13. Vztah dvou proměnných

Asociační tabulky

Korelace

Regrese

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek, S. Littnerová, J. Kalina

Fisherův exaktní (přímý) test

- Využití ve čtyřpolní tabulce s nízkými četnostmi, které znemožňují použití χ^2-testu.
- Patří mezi neparametrické testy pracující s daty na nominální škále, v nejjednodušší podobě ve dvou třídách: pozitivní/negativní, úspěch/neúspěch apod.
- Nulová hypotéza předpokládá rovnoměrné zastoupení sledovaného znaku u dvou nezávislých souborů.
- Slovo exaktní (přímý) znamená, že se přímo vypočítává pravděpodobnost odmítnutí, resp. platnosti nulové hypotézy.

Fisherův exaktní (přímý) test

- Výpočet probíhá v cyklu:
  - spočítá se parciální pravděpodobnost čtyřpolní tabulky p_1:

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

$$p_1 = \frac{(a+b)! * (c+d)! * (a+c)! * (b+d)!}{N! * a! * b! * c! * d!}$$

  - nejnižší hodnota v tabulce se sníží o jedna při zachování součtů řádků i sloupců,
  - postup se opakuje (výpočet parciálních pravděpodobností $p_2 \ldots p_n$)
  - cyklus končí ve chvíli, kdy je v nejnižším poli tabulky 0.
- p-hodnota testu je součtem parciálních pravděpodobností.

Vyjádření rizik ve čtyřpolní tabulce - motivace

- Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS). Výsledky dány v tabulce.
- Pomocí Pearsonova chí-kvadrátu 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 a poměr šancí.

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

Relativní riziko = Relative Risk (RR)

- 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		
	Experimentální	Kontrolní	Celkem
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 ve čtyřpolní tabulce:

SIDS	Věk matky		
	Do 25 let	25 a více let	Celkem
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 rodících ve vyšším věku.

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**

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

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

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

Příklad: odds ratio

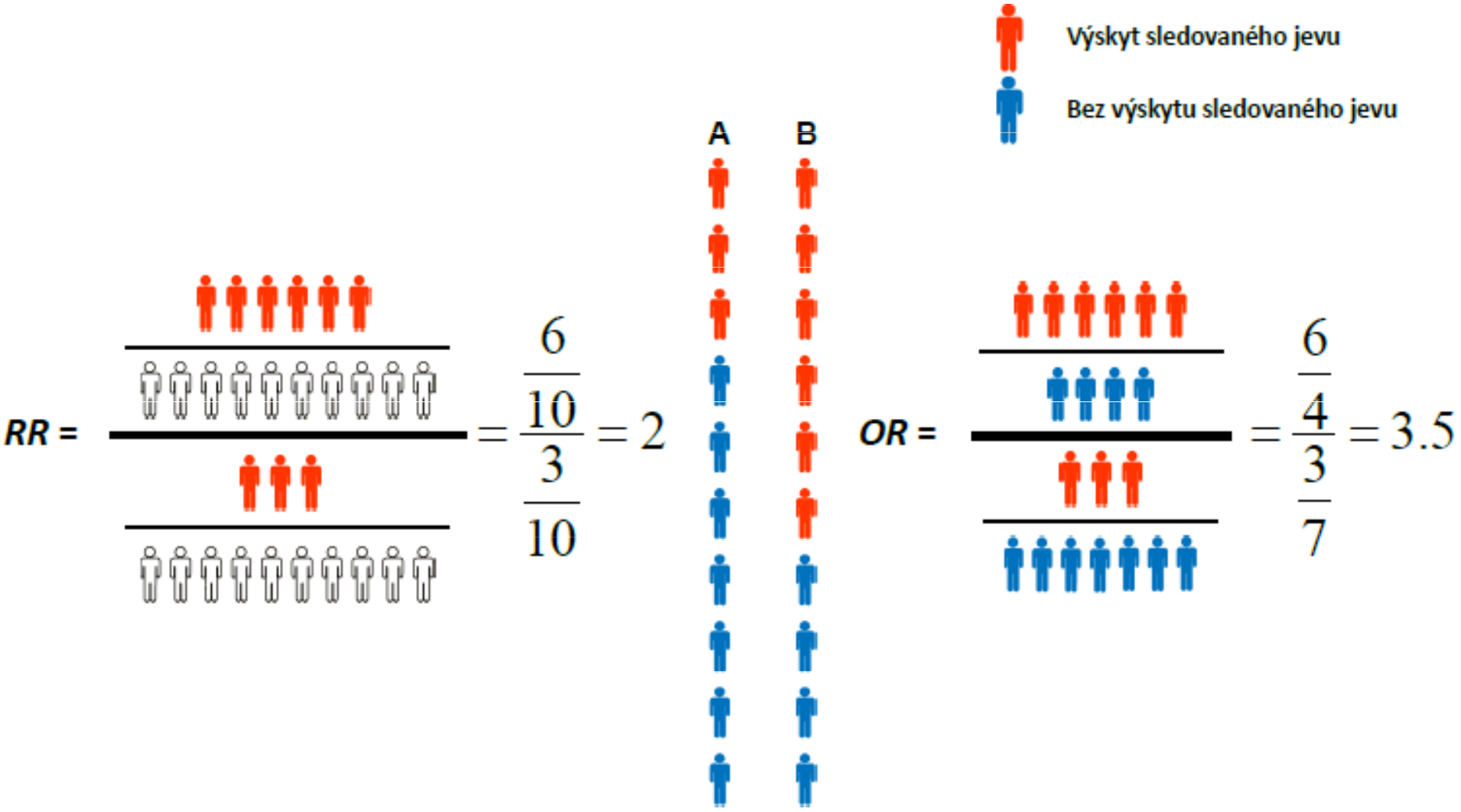
- Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS). Výsledky dány ve čtyřpolní tabulce:

SIDS	Věk matky		
	Do 25 let	25 a více let	Celkem
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 rodících 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:

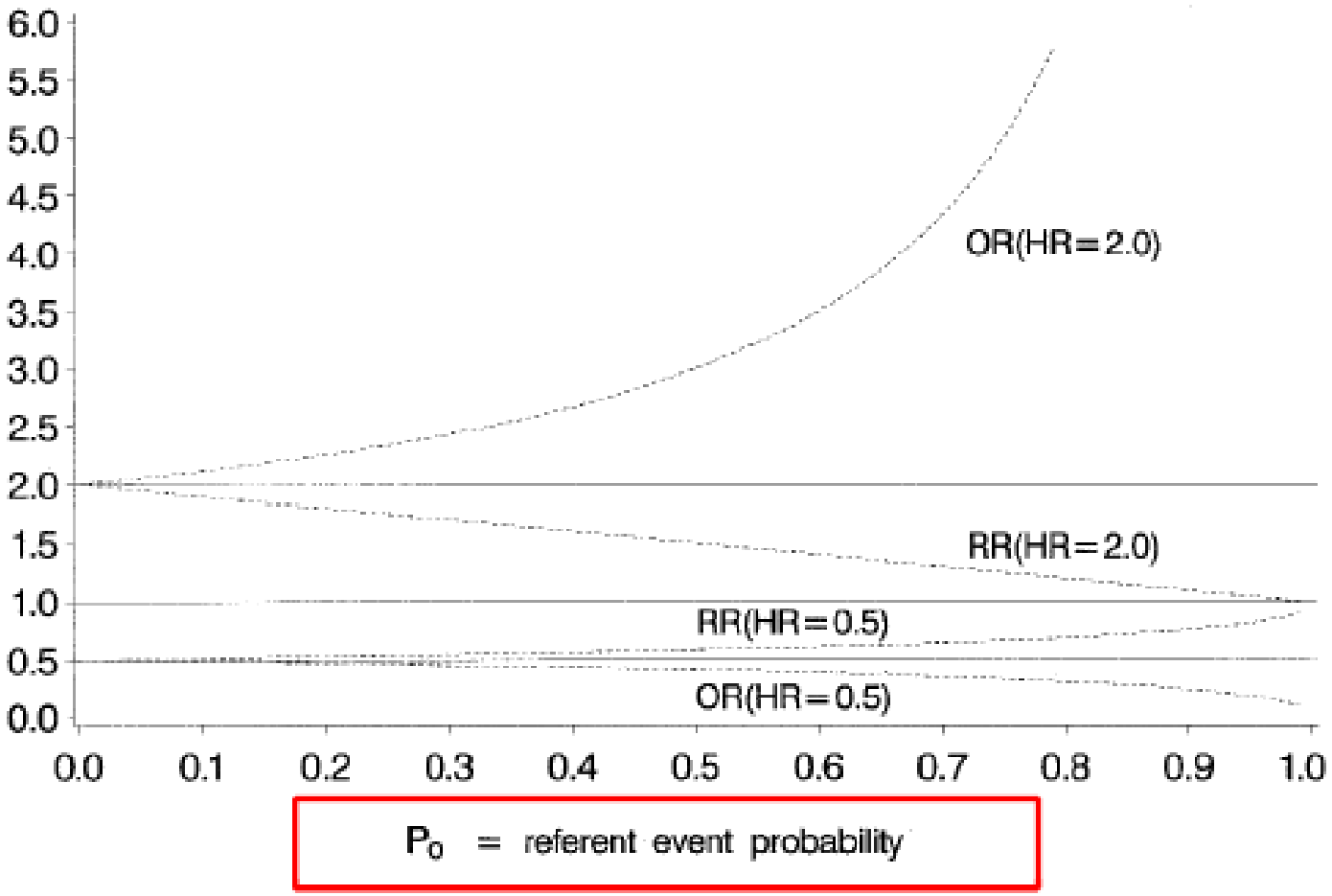
Zubní kaz	Pití slazených nápojů		
	Ano	Ne	Celkem
Ano	34	19	53
Ne	16	31	47
Celkem	50	50	100

