

# Korelace

Peter Spáč

3.11.2016

# Korelace

- Vzájemná souvislost mezi proměnnými
- Nárůst hodnot jedné proměnné je spojený s nárůstem / poklesem hodnot druhé proměnné
- Korelace neimplikuje kauzalitu

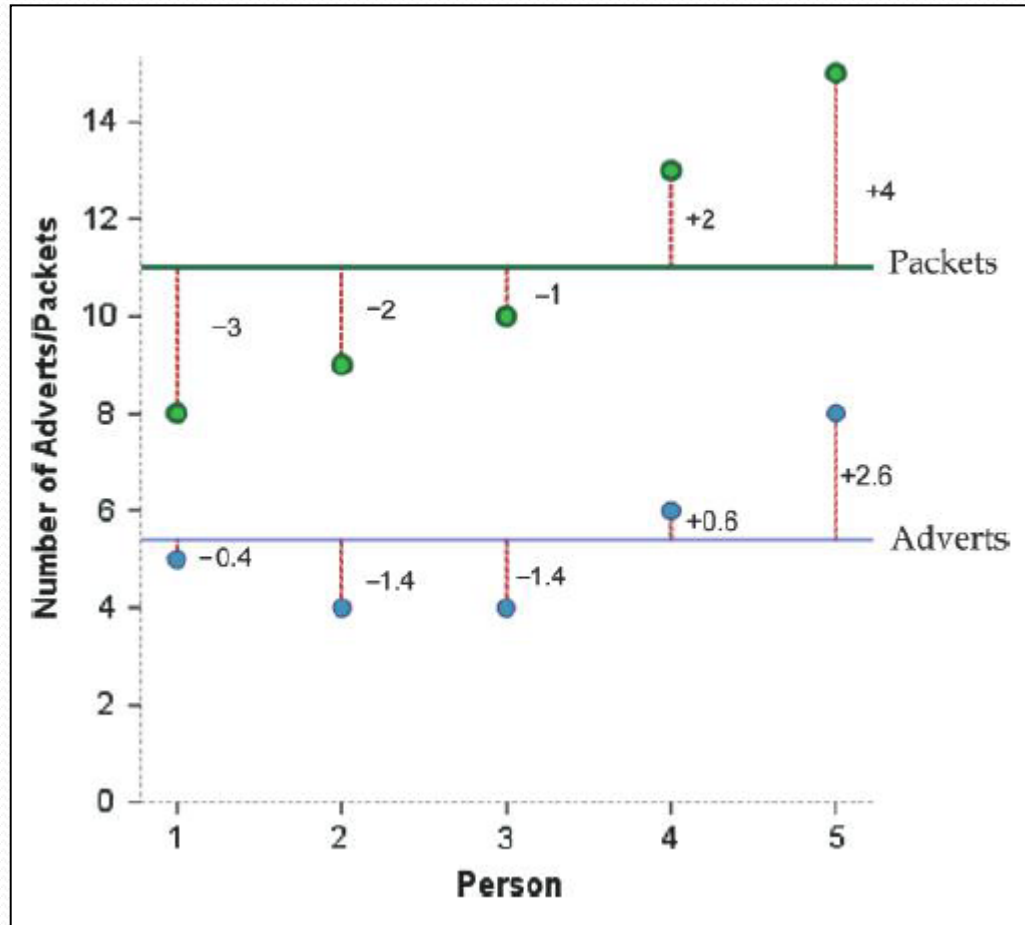
# Kovariance

- Nejjednodušší posouzení vzájemné souvislosti dvou proměnných
- Souvislost - změna v hodnotách jedné proměnné bude spojena s obdobnou změnou ve druhé proměnné
- Podobné odklony od průměru v obou proměnných

# Kovariance (Field 2009: 168)

Subject	1	2	3	4	5	Mean	S
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92

# Kovariance (Field 2009: 168)



# Výpočet

- Rozptyl (variance)
  - Suma umocněných odchylek od průměru vydělená počtem případů - 1

$$= \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{N - 1}$$

- Kovariance (covariance)
  - Totožný výpočet, do kterého se zakomponuje druhá proměnná

