FSS MUNI, katedra SPSP
Kvantitativní výzkum x118
Téma 11: Korelace

Miroslav Suchanec

Korelace - souvislost

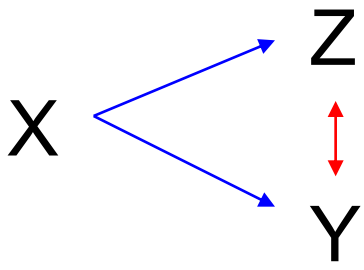
- vyjádření vzájemné souvislosti dvou nebo více jevů (nejméně ordinálního charakteru)
  - kondice a klidová tepová frekvence
  - tělesná výška a tělesná hmotnost
  - úroveň silových schopností a výsledky běhu na 100m
  - výška a rozpětí paží
- hledání skrytých souvislostí, vazeb, příčin, ...

Příčiny souvislosti dvou jevů

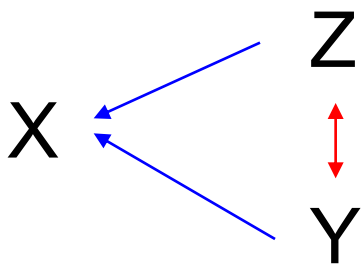
- přirozená sdružená proměnlivost
  - barva očí a vlasů
- teoretické zdůvodnění
  - tělesná hmotnost a výška: konstantní tělesná skladba ρ, objem je svázán s výškou, hmotnost = ρ * V
- jev X vyvolává jev Y
  - kouření a rakovina plic