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

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

Srovnání RR a OR

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



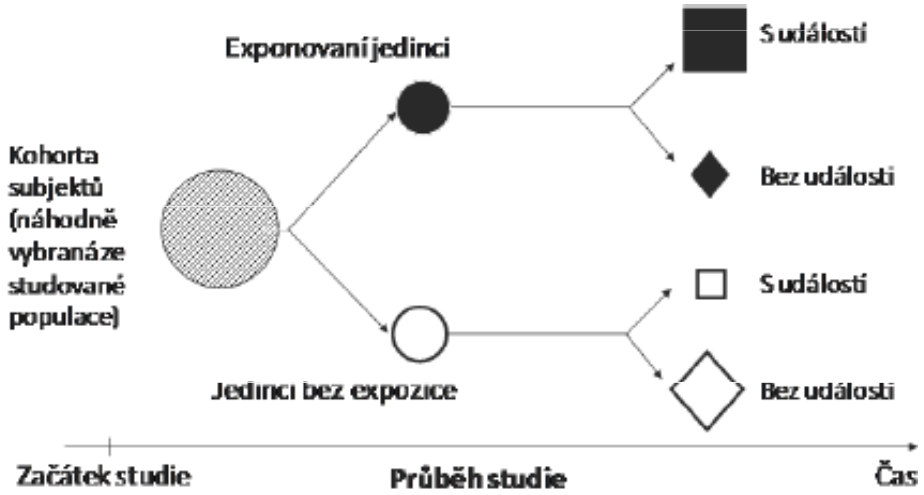
Výhody a nevýhody RR a OR

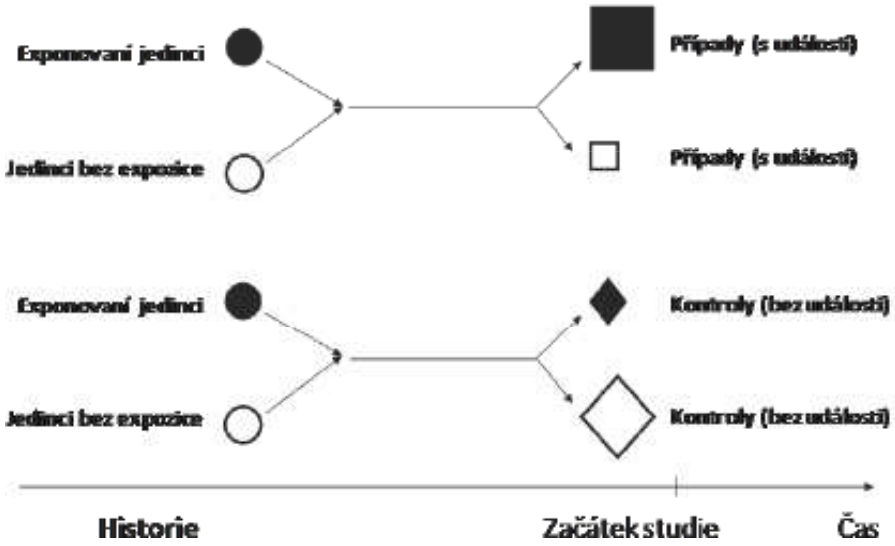
- Nevýhoda *OR:*
  - *obtížná interpretace.*