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Osoba	Reklamy (x)	Průměr	Rozdíl	Nákup (y)	Průměr	Rozdíl
1	5	5,4	-0,4	8	11	-3
2	4		-1,4	9		-2
3	4		-1,4	10		-1
4	6		0,6	13		2
5	8		2,6	15		4

$$\begin{aligned}
&= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1} \\
&= \frac{(-0.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (0.6)(2) + (2.6)(4)}{4} \\
&= \frac{1.2 + 2.8 + 1.4 + 1.2 + 10.4}{4} \\
&= \frac{17}{4} \\
&= 4.25
\end{aligned}$$

# Kovariance

- Ukazuje základní souvislost mezi proměnnými
- Je možné identifikovat kladní nebo záporní vztah
- Nevýhoda – nemožnost vzájemných srovnání
- Potřeba standardizace



# Kovariance

- Standardizace pro účely názornosti i srovnatelnosti (není možné spoléhat, že všechna měření budou v stejných jednotkách)
- Hodnota kovariance se vydělí součinem obou směrodatných odchylek
- Výsledkem je standardizovaná hodnota (vyjádřena v směrodatných odchylkách)
- **Pearsonův korelační koeficient**

# Pearsonův korelační koeficient

- Jeden ze základních korelačních koeficientů
- Značení -  $R$  (při populaci),  $r$  (při vzorce)
- Hodnoty koeficientu:
  - Rozsah od  $-1$  po  $1$
  - $+1$  = perfektní kladná souvislost
  - $-1$  = perfektní záporná souvislost
  - $0$  = žádná souvislost
- Čím více je hodnota vzdálena od nuly, tím je souvislost silnější

# Pearsonův korelační koeficient

- Síla vztahu:
  - $\pm 0,1$  – slabý
  - $\pm 0,3$  – střední
  - $\pm 0,5$  – silný
- Spíše arbitrabilní hodnoty (mezi  $r = 0,29$  a  $r = 0,31$  žádný zásadný rozdíl není)

# Druhy korelace

- Bivariační – souvislost mezi dvěma proměnnými
- Parciální (partial) – souvislost mezi dvěma proměnnými za jisté kontroly vlivu jiných proměnných

# Bivariační korelace

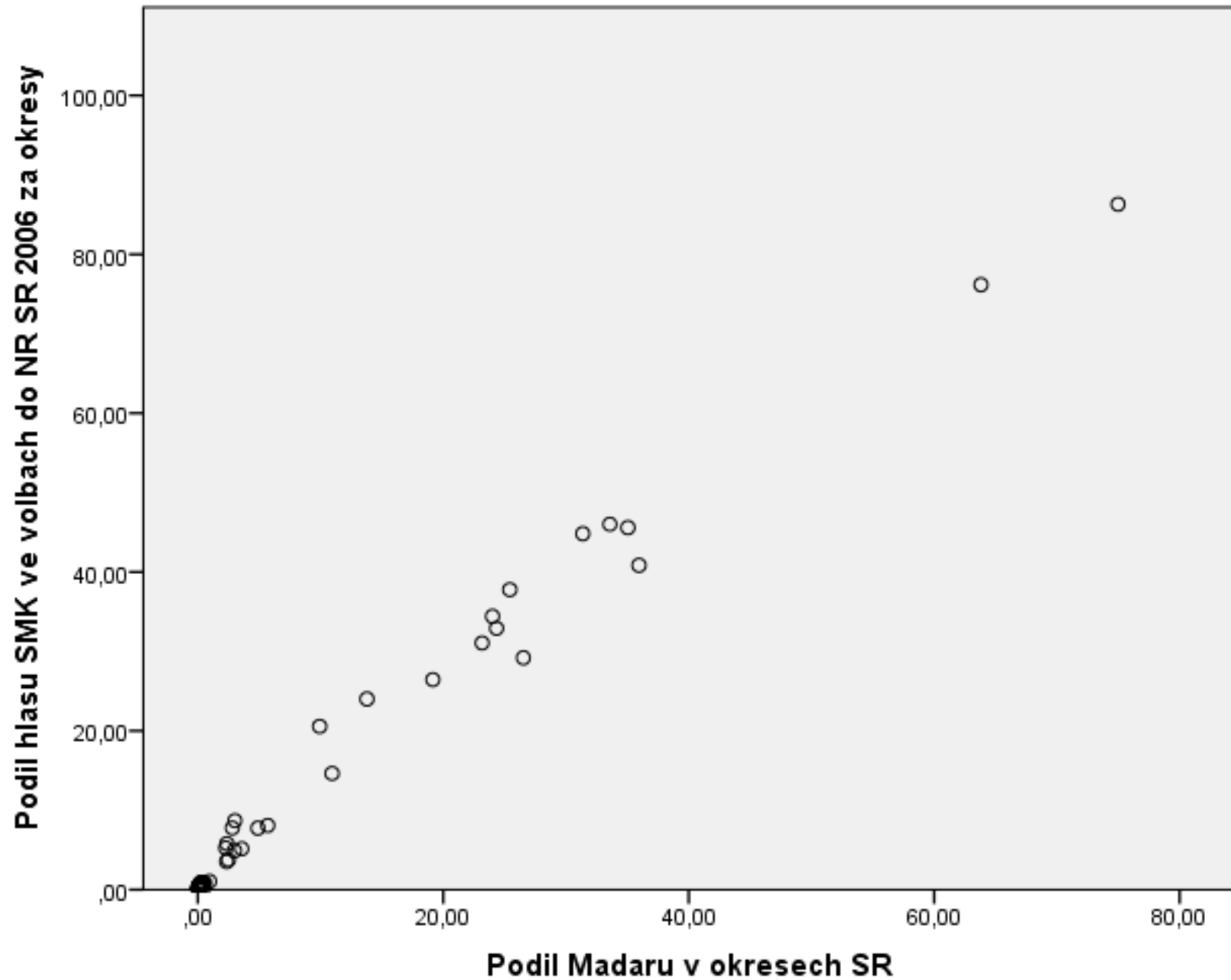
- „Jednodušší“ forma
- Posuzuje souvislost mezi dvěma proměnnými bez dalšího
- Tři základní postupy:
  - Pearsonův korelační koeficient
  - Spearmanovo rho
  - Kendalovo tau