Příčiny souvislosti dvou jevů

- jevy spolu souvisejí, protože mají společnou příčinu

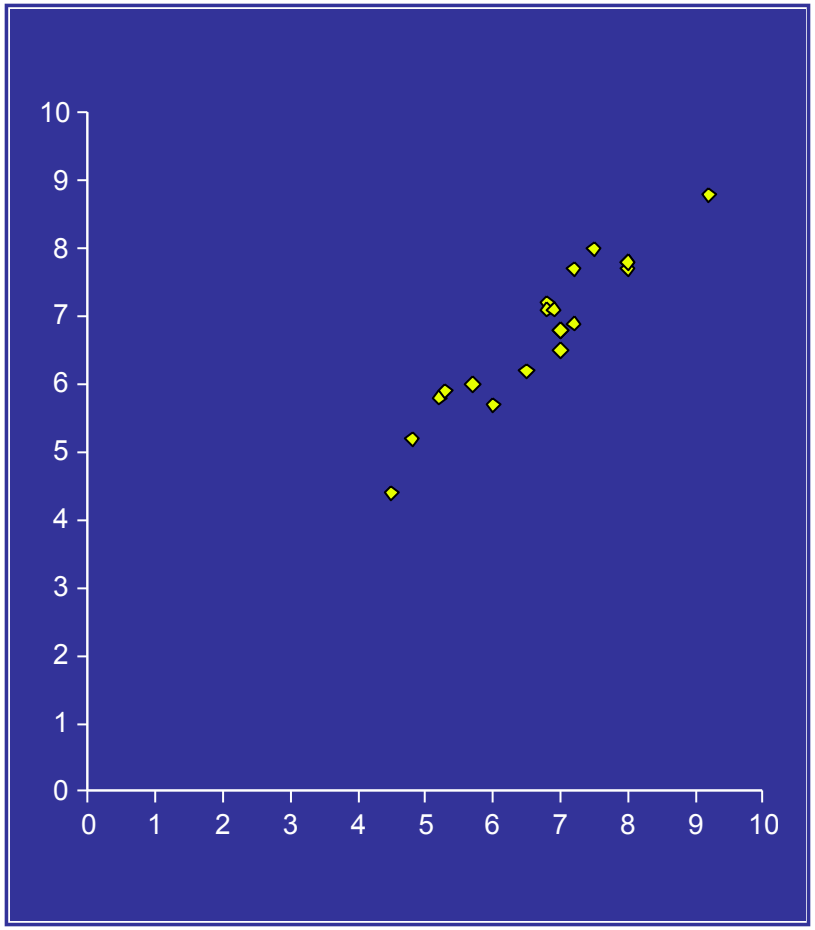


- jevy spolu souvisejí, protože měří něco společného

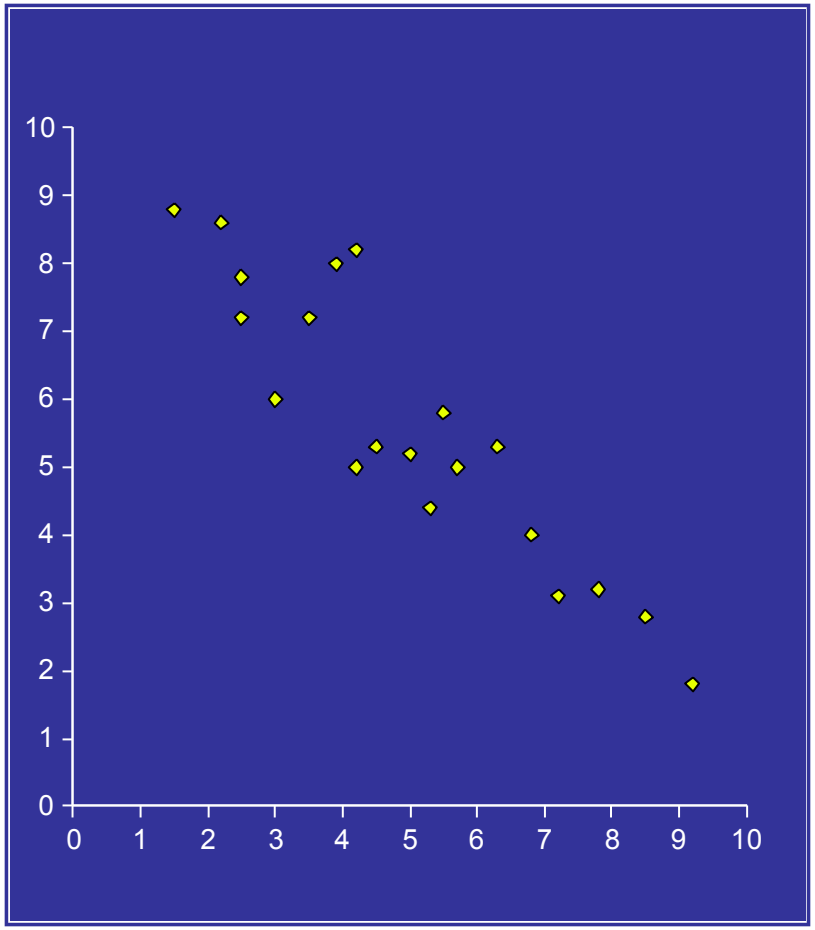


Typy souvislostí

pozitivní



negativní

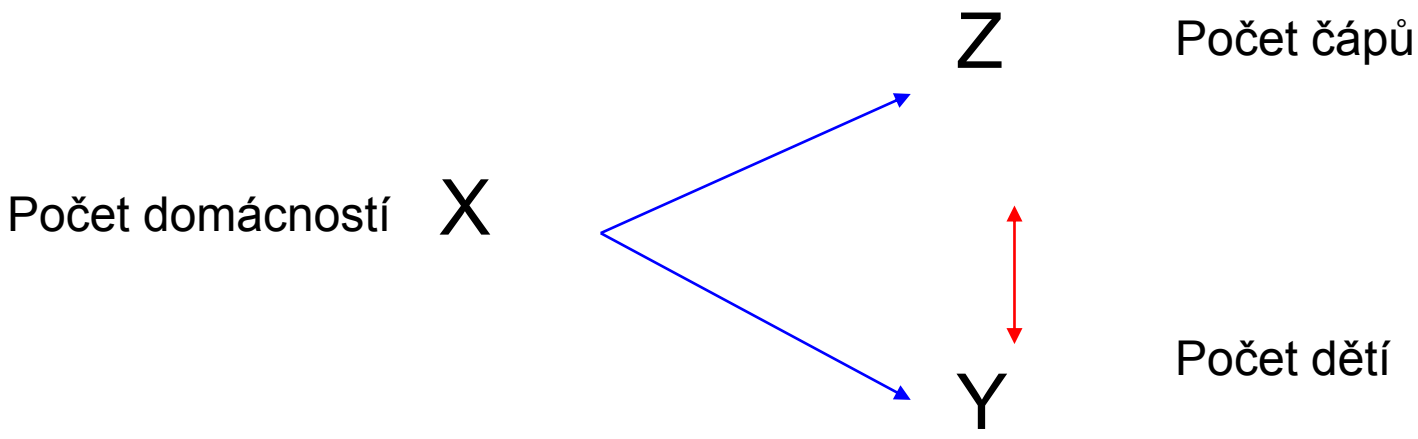


Korelace vs. kauzalita

- Korelace není kauzalita, pouze jedním z jejich předpokladů (dále časová následnost, neexistence alternativní příčiny)

Nepravá korelace (spuriousness)

Znamená vysoká korelace mezi počtem čápů a počtem dětí, že čápi nosí děti?



Velikost souvislosti

- koeficient korelace

$$r \in \langle -1 ; 1 \rangle$$

- mezní hodnoty -1 a 1 značí absolutní souvislost
- hodnota 0 značí absolutní nezávislost

Vyjádření souvislosti

- různé druhy korelačních koeficientů
- použití se liší podle druhu dat, typu závislosti a typu rozložení
- nejčastěji používané:
  - Pearsonův koeficient součinové korelace
  - Spearmanův koeficient pořadové korelace

Pearsonův korelační koeficient (R)

= nástroj pro měření míry lineární souvislosti (vztahu) mezi dvěma intervalovými nebo poměrovými proměnnými (které tudíž mají svůj průměr, rozptyl a odchylku) (Pro ostatní typy proměnných se užívá jiný nástroj pro měření vztahu) *např. souvisí četnost sledování televizních reklam (v min. týdně) s průměrnou týdenní útratou?*

- pohybuje se v rozmezí od -1 do 1 *(-1 =< R xy =< 1),*

  kdy R xy=1 je absolutní pozitivní souvislost (*např. čím více počtu let vzdělání, tím vyšší mzda);*

  R xy=0 je absolutní nezávislost, (např. víme-li člověkovu výšku postavy, nepomůže nám to určit jeho hodnotu IQ);

  a R xy=-1 je absolutní negativní souvislost *(např. čím více mailů sekretářka dostává, tím kratší dobu jí trvá než na ně odpoví).*

