (x-y graf).
- ➔ Vztah výšky a váhy studentů Biostatistiky pro matematické biology – jaro 2010:



# Korelace

- **Korelační koeficient** – kvantifikuje míru vztahu mezi dvěma spojitými veličinami ( $X$  a  $Y$ ).
- Standardní metodou je výpočet Pearsonova korelačního koeficientu ( $r$ ).
  - Nabývá hodnot od -1 do 1.
  - Hodnota  $r$  je kladná, když vyšší hodnoty  $X$  souvisí s vyššími hodnotami  $Y$ , a naopak je záporná, když nižší hodnoty  $X$  souvisí s vyššími hodnotami  $Y$ .
  - Charakterizuje linearitu vztahu mezi  $X$  a  $Y$  – jinak řečeno variabilitu kolem lineárního trendu.
  - Hodnoty 1 nebo -1 získáme, když body  $x$ - $y$  grafu leží na přímce.



# Pearsonův korelační koeficient ( $r$ )

→ Předpokládáme realizaci dvourozměrného náhodného vektoru o rozsahu  $n$ :

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix} \quad (\text{máme dvojice hodnot, které patří k sobě – charakterizují } i\text{-tý subjekt})$$

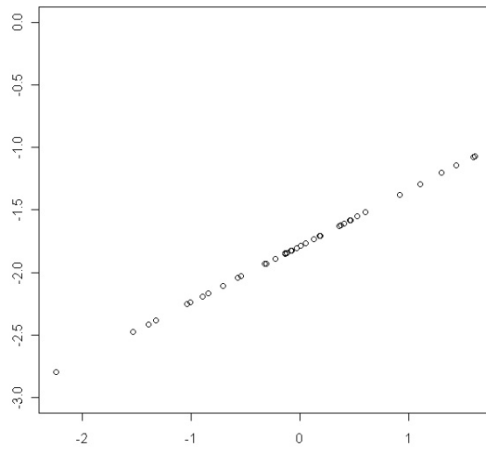
→ Pearsonův korelační koeficient:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

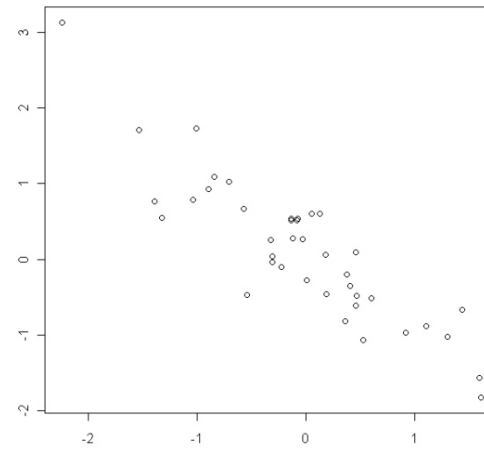
→ kde  $\bar{x}$  a  $\bar{y}$  jsou výběrové průměry,  $s_x$  a  $s_y$  jsou výběrové směrodatné odchylky.



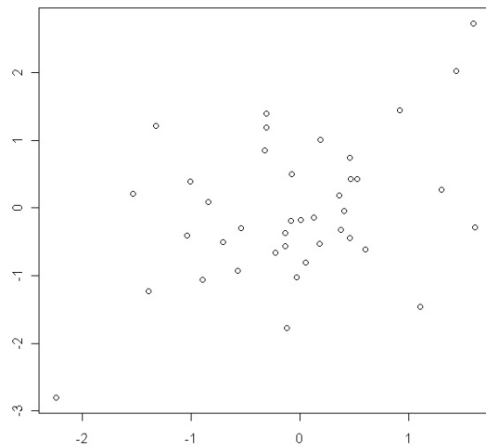
# Pearsonův korelační koeficient ( $r$ )



