

Hodnocení vztahu dvou spojitých veličin

– základy korelace

Proč hodnotit vztah dvou spojitých veličin?

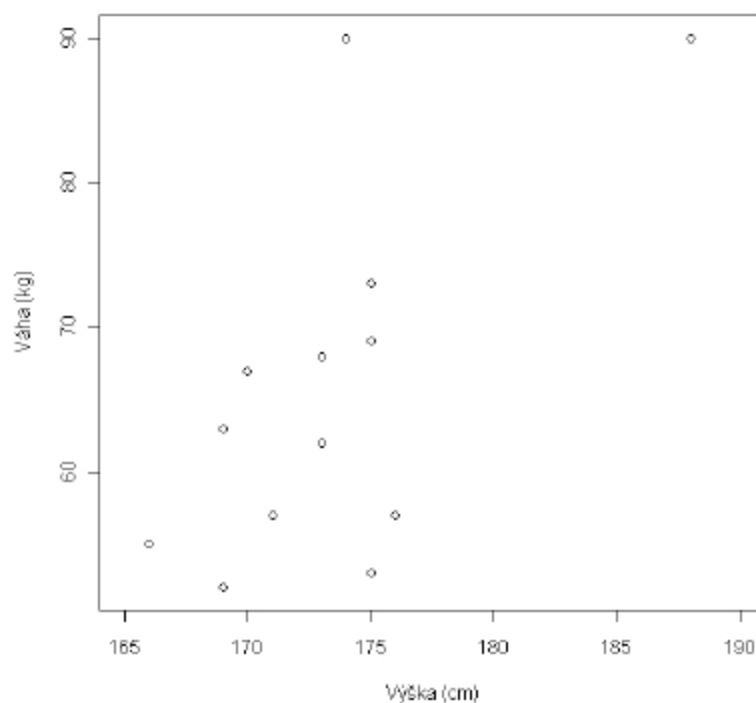
➤ Zatím jsme se zabývali spojitou veličinou v jedné skupině, spojitou veličinou ve více skupinách, diskrétní veličinou v jedné skupině, diskrétní veličinou ve více skupinách, dvěma diskrétními veličinami v jedné skupině.

➤ 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 biologie – 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 :

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

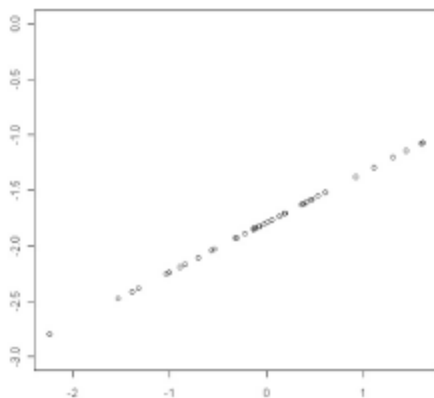
- Pearsonův korelační koeficient:

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

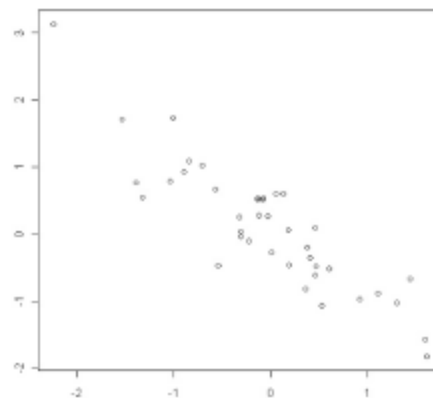
- kde \bar{x} a \bar{y} jsou výběrové průměry, s_x a s_y jsou výběrové směrodatné odchylky.



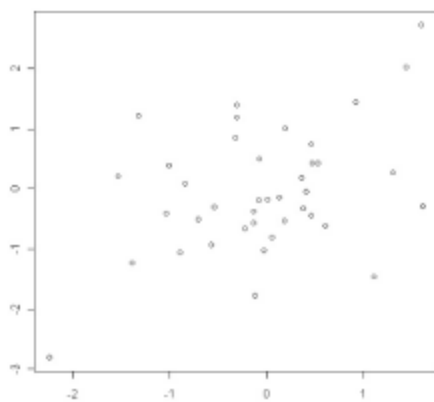
Pearsonův korelační koeficient (r)