Předpoklady použití Pearsonova korelačního koeficientu

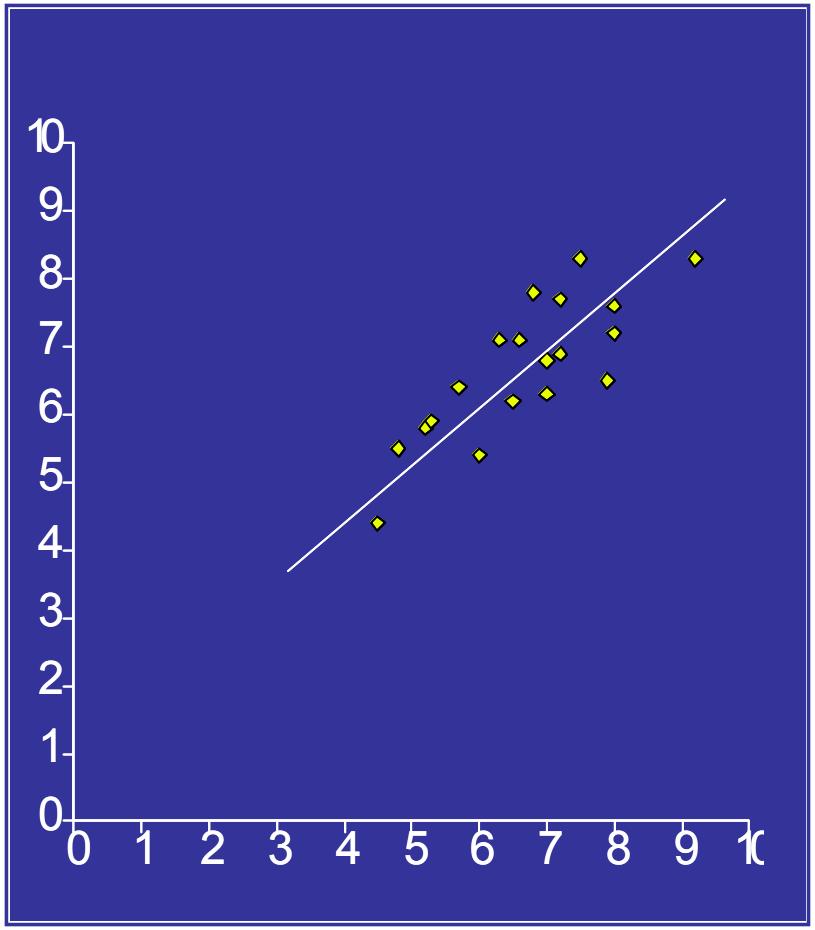
- 1) nejméně intervalová data
- 2) normální rozložení v populaci
- 3) neexistence extrémních případů
- 4) linearita vztahu

2, 3 a 4 třeba ověřit / otestovat

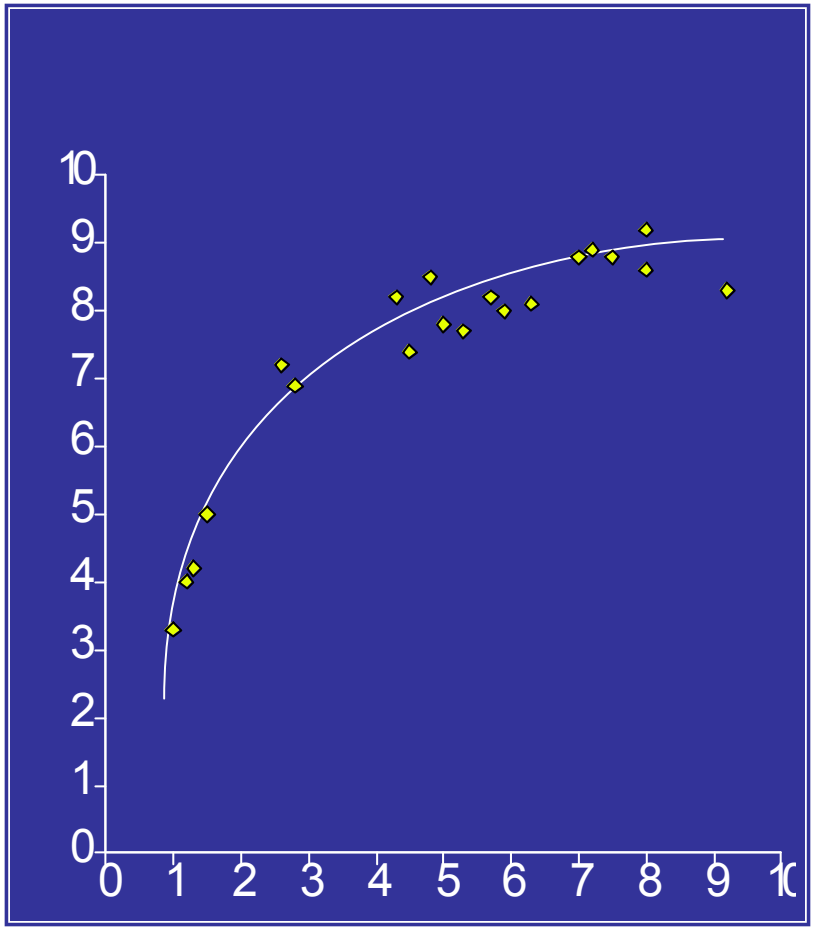
Není-li jeden z předpokladů naplněn a máme-li alespoň ordinální data, používáme Spearmanův koeficient