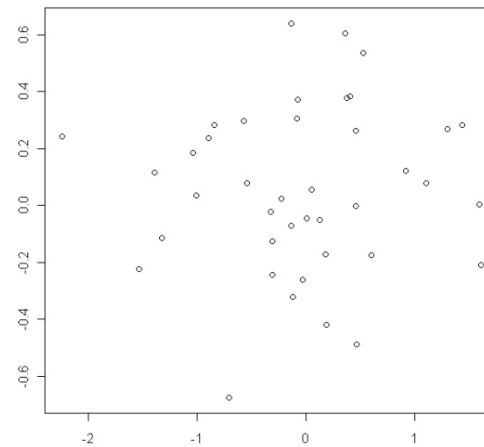
$r = 1,0$



$r = -0,9$



$r = 0,4$

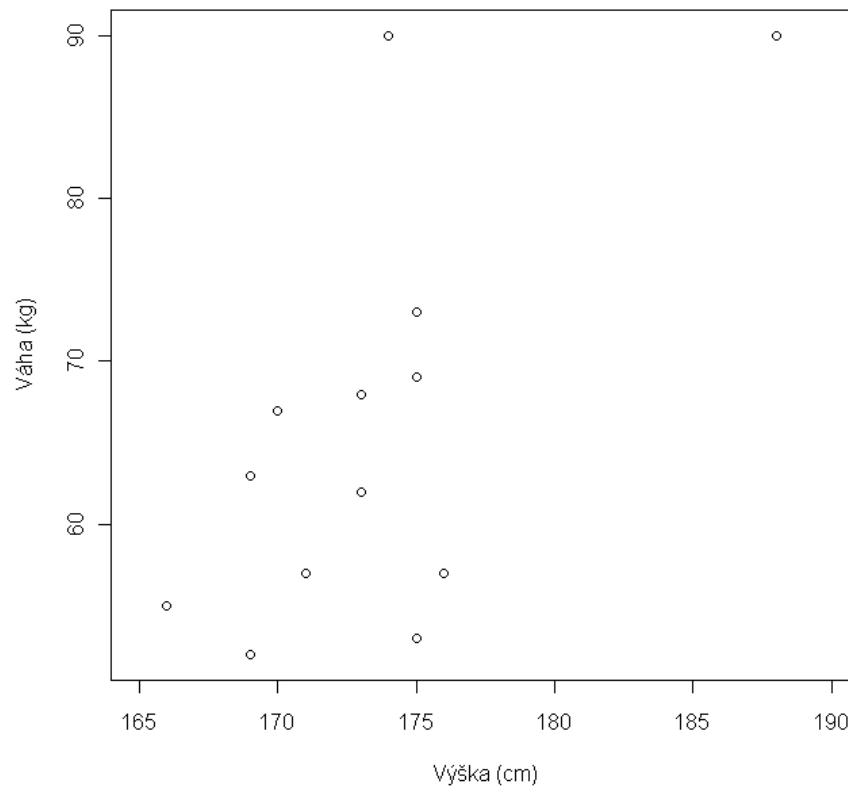


$r = 0,05$



# Příklad – Pearsonův korelační koeficient ( $r$ )

➔ Vztah výšky a váhy studentů Biostatistiky pro matematické biologie – jaro 2010:



$$r = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{(n-1) s_x s_y}$$

$$\sum_{i=1}^n x_i y_i = 148\,929$$

$$n \bar{x} \bar{y} = 148\,417,2$$

$$s_x = 5,3$$

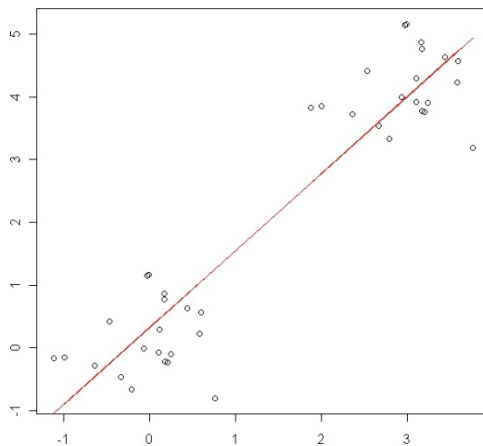
$$s_y = 12,5$$

$$r = \frac{148\,929 - 148\,417,2}{(13-1) * 5,3 * 12,5} = 0,64$$

# Problémy s výpočtem $r$

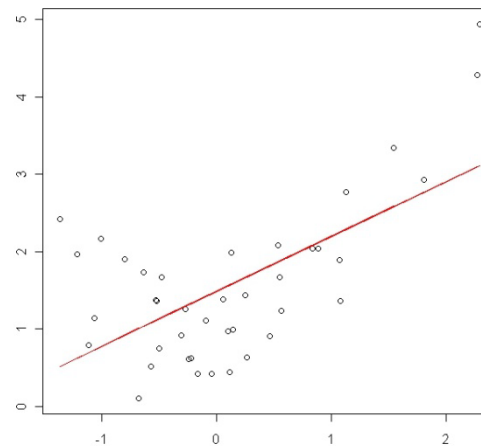
- ➔ Pearsonův korelační koeficient lze vypočítat na jakýchkoliv datech.
- ➔ Pokud však budeme chtít jakkoliv rozhodovat o vlastnostech  $r$  (interval spolehlivosti, testování hypotéz), musíme učinit předpoklad o normalitě hodnocených veličin.

### Více skupin



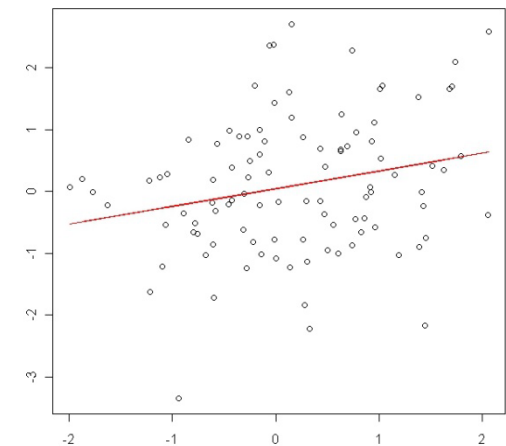
$$r = 0,93$$
$$p < 0,001$$

### Nelineární vztah



$$r = 0,63$$
$$p < 0,001$$

### Velikost výběru



$$r = 0,23$$
$$p = 0,019$$



# Interval spolehlivosti pro $r$

- Výběrové rozdělení koeficientu  $r$  není normální, pro výpočet IS je třeba ho transformovat:

$$w = \frac{1}{2} \ln \frac{1+r}{1-r}$$

- Veličina  $w$  má normální rozdělení se standardní chybou přibližně:  $SE(w) = 1/\sqrt{n-3}$