# Pearsonův korelační koeficient

- Předpoklady:
  - Kardinální data (možná výjimka)
  - Pokud zjišťujeme i statistickou signifikanci, tak i normální rozložení (nebo dostatečná velikost vzorku)
- Výjimka – jedna z proměnných může být kategorická (dichotomická)
- Citlivost na odlehlé případy

# Práce v SPSS

- Před analýzou je vhodné si data graficky zobrazit (netýká se pouze Pearsonova korel. koeficientu)
- Bodový graf (scatter/dot)
- Graphs → Chart builder :
  - Zvolit Scatter/Dot
  - Vložit proměnné





# Práce v SPSS

- Analyze → Correlate → Bivariate:
  - Zvolit proměnné
  - Pearsonův koeficient je přednastavený
  - Pro sledování signifikance zvolit *Flag significant correlations*
- Options:
  - Možnost spočítat základní statistiky a kovarianci
  - Vynechání hodnot / případů

# Pearsonův korelační koeficient

		Podíl Madaru v okresech SR	Podíl hlasu SMK ve volbach do NR SR 2006 za okresy
Podíl Madaru v okresech SR	Pearson Correlation	1	,992
	Sig. (2-tailed)		,000
	N	79	79
Podíl hlasu SMK ve volbach do NR SR 2006 za okresy	Pearson Correlation	,992	1
	Sig. (2-tailed)	,000	
	N	79	79

# Pearsonův korelační koeficient

		Time Spent Revising	Exam Performance (%)	Exam Anxiety
Time Spent Revising	Pearson Correlation	1	,397**	-,709**
	Sig. (2-tailed)		,000	,000
	N	103	103	103
Exam Performance (%)	Pearson Correlation	,397**	1	-,441**
	Sig. (2-tailed)	,000		,000
	N	103	103	103
Exam Anxiety	Pearson Correlation	-,709**	-,441**	1
	Sig. (2-tailed)	,000	,000	
	N	103	103	103

\*\* . Correlation is significant at the 0.01 level (2-tailed).