- 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*.**

Korelace a regrese

- 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 namísto druhé veličiny.**

Korelace a regrese

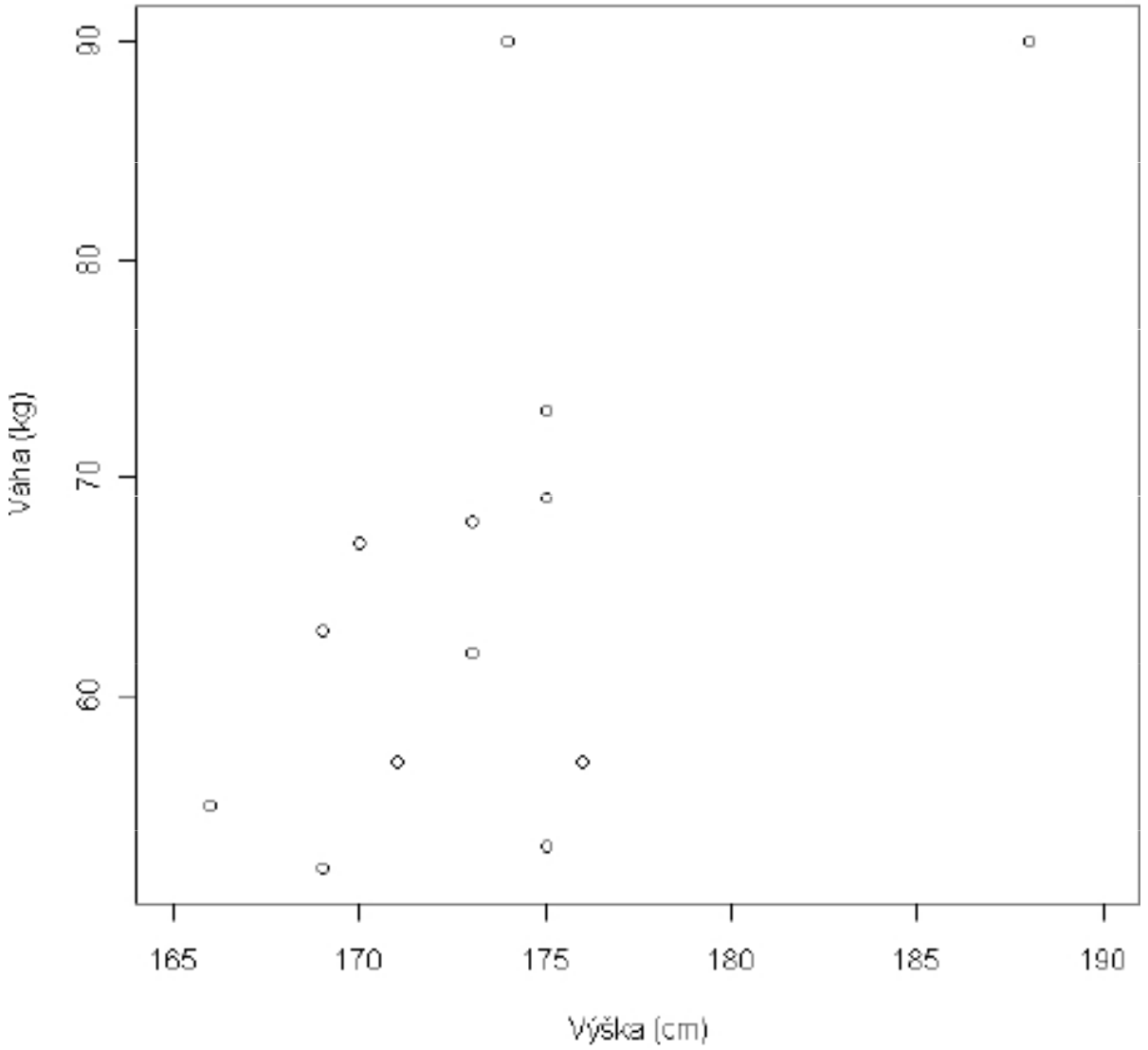
- Korelační analýza je využívána pro vyhodnocení míry vztahu dvou spojitých proměnných. Obdobně jako jiné statistické metody, i korelace mohou být parametrické nebo neparametrické.
- Regresní analýza vytváří model vztahu dvou nebo více proměnných, tedy jakým způsobem jedna proměnná (vysvětlovaná) závisí na jiných proměnných (prediktorech). Regresní analýza je obdobně jako ANOVA nástrojem pro vysvětlení variability hodnocené proměnné.