Předpoklad linearity vztahu

lineární

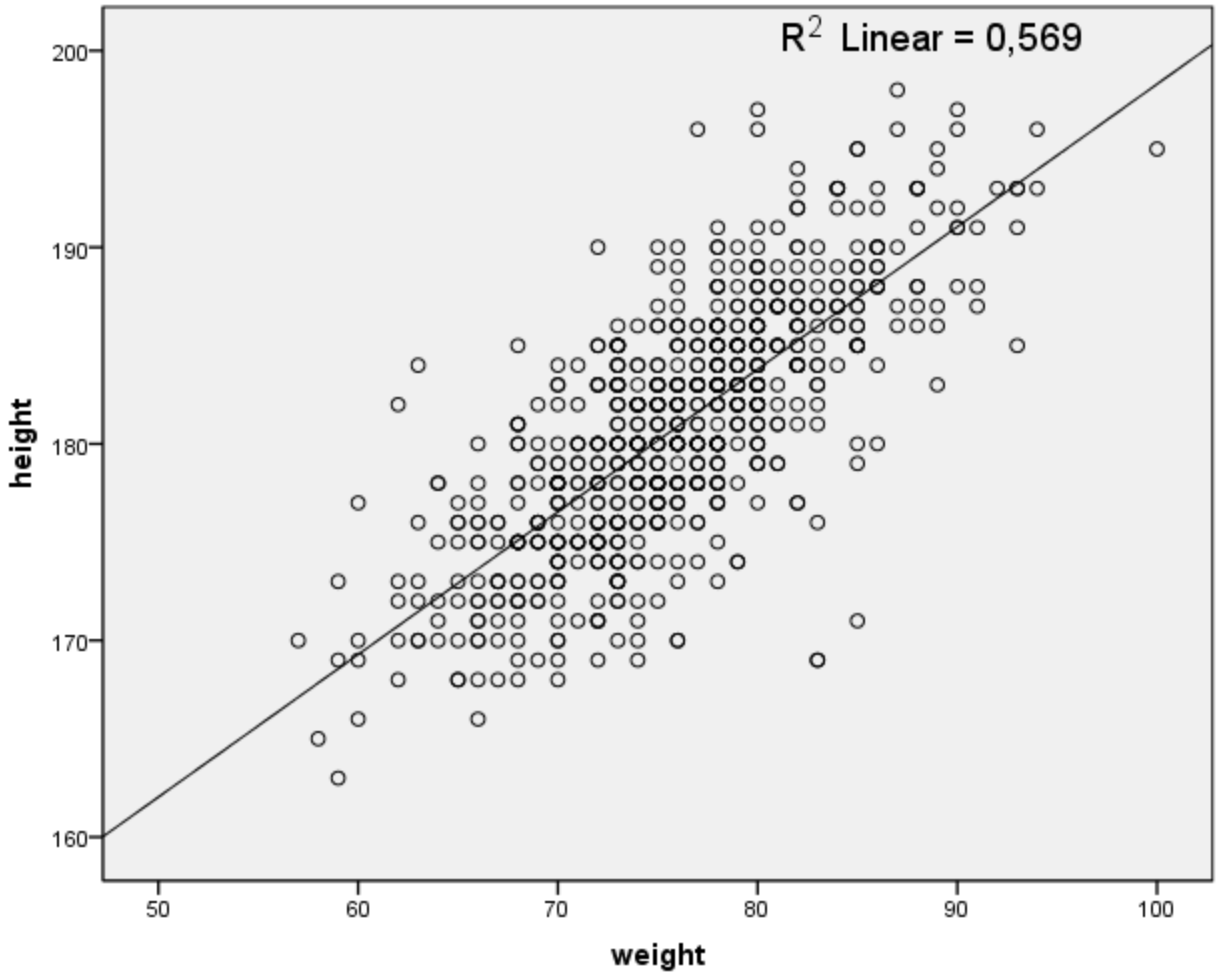


nelineární



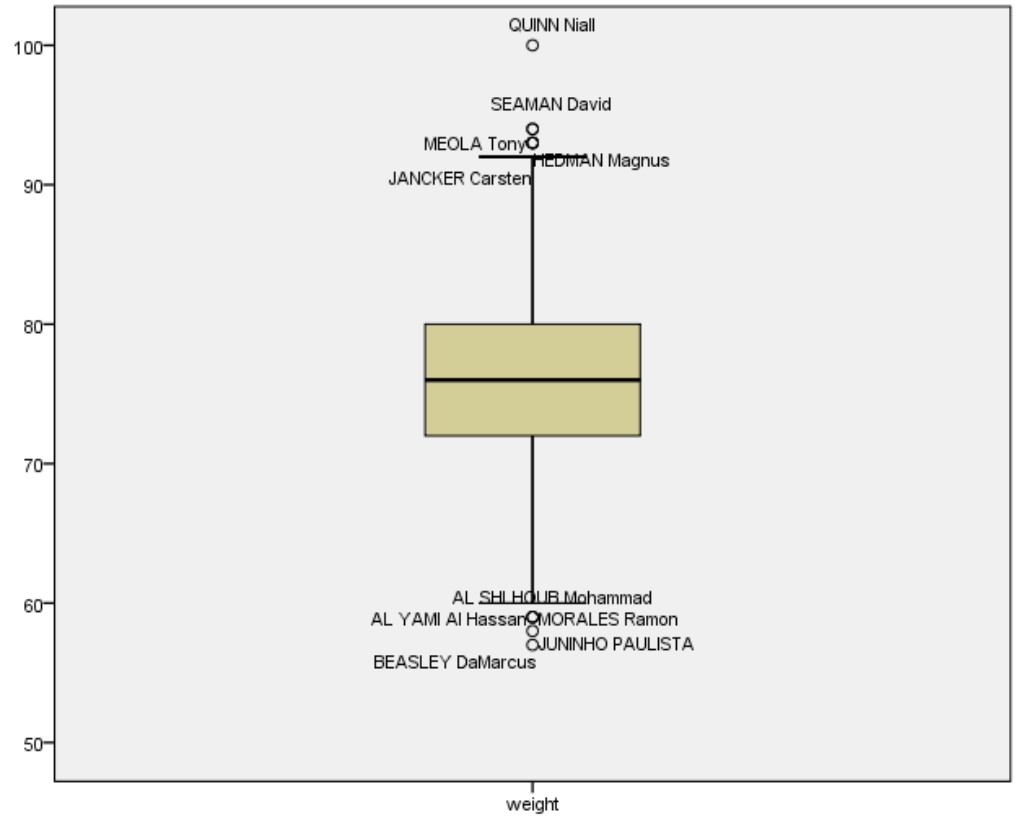
Ověřování předpokladu linearity vztahu

- Nejlépe pomocí bodového rozptýlení (scatterplot)



Ověřování předpokladu neexistence extrémních hodnot

Např. pomocí krabicového diagramu (boxplot) nebo jiného zobrazení extrémních hodnot…