# Pearsonův korelační koeficient

- Se zjištěným  $R$  je možné dál pracovat
- Po umocnění získáváme tzv. Index determinace ( $R^2$ )
- $R^2$  vymezuje, jaký podíl variability jedné proměnné je sdílený s druhou proměnnou
- Pro názornost se  $R^2$  násobí číslem 100 a vyjadřuje v procentech
- Nadále však daná hodnota neříká nic o kauzalitě

# Pearsonův korelační koeficient

- Výjimka z kardinálních dat → korelace jedné kardinální proměnné a jedné dichotomické
- Tzv. point-biserial korelace
- Úplně stejný postup
- Kladní / záporní výsledné hodnoty plně závisí od kódování dichotomické proměnné

# Pearsonův korelační koeficient

## Correlations

		Pocet shlednuti	Pohlavi
Pocet shlednuti	Pearson Correlation	1	,677**
	Sig. (2-tailed)		,000
	N	37	37
Pohlavi	Pearson Correlation	,677**	1
	Sig. (2-tailed)	,000	
	N	37	37

\*\* . Correlation is significant at the 0.01 level (2-tailed).

# Pearsonův korelační koeficient

## Correlations

		Pocet shlednuti	pohl2
Pocet shlednuti	Pearson Correlation	1	-,677 <sup>**</sup>