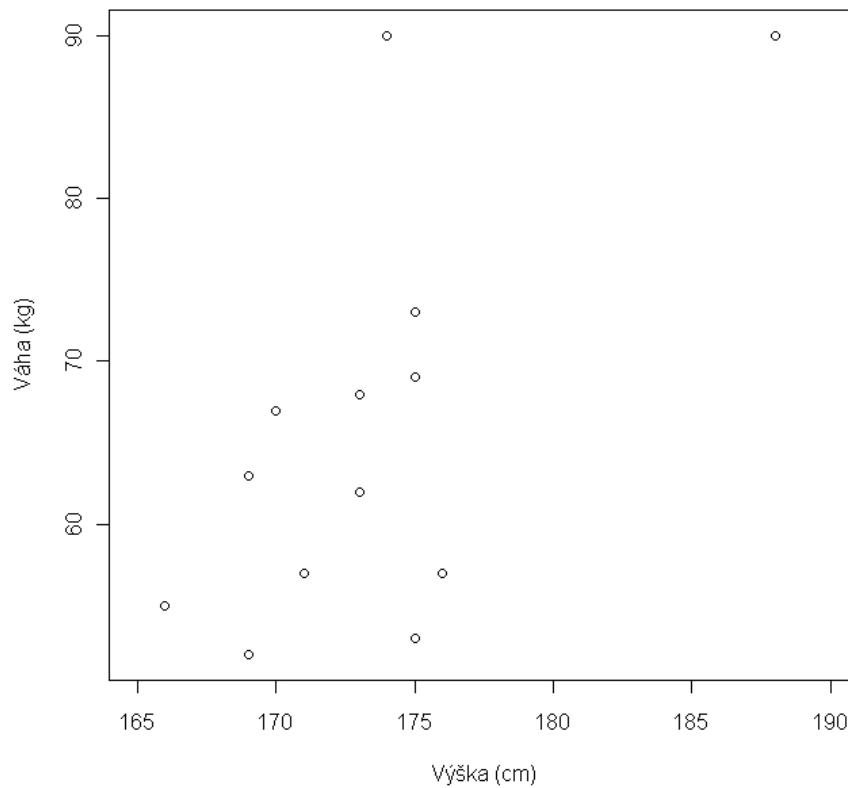
- 100(1- $\alpha$ )% IS pro  $w$  má tvar:  $(d^*, h^*) = w \pm z_{1-\alpha/2} / \sqrt{n-3}$

- 100(1- $\alpha$ )% IS pro  $r$  pak dostaneme zpětnou transformací:

$$(d, h) = \left( \frac{\exp(2d^*) - 1}{\exp(2d^*) + 1}, \frac{\exp(2h^*) - 1}{\exp(2h^*) + 1} \right)$$

# Příklad – interval spolehlivosti pro $r$

➔ Vztah výšky a váhy studentů Biostatistiky pro matematické biologie – jaro 2010:



$$r = 0,64$$

$$w = \frac{1}{2} \ln \frac{1 + 0,64}{1 - 0,64} = 0,758$$

$$SE(w) = 1 / \sqrt{10} = 0,316$$

$$(d^*, h^*) = w \pm z_{1-\alpha/2} SE(w) = (0,138; 1,377)$$

$$(d, h) = \left( \frac{\exp(2d^*) - 1}{\exp(2d^*) + 1}, \frac{\exp(2h^*) - 1}{\exp(2h^*) + 1} \right)$$

$$(d, h) = (0,14; 0,88)$$

# Test hypotézy $H_0: r = 0$

→ Předpokládáme realizaci dvourozměrného náhodného vektoru o rozsahu  $n$ :

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix} \quad \text{Předpokládáme normalitu } X \text{ i } Y!$$