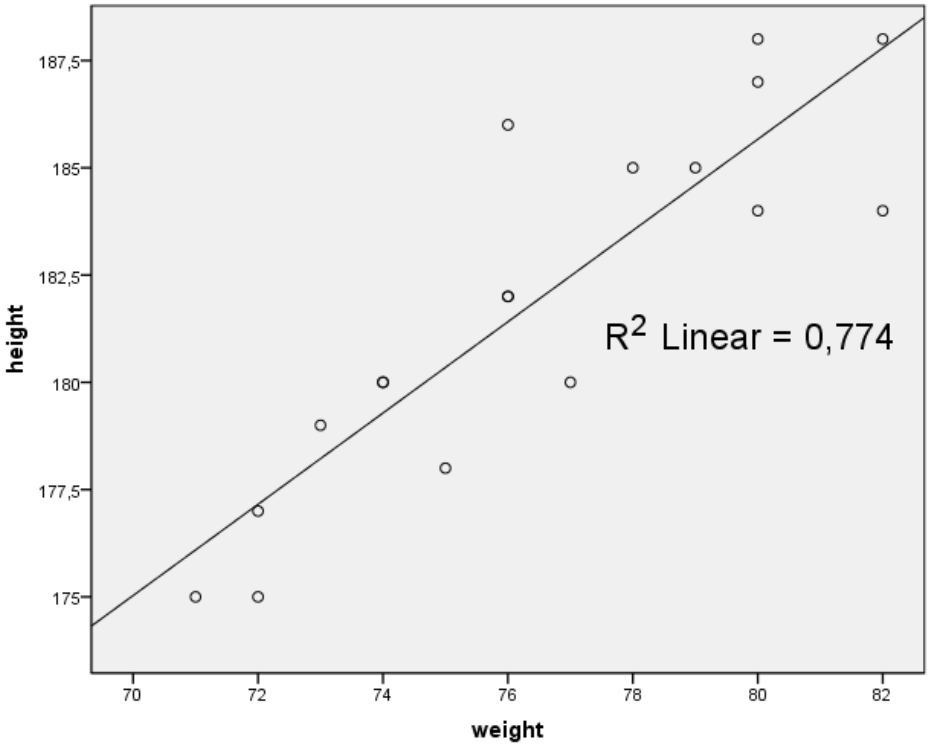


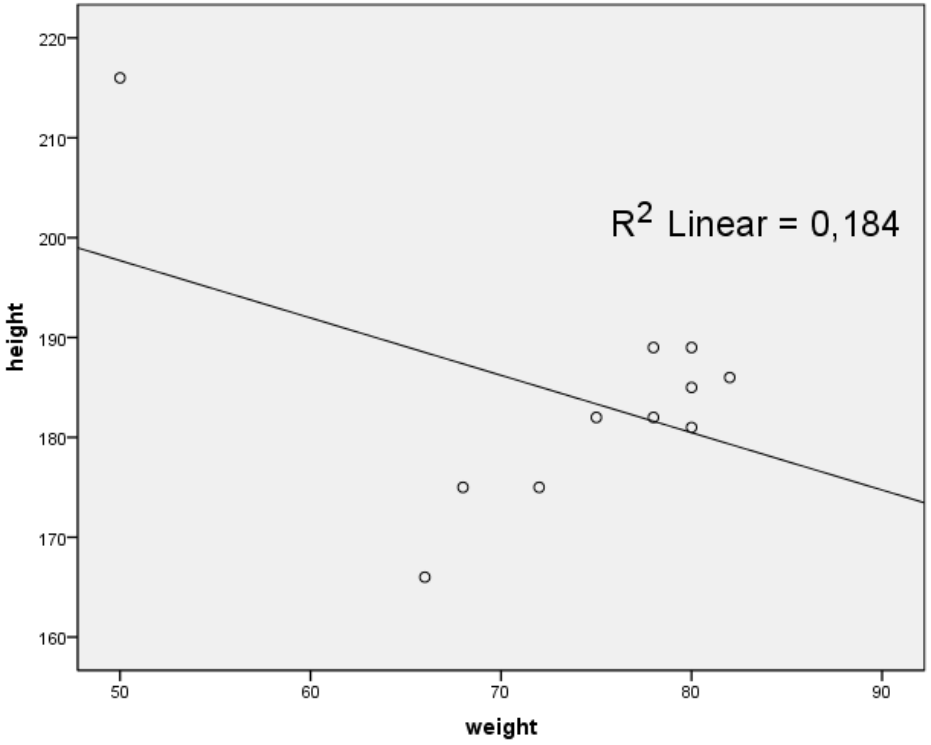
Extreme Values

			Case Number	name	Value
weight	Highest	1	316	QUINN Niall	100
		2	229	JAMES David	94
		3	610	DABANOVIC Mladen	94
		4	208	SEAMAN David	93
		5	285	JANCKER Carsten	93[a]
	Lowest	1	730	BEASLEY DaMarcus	57
		2	65	JUNINHO PAULISTA	58
		3	421	MORALES Ramon	59
		4	411	AL YAMI Al Hassan	59
		5	401	AL SHLHOUB Mohammad	59

a. Only a partial list of cases with the value 93 are shown in the table of upper extremes.

Jak nenaplnění předpokladu neexistence extrémních hodnot ovlivní Pearsonův r?

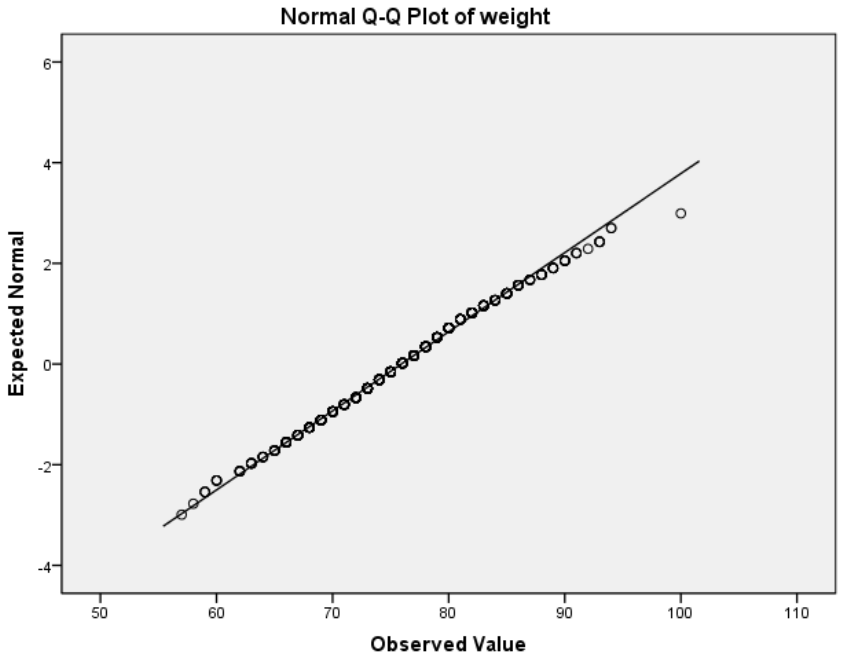