	Sig. (2-tailed)		,000
	N	37	37
pohl2	Pearson Correlation	-,677 <sup>**</sup>	1
	Sig. (2-tailed)	,000	
	N	37	37

\*\* . Correlation is significant at the 0.01 level (2-tailed).

# Spearmanovo rho

- Neparametrický postup
- Použitelný pro neparametrická data (ordinální, porušení normality apod.)
- Data nejdřív seřadí a následně toto pořadí využívá pro výpočet korelačního koeficientu
- Výsledné hodnoty jsou ve stejném pásmu jako u PKK (od -1 po 1)



# Spearmanovo rho

- Analyze → Correlate → Bivariate:
  - Zvolit proměnné
  - Vybrat *Spearman*
- Vše ostatní je stejné, pouze v *Options* není možnost spočítat statistiky (mají smysl pouze pro Pearsonův korelační koeficient)

# Spearmanovo rho

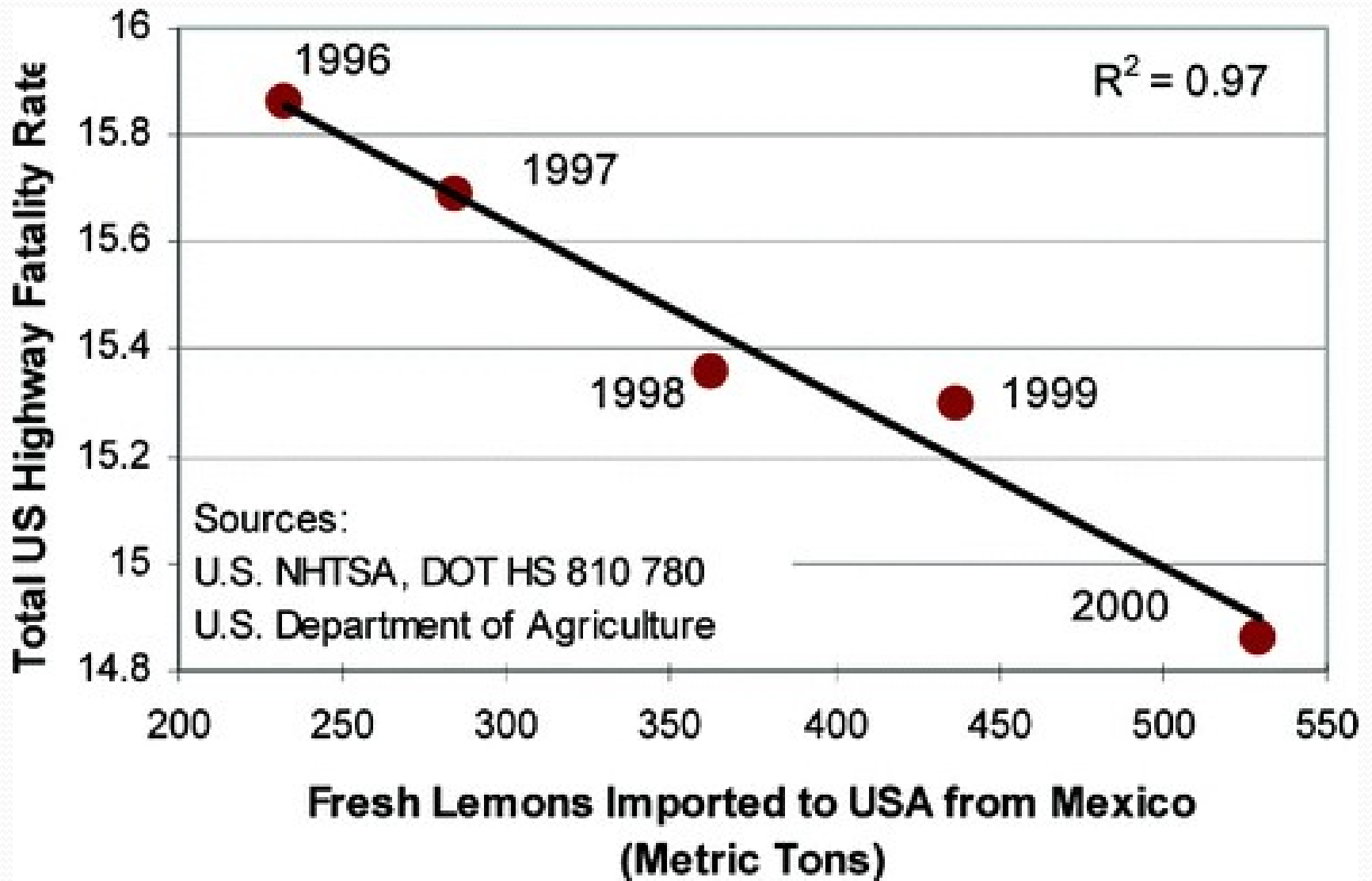
- Podobně jako u PKK, i zde je možné výsledný koeficient umocnit  $\rightarrow R_S^2$
- Interpretace je částečně odlišná – Spearmanovo rho je založené na pořadí  $\rightarrow R_S^2$  vyjadřuje podíl sdílených pořadí mezi proměnnými