Korelace

- K měření těsnosti lineárního vztahu 2 spojitých proměnných
  r = 0 → nekorelované
  r > 0 → kladně korelované
  r < 0 → záporně korelované
- H_0: proměnné X, Y jsou stochasticky nezávislé náhodné veličiny
  (r = 0)
  H_A: proměnné X, Y nejsou stochasticky nezávislé náhodné veličiny
  (r ≠ 0)
- Parametrický korelační koeficient:
  Pearsonův kor. koef. (dvourozměrné normální rozložení)
- Neparametrické korelační koeficienty:
  Spearmanův (pořadový) kor. koef., Kendallovo tau.

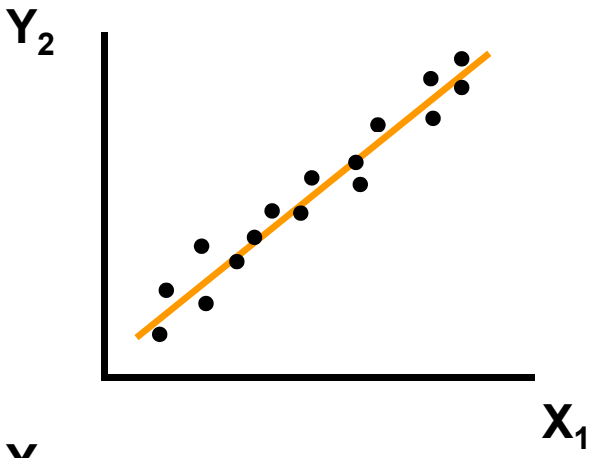
Vizuální hodnocení vztahu dvou proměnných

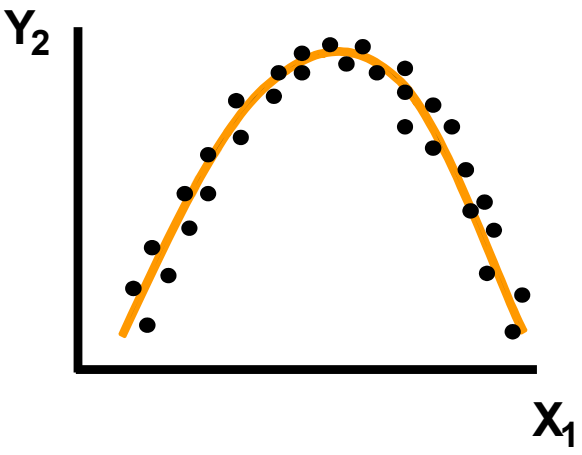
- Nejjednodušší formou je bodový graf (x-y graf), tzv. scatterplot.
- Vztah výšky a váhy studentů Biostatistiky pro matematické biology – jaro 2010:

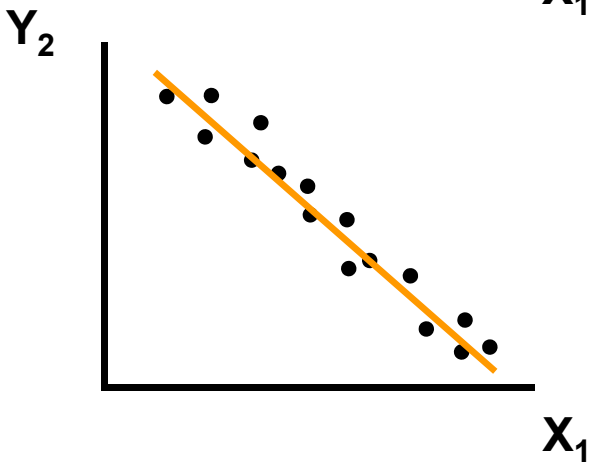


Základy korelační analýzy - I.