Ověřování předpokladu normality

a) Graficky – pozorované hodnoty ve vzorku vs. očekávané hodnoty pokud je populace normálně rozložená

b) Kolmogorov-Smirnov test normality rozložení



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
weight	,060	723	,000	,994	723	,009

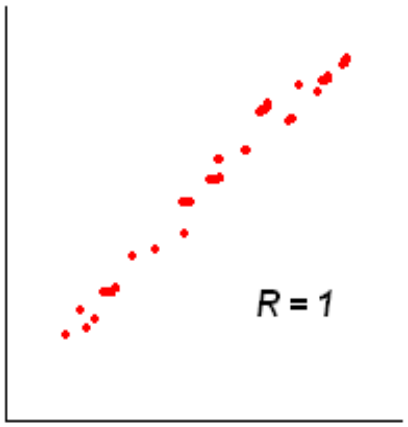
a. Lilliefors Significance Correction

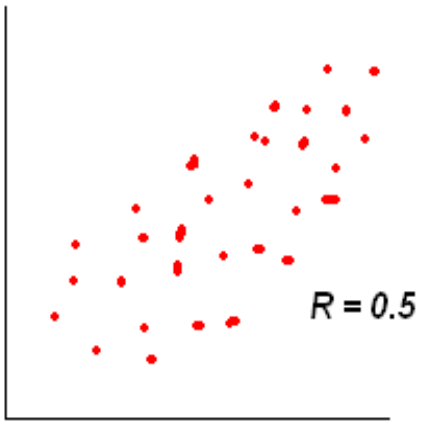
Hypothesis Test Summary

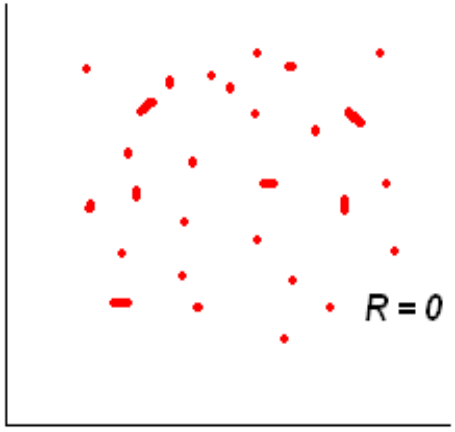
	Null Hypothesis	Test	Sig.	Decision
1	The distribution of weight is normal with mean 75.918 and standard deviation 6.364.	One-Sample Kolmogorov-Smirnov Test	.011	Reject the null hypothesis.