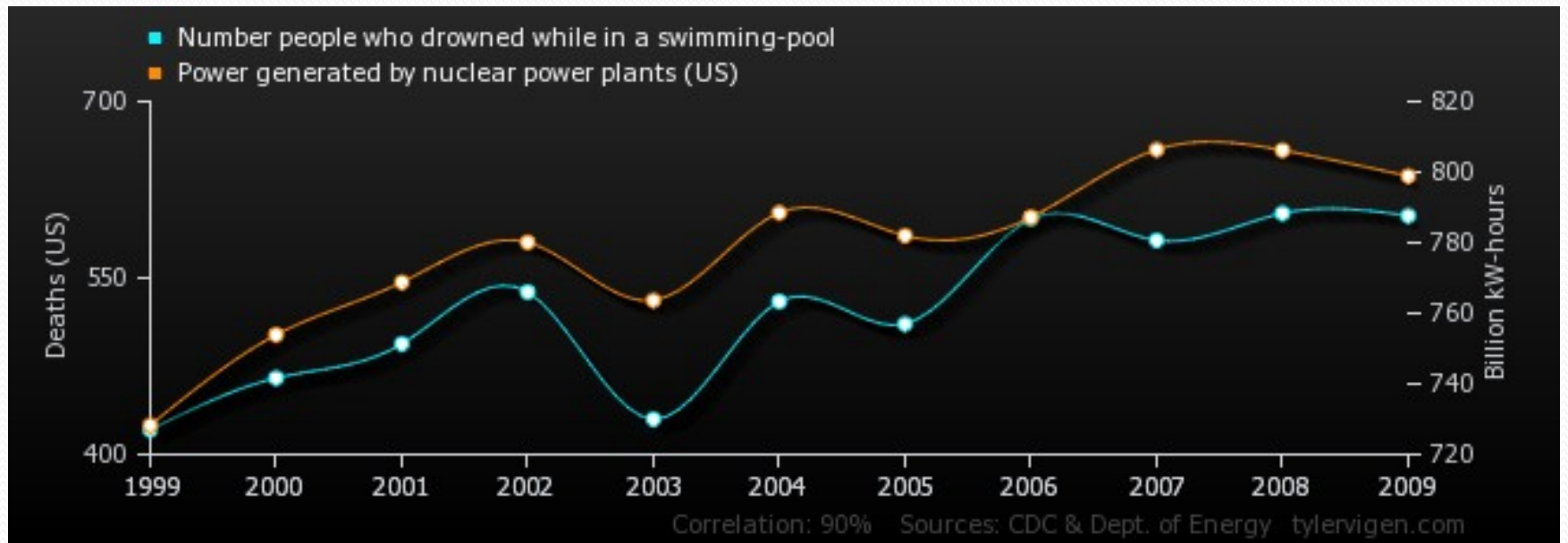
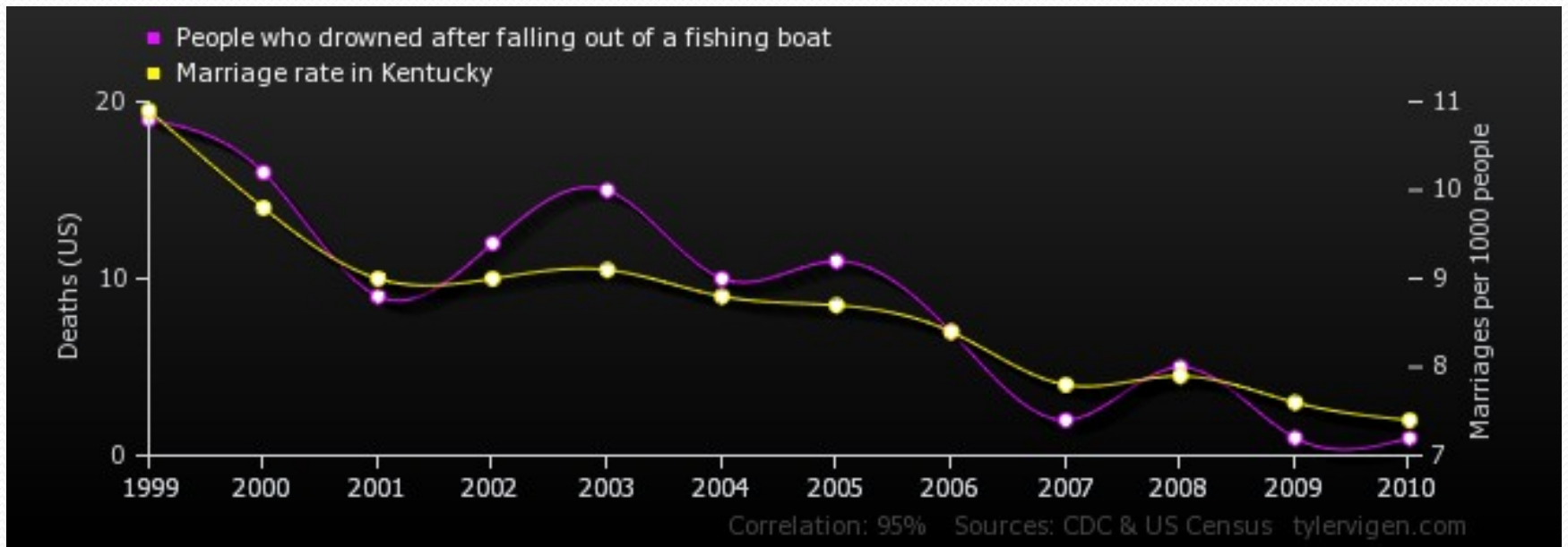
# Kendalovo tau