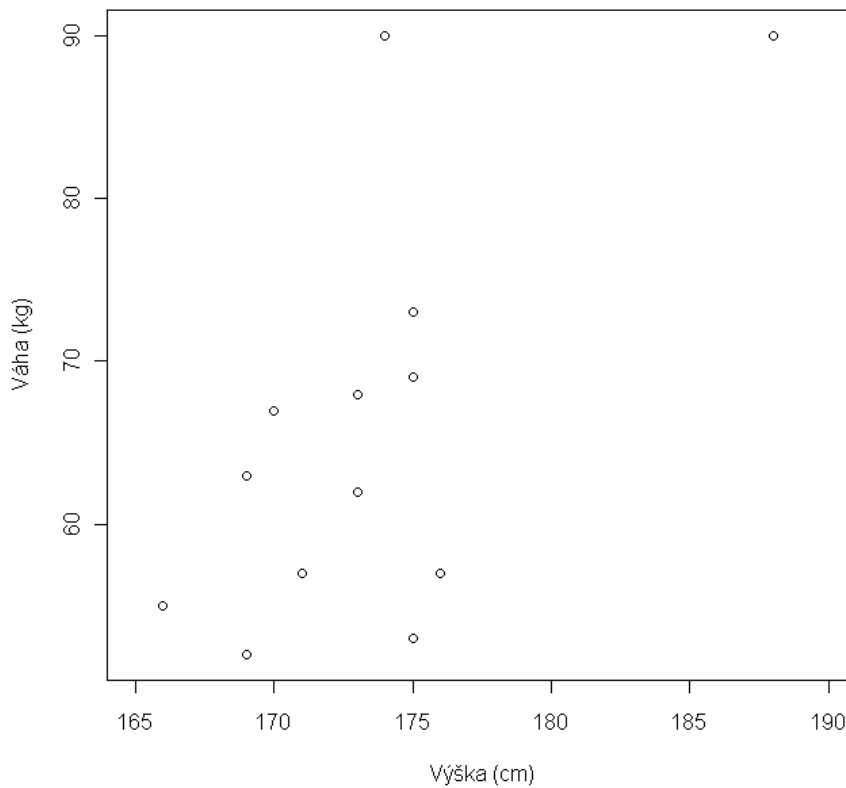
→ Za platnosti nulové hypotézy má statistika  $T = r \sqrt{\frac{n-2}{1-r^2}}$   $t$  rozdělení pravděpodobnosti s  $n - 2$  stupni volnosti.

→ Pro oboustrannou alternativu zamítáme  $H_0$  na hladině významnosti  $\alpha = 0,05$ , když hodnota testové statistiky přesáhne v absolutní hodnotě kvantil  $t_{1-\alpha/2}^{(n-2)}$

→ Tuto testovou statistiku nelze použít pro testování hypotézy  $H_0: r = r_0 \neq 0$

# Příklad – test hypotézy $H_0: r = 0$

➔ Vztah výšky a váhy studentů Biostatistiky pro matematické biologie – jaro 2010:



$$r = 0,64$$

$$T = r \sqrt{\frac{n-2}{1-r^2}} = 0,64 \sqrt{\frac{13-2}{1-0,64^2}} = 2,76$$

$$H_1 : r \neq 0 \quad \longrightarrow \quad t_{1-\alpha/2}^{(n-2)} = t_{0,975}^{(11)} = 2,20$$

$$T = 2,76 > 2,20 = t_{0,975}^{(11)}$$

➔ **Zamítáme  $H_0: r = 0$ .**

# Spearmanův korelační koeficient ( $r_s$ )

→ Pearsonův korelační koeficient je náchylný k odlehlým hodnotám a obecně odchylným od normality. **Spearmanův korelační koeficient** stejně jako řada dalších neparametrických metod **pracuje pouze s pořadími** pozorovaných hodnot.

→ Máme náhodný výběr rozsahu  $n$ :  $\left(\begin{matrix} x_1 \\ y_1 \end{matrix}\right), \left(\begin{matrix} x_2 \\ y_2 \end{matrix}\right), \dots, \left(\begin{matrix} x_n \\ y_n \end{matrix}\right)$

→ Definujeme:

$x_{ri}$  – pořadí  $x_i$  mezi hodnotami  $x$ ;  $y_{ri}$  – pořadí  $y_i$  mezi hodnotami  $y$ ;  $d_i = x_{ri} - y_{ri}$ .

→ Spearmanův korelační koeficient:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

→ Vyskytují-li se shodné hodnoty, je nutné použít výpočet pomocí Pearsonova korelačního koeficientu na pořadích.

→ Hodnoty  $r_s$  se pohybují stejně jako u  $r$  od -1 do 1.