Korelace – vztah (závislost) dvou znaků (parametrů)







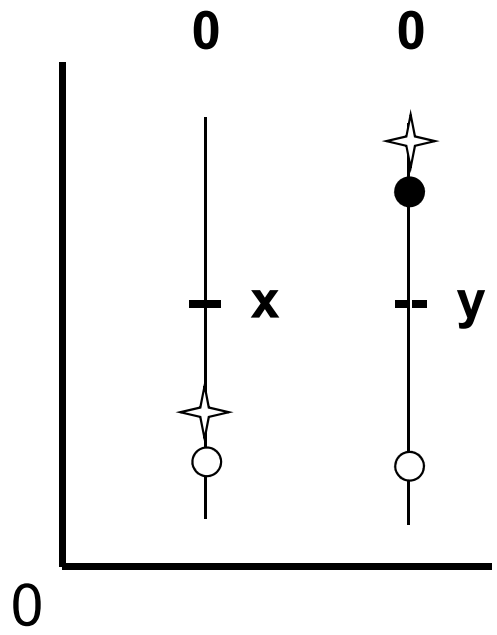
X_2 \ X_1	ANO	NE
ANO	a	b
NE	c	d

Základy korelační analýzy - II.

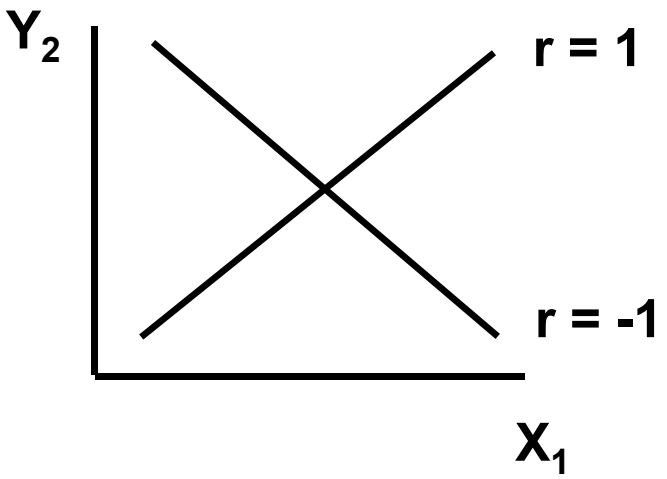
Parametrické míry korelace

Kovariance

$$Cov(x,y) = E(x_i - \bar{x}).(y_i - \bar{y})$$



Pearsonův koeficient korelace



Základy korelační analýzy - III.

P_I (zem)	10	14	15	32	40	20	16	50
P_I (rostl.)	19	22	26	41	35	32	25	40

$$I = 1,\ldots..,\ n; n = 8; v = 6$$

$$r = \frac{Cov\,(x,y)}{S_x.S_y} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]}} = 0,7176$$

I. $H_0 : \rho = \phi : \alpha = 0,05$

tab : $r(v=6) = 0,7076$

II. $H_0 : \rho = \phi$ $\qquad t = \left[\frac{r}{\sqrt{1-r^2}}\right] \cdot \sqrt{n-2} \qquad v = n-2$

$$t = \frac{0,7176}{0,6965} \cdot \sqrt{6} = 2,524$$

tab : $t_{0,975}^{(n-2)} = 2,447$

$P \leq 0,05$

Základy korelační analýzy - IV.
Srovnání dvou korelačních koeficientů (r)

1. $n_1 = 1258$

$r_1 = 0{,}682$

2. $n_2 = 462$

$r_2 = 0{,}402$

Krevní tlak **x** koncentrace kysl. radikálů

$$Z_i = 1.1513 \cdot \log \frac{(1 + r_i)}{(1 - r_i)}$$

$Z_1 = 0{,}833$ $\qquad$ $Z_2 = 0{,}426$

Test: $H_0 : \rho_1 = \rho_2$; $\alpha = 0{,}05$

$$Z = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0{,}407}{0{,}0545} = 7{,}461$$

tabulky : $Z_{0{,}975} = 1{,}96$

7,461 >> 1,96 => P << 0,01

Základy korelační analýzy - V. Neparametrická korelace (rs)