- Neparametrický postup
- Použitelný jako Spearmanovo rho (totožný postup i v SPSS – pouze se zvolí *Kendall's tau-b* namísto *Spearman*)
- Kdy upřednostnit před Spearmanem:
  - Menší počet dat
  - Mnoho totožných hodnot

# Interpretace výsledků

- Základní pravidlo – **korelace  $\neq$  kauzalita**
- Korelace vyjadřuje pouze souvislost mezi proměnnými, neukazuje na žádnou příčinu a následek
- Vliv třetích proměnných
- Korelace neuvádí směr působení proměnných - ty jsou ve výpočte plně rovnocenné (žádná nezávislá a závislá proměnná)
- Nemožnost konstatovat kauzalitu trvá i pokud se příčinný vztah jeví jako „logický“ – korelace nemá potenciál ani nástroj to odhalit
- Statistické zjištění nemá automaticky věcný význam

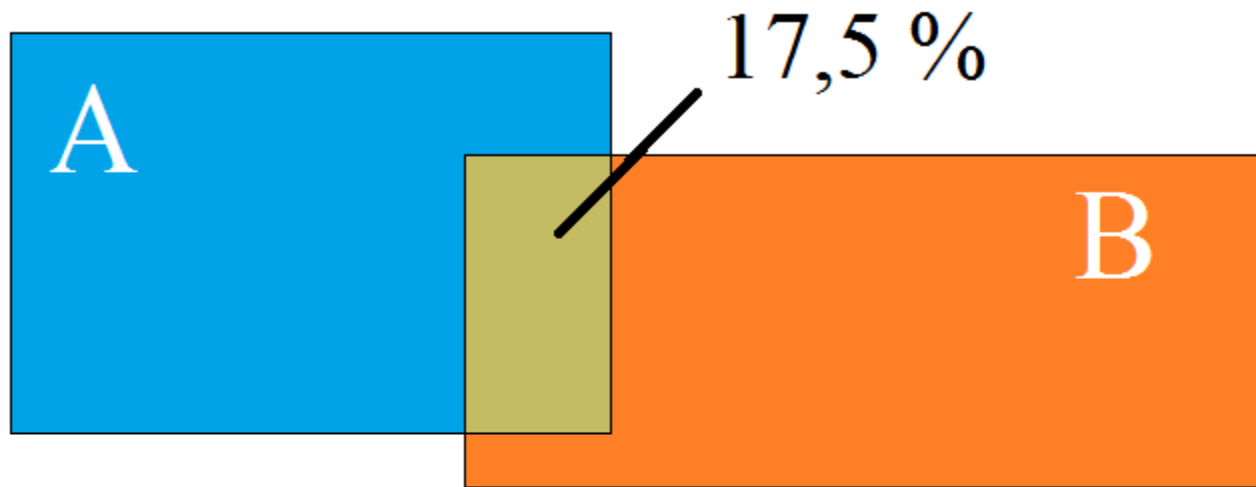




# Parciální (partial) korelace

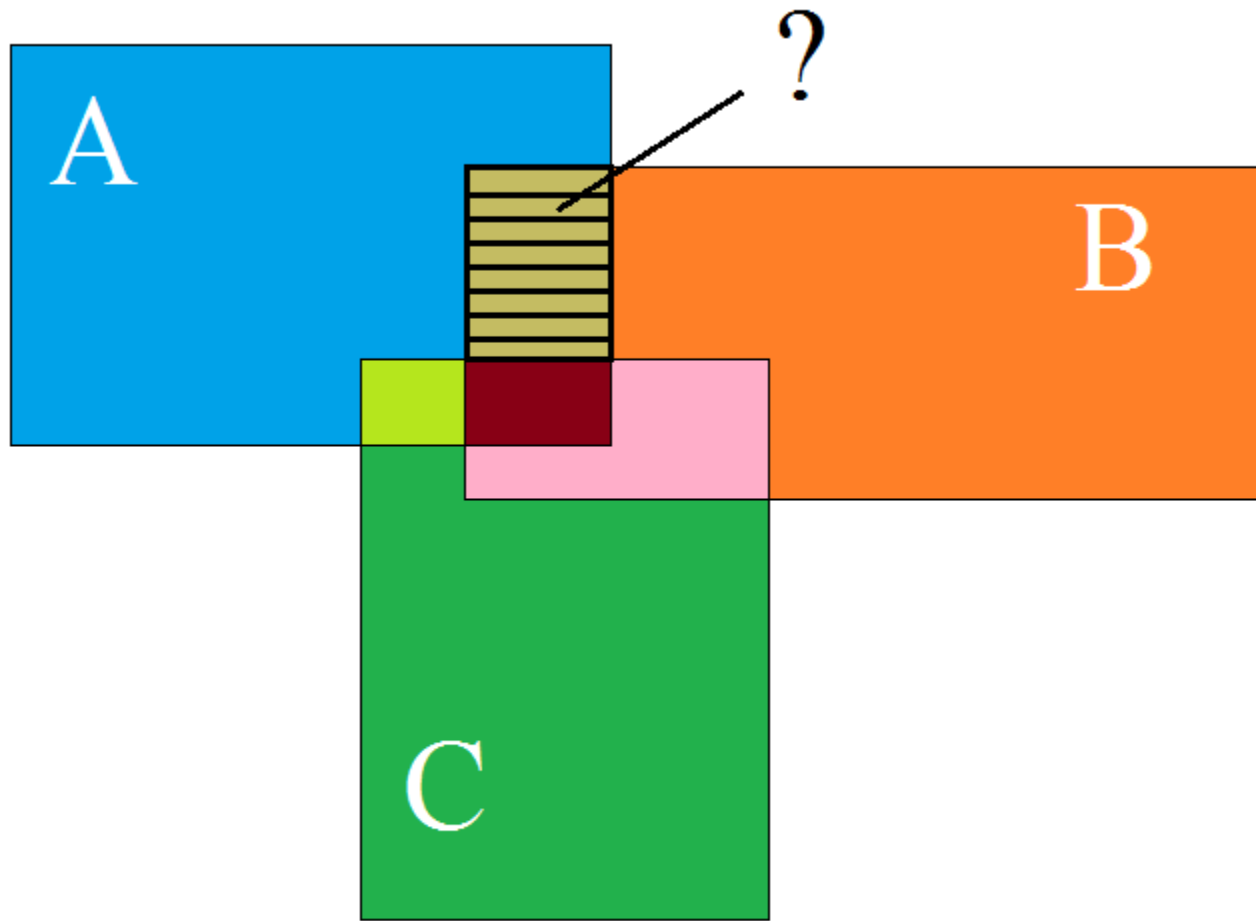
- Souvislost mezi dvěma proměnnými za jisté kontroly vlivu jiných proměnných (třetí proměnná je konstantní)
- „Očištění“ souvislosti od jiných proměnných
- Snaha o identifikaci „čistého“ podílu sdílené variability pouze mezi dvěma proměnnými

# Parciální (partial) korelace





# Parciální (partial) korelace



# Parciální korelace

- Analyze → Correlate → Partial:
  - Korelované proměnné do *Variables*
  - Kontrolní proměnné do *Controlling for*
  - Pro sledování numericky vyjádřené signifikance zvolit *Display actual significance level*