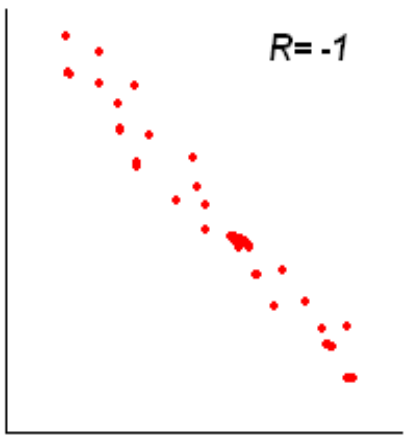
Asymptotic significances are displayed. The significance level is .05.

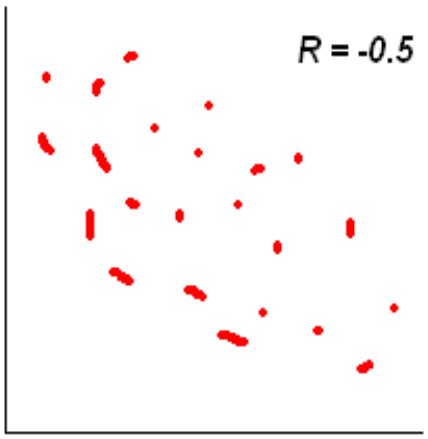
Korelační koeficient a bodové rozptýlení proměnných x a y