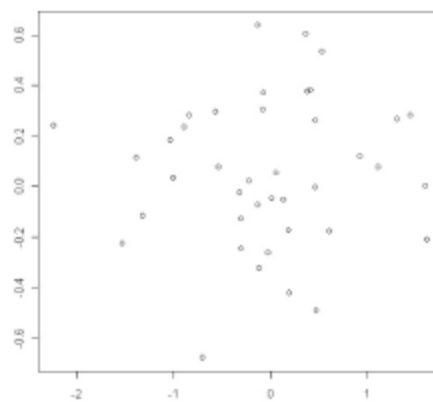
$r = 1,0$



$r = -0,9$



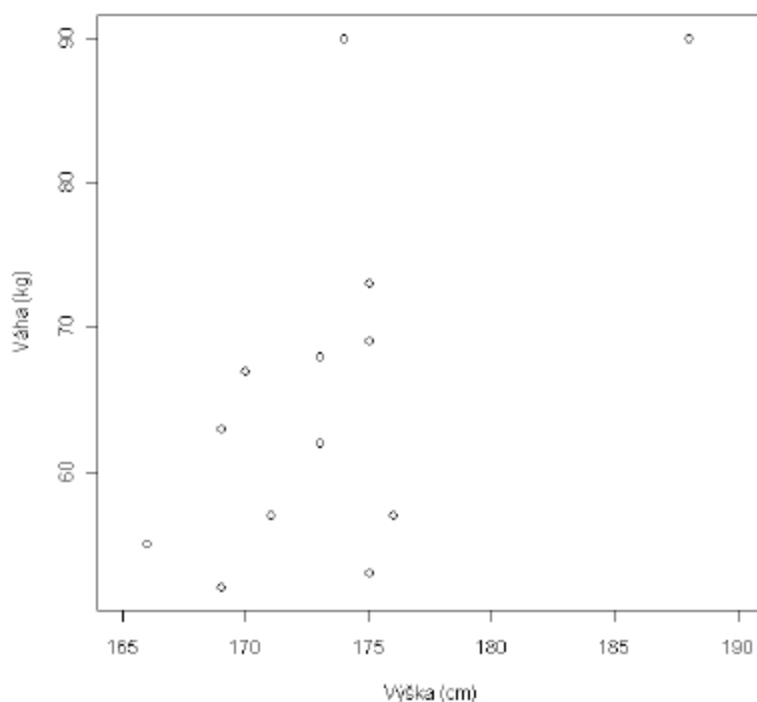
$r = 0,4$



$r = 0,05$

Příklad – Pearsonův korelační koeficient (r)

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



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

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

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

$$s_x = 5,3$$

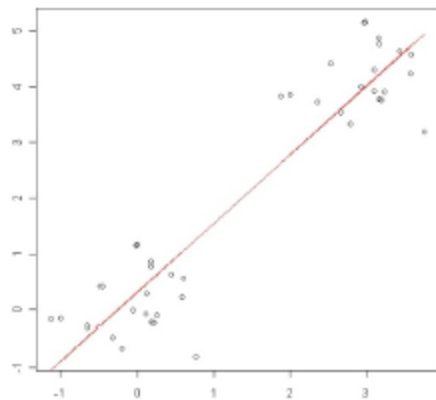
$$s_y = 12,5$$

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

Problémy s výpočtem r

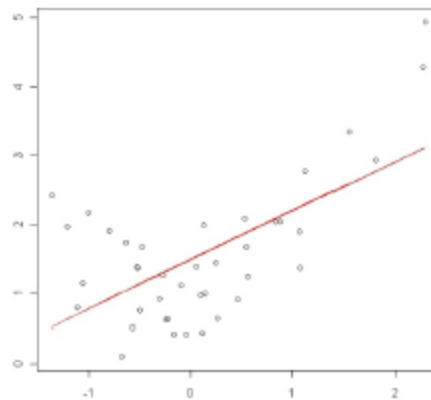
- Pearsonův korelační koeficient lze vypočítat na jakýchkoliv datech.
- Pokud však budeme chtít jakkoliv rozhodovat o vlastnostech r (interval spolehlivosti, testování hypotéz), musíme učinit předpoklad o normalitě hodnocených veličin.