P_I v půdě	1	2	3	6	7	5	4	8
P_I v rostl.	1	2	4	8	6	5	3	7
d_I	0	0	1	2	-1	0	-1	-1

i = 1, ….. n; n = 8 => v = 6

$$r_s = 1 - \frac{6 \cdot \sum di^2}{n\left(n^2 - 1\right)} = 0,9048$$

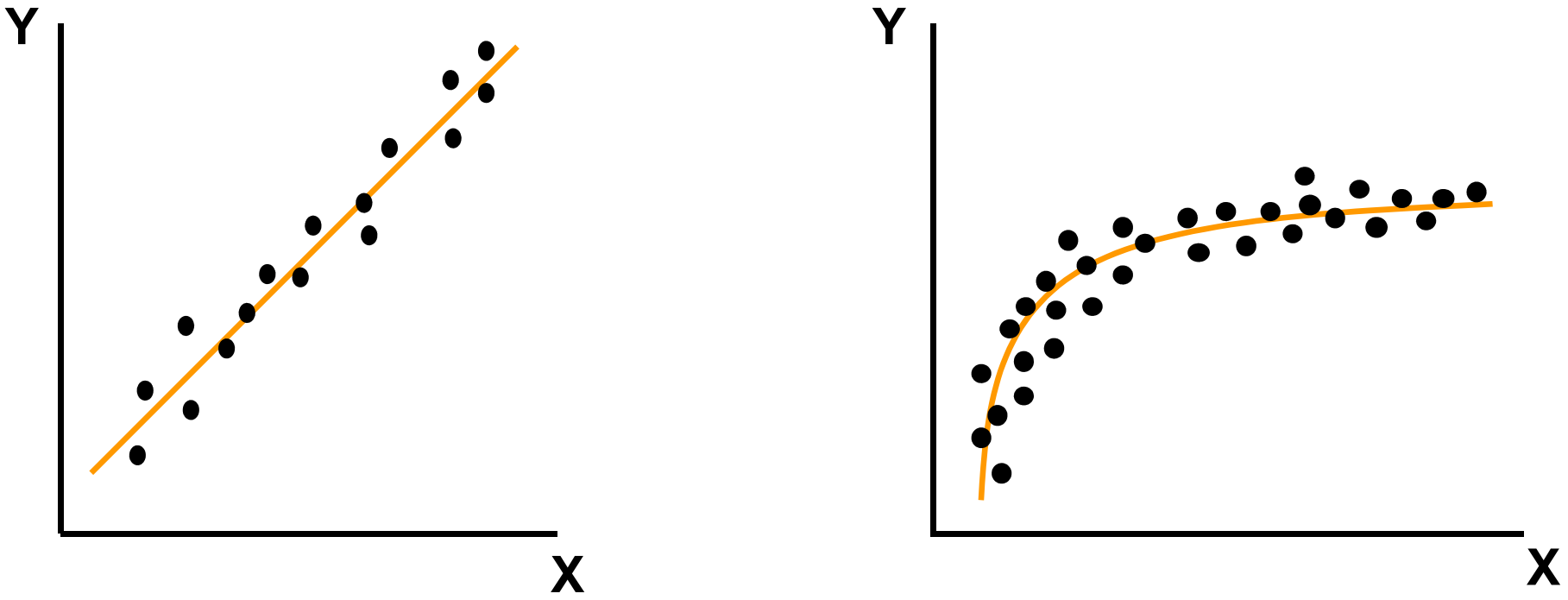
tab : $r_s(v = 6) = 0,89$

Pacient č.	1	2	3	4	5	6	7
Lékař 1	4	1	6	5	3	2	7
Lékař 2	4	2	5	6	1	3	7
d_I	0	-1	1	-1	2	-1	0

$$r_s = 1 - \frac{6 \cdot 8}{7\left(49 - 1\right)} = 0,857$$

P = 0,358

Korelace v grafech I.



Vztahy velmi často implikují funkční vztah mezi Y a X.

$$Y = a + b \,.\, X$$

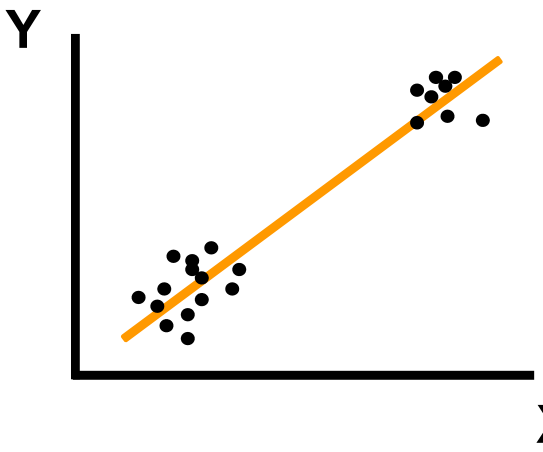
$$Y = a + b_1 \,.\, X_1 + b_2 \,.\, X_2 + b_3 \,.\, X_3$$