# Příklad – Spearmanův korelační koeficient ( $r_s$ )

➔ Vztah výšky a váhy studentů Biostatistiky pro matematické biologie – jaro 2010:

Student	Výška $x_i$	Pořadí výška	Váha $y_i$	Pořadí váha	Rozdíl $d_i$	$d_i^2$
1	175	10	69	10	0	0
2	166	1	55	3	-2	4
3	170	4	67	8	-4	16
4	169	2,5	52	1	1,5	2,25
5	188	13	90	12,5	0,5	0,25
6	175	10	53	2	8	64
7	176	12	57	4,5	7,5	56,25
8	171	5	57	4,5	0,5	0,25
9	173	6,5	68	9	-2,5	6,25
10	175	10	73	11	-1	1
11	173	6,5	62	6	0,5	0,25
12	174	8	90	12,5	-4,5	20,25
13	169	2,5	63	7	-4,5	20,25



# Příklad – Spearmanův korelační koeficient ( $r_s$ )

→ V souboru je hodně shodných hodnot → musíme použít Pearsonovo  $r$  na pořadí.

Student	Pořadí výška	Pořadí váha	Rozdíl $d_i$	$d_i^2$
1	10	10	0	0
2	1	3	-2	4
3	4	8	-4	16
4	2,5	1	1,5	2,25
5	13	12,5	0,5	0,25
6	10	2	8	64
7	12	4,5	7,5	56,25
8	5	4,5	0,5	0,25
9	6,5	9	-2,5	6,25
10	10	11	-1	1
11	6,5	6	0,5	0,25
12	8	12,5	-4,5	20,25
13	2,5	7	-4,5	20,25

$$r = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{(n-1) s_x s_y}$$

$$\sum_{i=1}^n x_i y_i = 721,5$$

$$n \bar{x} \bar{y} = 637$$

$$s_x = 3,86$$

$$s_y = 3,88$$

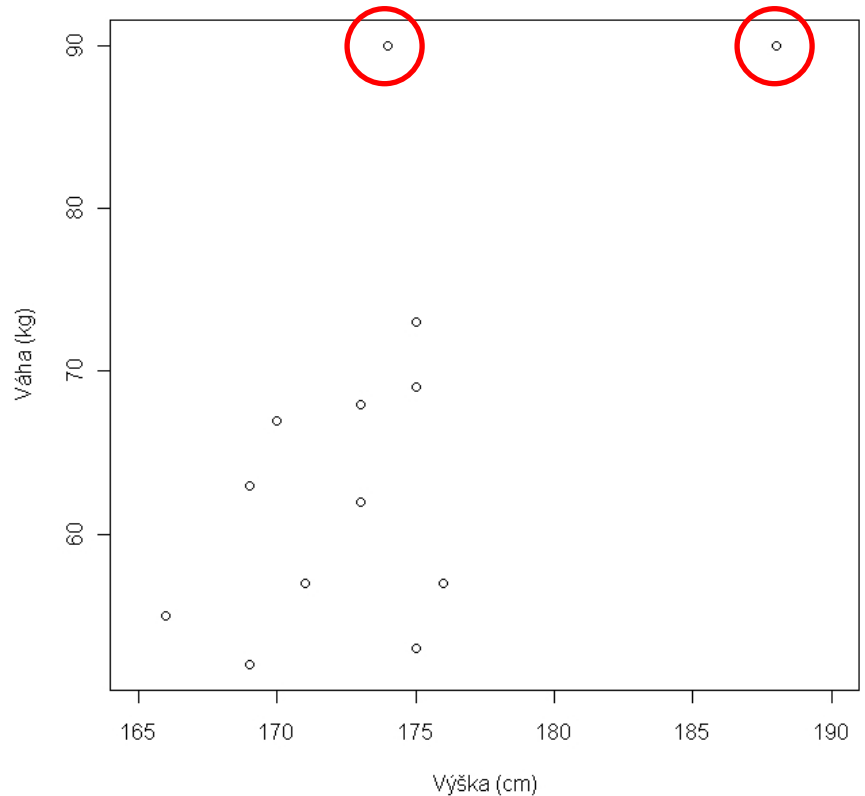
$$r = \frac{721,5 - 637}{(13-1) * 3,86 * 3,88} = 0,47$$