Více skupin



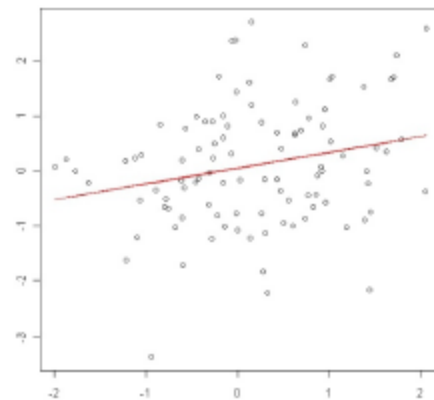
$$r = 0,93$$
$$p < 0,001$$

Nelineární vztah



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

Velikost výběru



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



Test hypotézy $H_0: r = 0$

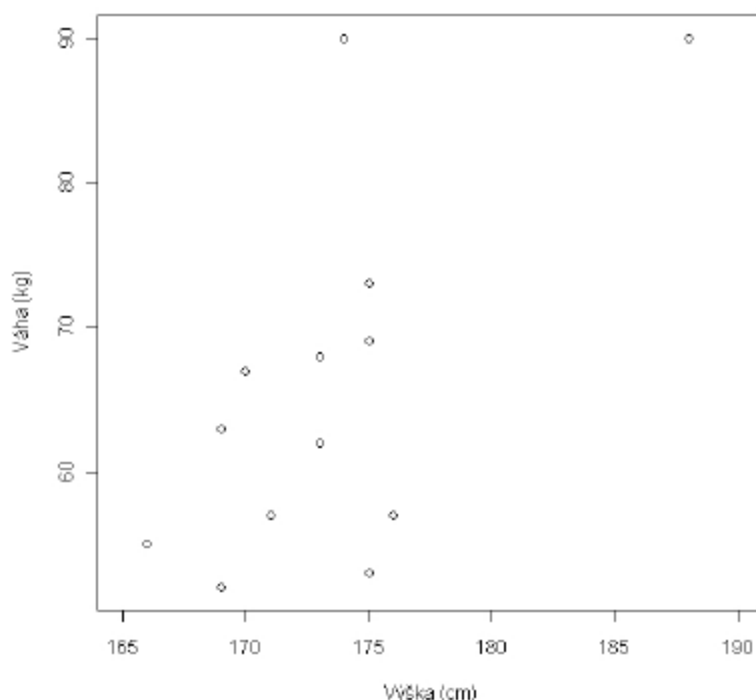
- Předpokládáme realizaci dvourozměrného náhodného vektoru o rozsahu n :

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

- 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 odchylnám 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, doporučuje se použití 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 → lépe 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

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

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

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

$$s_x = 3,86$$

$$s_y = 3,88$$

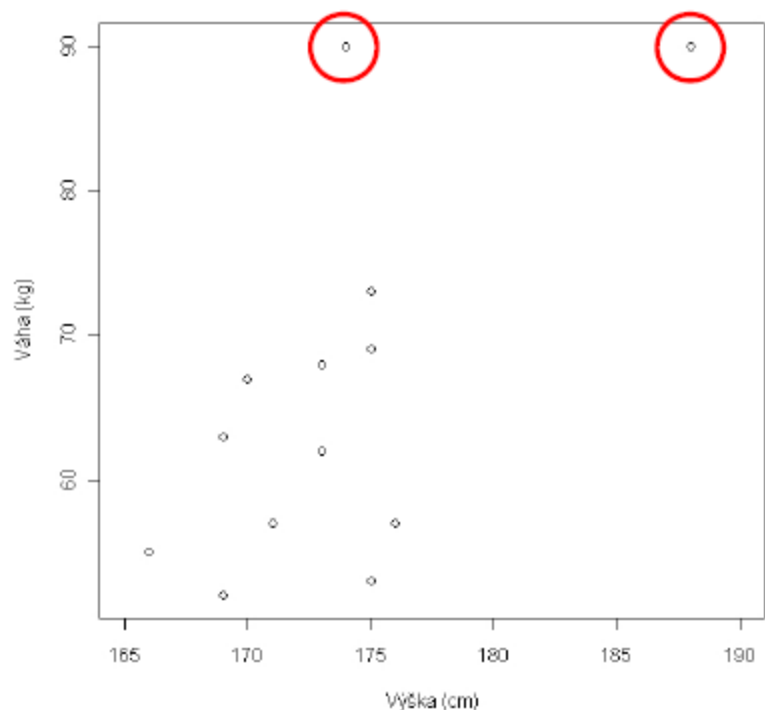
$$r = \frac{721,5 - 637}{(13-1) * 3,86 * 3,88} = 0,47$$

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

Jak to, že nám r a r_s vyšly různě?

🌱 Původní hodnoty: $r = 0,64$

🌱 Pořadí: $r = 0,47$
 $r_s = 0,48$



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.