- Options:
  - Možnost spočítat základní statistiky a bivariační korelace
  - Vynechání hodnot / případů

# Parciální korelace

Control Variables			Exam Performance (%)	Exam Anxiety	Time spent Revising
-none <sup>a</sup>	Exam Performance (%)	Correlation	1,000	-,441	,397
		Significance (2-tailed)	.	,000	,000
		df	0	101	101
	Exam Anxiety	Correlation	-,441	1,000	-,709
		Significance (2-tailed)	,000	.	,000
		df	101	0	101
	Time Spent Revising	Correlation	,397	-,709	1,000
		Significance (2-tailed)	,000	,000	.
		df	101	101	0
Time Spent Revising	Exam Performance (%)	Correlation	1,000	-,247	
		Significance (2-tailed)	.	,012	
		df	0	100	
	Exam Anxiety	Correlation	-,247	1,000	
		Significance (2-tailed)	,012	.	
		df	100	0	

a. Cells contain zero-order (Pearson) correlations.

# Práce s koeficienty

- $R^2$  (Pearson) a  $R_s^2$  (Spearman) je možné srovnávat, zvláště pokud se distribuce hodnot blíží normální
- Kendallovo tau se svou hodnotou neblíží ani Pearsonovmu  $R$ , ani Spearmanovmu  $\rho$  (je o 66-75 % nižší)

# Práce s koeficienty

- Na místě je opatrná interpretace
- Nikdy nepoužívat obraty typu „korelační koeficient ukázal vliv proměnné A na proměnnou B...“
- Co uvádět:
  - Korelační koeficient (pozor na odlišné značení P, S a K koeficientů)
  - Signifikantnost (pokud má smysl) a její hladinu