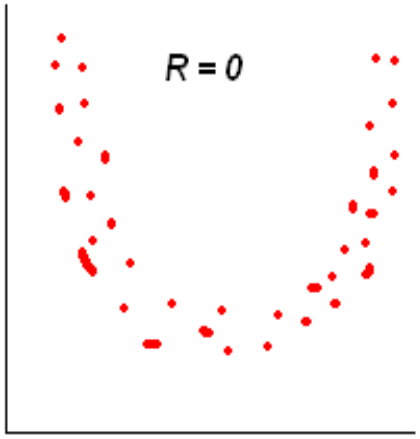




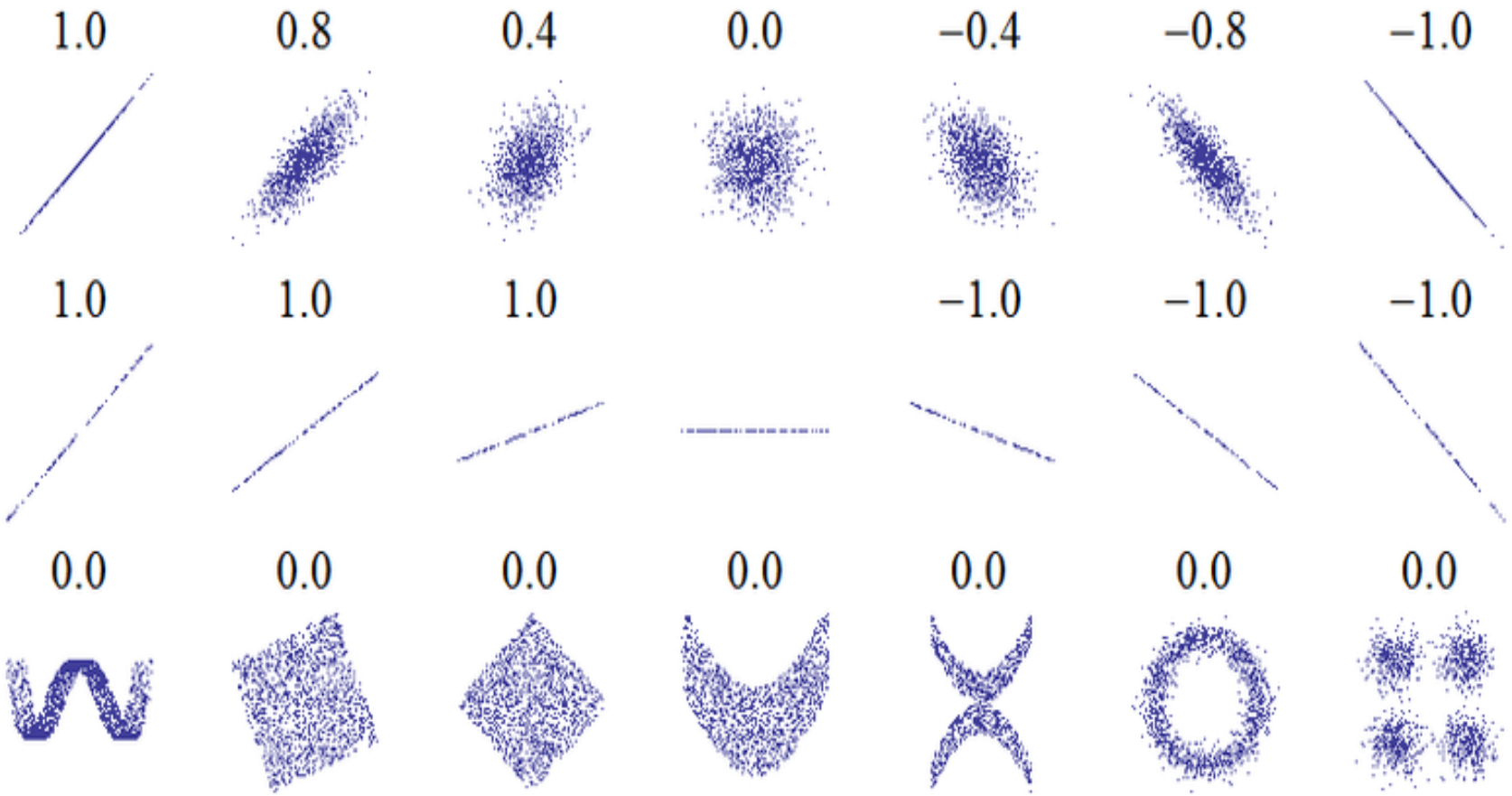








…další příklady



(zdroj: wikipedia)

x	y
2	0
2	2
3	1
3	3
4	2
4	4
5	3
5	5
6	4
6	6

korelace mezi x a y, neboli R xy = cov(x,y) / s(x) * s(y),

(Databáze korelace a regrese.sav)

$$Cov\ (x,y) = \Sigma\ dx * dy / n - 1 = \Sigma\ (xi - \bar{x})*(yi - \bar{y}) / n - 1$$

$$\bar{X} = \Sigma\ x_i / n = 40 / 10 = 4$$

$$\bar{Y} = \Sigma\ y_i / n = 30 / 10 = 3$$

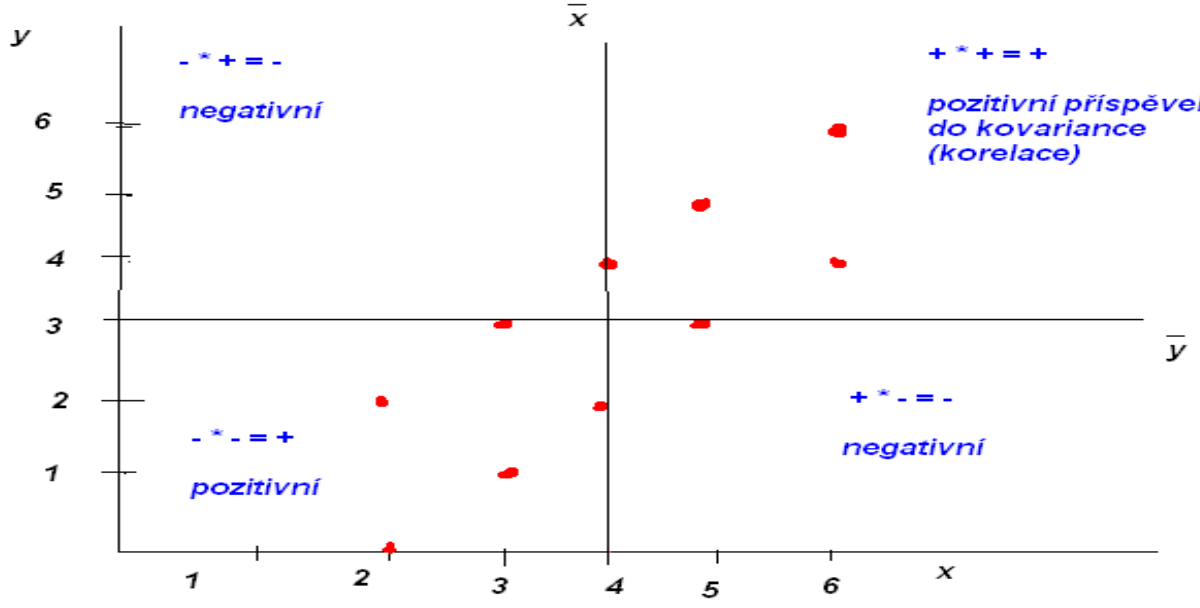
$$Cov\ (x,y) = \Sigma\ dx * dy / n - 1 = \Sigma\ (xi - \bar{x})*(yi - \bar{y}) / n - 1 = (-2*-3) + (-2*-1) + (-1*-2) + 0 + 0 + 0 + 0 + (1*2) + (2*1) + (2*3) = 20 / 9 = 2,22$$

$$s(x) * s(y) = \sqrt{}\ var\ (x) * \sqrt{}\ var\ (y)$$

$$var\ (x) = \Sigma\ (x_i - \bar{x})^2 / n - 1 = 20 / 9 = 2,22$$

$$var\ (y) = \Sigma\ (y_i - \bar{y})^2 / n - 1 = 30 / 9 = 3,33$$

$$R\ xy = cov(x,y) / s(x)\ s(y) = 2,22 / \sqrt{}\ 2,22 * \sqrt{}\ 3,33 = 2,22 / 2,72 = 0,81$$



Legenda

Xi = hodnota X pro jednotlivá individua

X = průměr pro x

d = absolutní odchylka

var(x)=rozptyl x

s(x)=směrodatná odchylka

Cov (x,y)=kovariance mezi x a y

R (x,y)= korelace mezi x a y

Korelace výpočet s příkladem

Rozdíl mezi Pearsonovým a Spearmanovým koeficientem