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{6 * 191}{13(13^2 - 1)} = 0,48$$

# Jak to, že nám $r$ a $r_s$ vyšly různě?

→ Původní hodnoty:  $r = 0,64$

→ Pořadí:  $r = 0,47$   
 $r_s = 0,48$



# IS pro $r_s$ a test hypotézy $H_0: r_s = 0$

- ➔ Výběrové rozdělení  $r_s$  je pro výběry s  $n > 10$  stejné jako výběrové rozdělení  $r$ , proto je možné pro konstrukci  $100(1-\alpha)\%$  IS použít metodu pro Pearsonův koeficient.
- ➔ Pro větší vzorky,  $n > 30$ , je možné použít pro ověření hypotézy  $H_0: r_s = 0$  stejnou testovou statistiku jako v případě  $r$ :

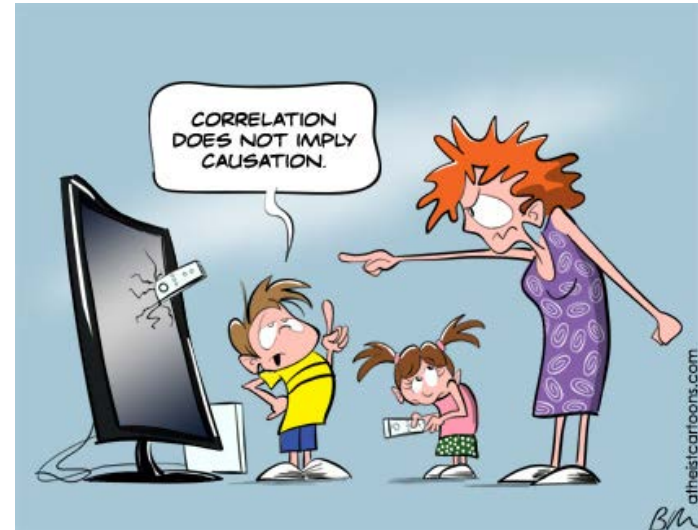
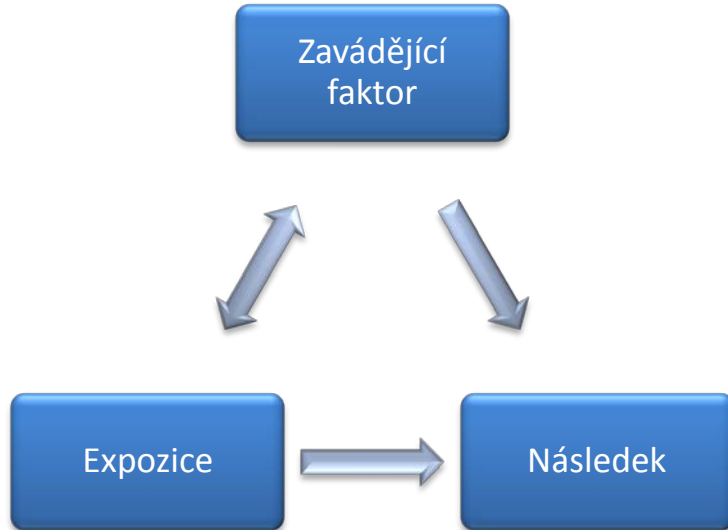
$$T = r_s \sqrt{\frac{n-2}{1-r_s^2}} \sim t^{(n-2)}$$

# Poznámka o $r^2$

- Korelace dvou náhodných veličin se často interpretuje pomocí druhé mocniny Pearsonova korelačního koeficientu:  $r^2$ .
- Hodnota  $r^2$  vyjadřuje, kolik % své variability sdílí jedna veličina s druhou, jinak řečeno, kolik % variability jedné veličiny může být predikováno pomocí té druhé.
- S hodnotou  $r^2$  se setkáte v lineárních modelech.