$$Y = a + b_1 \,.\, X_1 + b_2 \,.\, X_2$$

$$Y = a + b_1 \,.\, X_1 + b_2 \,.\, X_2 + b_3 \,.\, X_1 \,.\, X_2$$

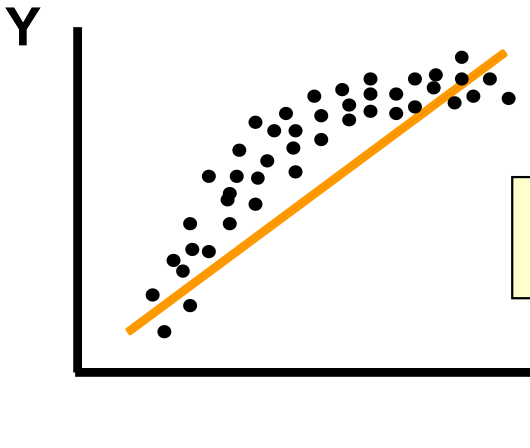
Korelace v grafech II.

Problém rozložení hodnot



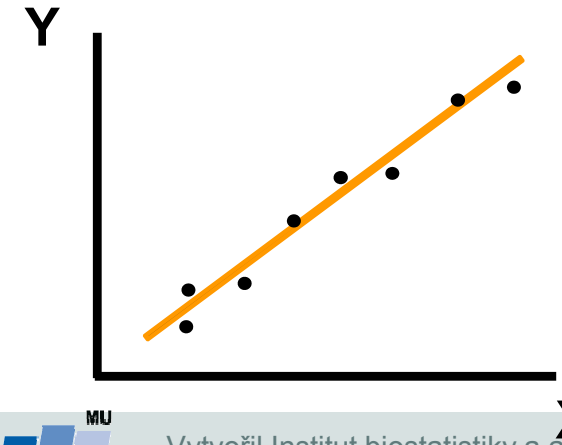
r = 0,981
(p < 0,001)

Problém typu modelu

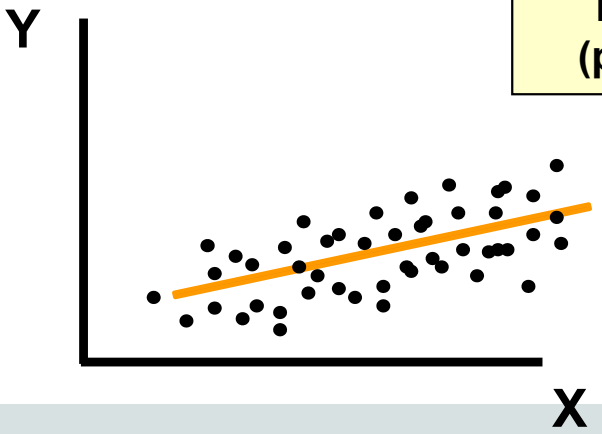


r = 0,761
(p < 0,032)

Problém velikosti vzorku



r = 0,891
(p < 0,214)



r = 0,212
(p < 0,008)

Jednoduchá lineární regrese

- V případě existence vzájemného vztahu (korelace) lze tento vztah podrobněji popsat.
- Cíl regresní analýzy: popsat závislost hodnot proměnné Y na hodnotách proměnné X.
- V případě lineární regrese je tento popis dán lineárním modelem tvaru $y = ax + b$.
- Existují i techniky nelineární regrese.
- Nemáme-li dostatek informací k teoretickému souboru, snažíme se odhadnout typ funkce pomocí dvourozměrného diagramu.

Předpoklady lineární regrese

- Hlavním předpokladem je splnění požadavků Gauss-Markovovy věty:

1. $E(\varepsilon_i) = 0,$
2. $V(\varepsilon_i) = \sigma^2 < \infty,$
3. $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$

- Splnění těchto předpokladů je zajištěno v případě, kdy jsou rezidua normálně rozdělena, nezávislá na prediktorech (které jsou nezávislé).