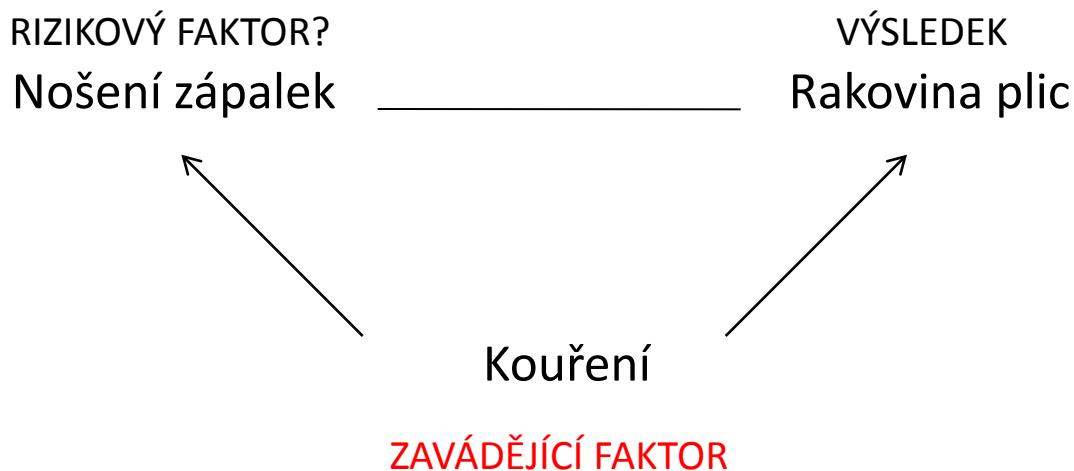
# Klíčové principy – zkreslení

- Pojem **zavádějící faktor** – pro zavádějící faktor současně platí, že
  - přímo nebo nepřímo ovlivňuje sledovaný následek,
  - je ve vztahu se studovanou expozicí ,
  - není mezikrokem mezi expozicí a následkem.



# Zavádějící faktor (confounder)

- Proměnná asociovaná s rizikovým faktorem a kauzálně spojená s výsledkem



- může zcela zatemnit skutečný vztah mezi rizikovým faktorem a výsledkem

# Jak na zavádějící faktory: stratifikace

Rakovina plic	Konzumace alkoholu		Celkem
	Vysoká	Nízká	
Ano	33	27	60
Ne	1667	2273	3940
Celkem	1700	2300	4000

$$OR = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{\frac{33}{1667}}{\frac{27}{2273}} = 1,67$$



**Vysoká konzumace alkoholu je rizikovým faktorem pro vznik rakoviny plic...**

Zdroj: Fundamentals of biostatistics, Rosner 2006

# Jak na zavádějící faktory: stratifikace

## Skupina kuřáků

Rakovina plic	Konzumace alkoholu		Celkem
	Vysoká	Nízká	
Ano	24	6	30
Ne	776	194	970
Celkem	800	200	1000

$$OR = \frac{\frac{24}{6}}{\frac{776}{194}} = 1,00$$

## Skupina nekuřáků

Rakovina plic	Konzumace alkoholu		Celkem
	Vysoká	Nízká	
Ano	9	21	30
Ne	891	2079	2970
Celkem	900	2100	3000

$$OR = \frac{\frac{9}{21}}{\frac{891}{2079}} = 1,00$$

**Ve skutečnosti ani u kuřáků ani u nekuřáků konzumace alkoholu riziko vzniku rakoviny plic nezvyšuje**

Zdroj: Fundamentals of biostatistics, Rosner 2006



# Poděkování...

Rozvoj studijního oboru „Matematická biologie“ PŘF MU Brno je finančně podporován prostředky projektu ESF č. CZ.1.07/2.2.00/07.0318 „Víceoborová inovace studia Matematické biologie“ a státním rozpočtem České republiky



Tomáš Pavlík



Biostatistika