PSY117/454
Statistická analýza dat v psychologii
Přednáška 5

ZOBRAZENÍ DVOUROZMĚRNÝCH DAT
KORELAČNÍ KOEFICIENT

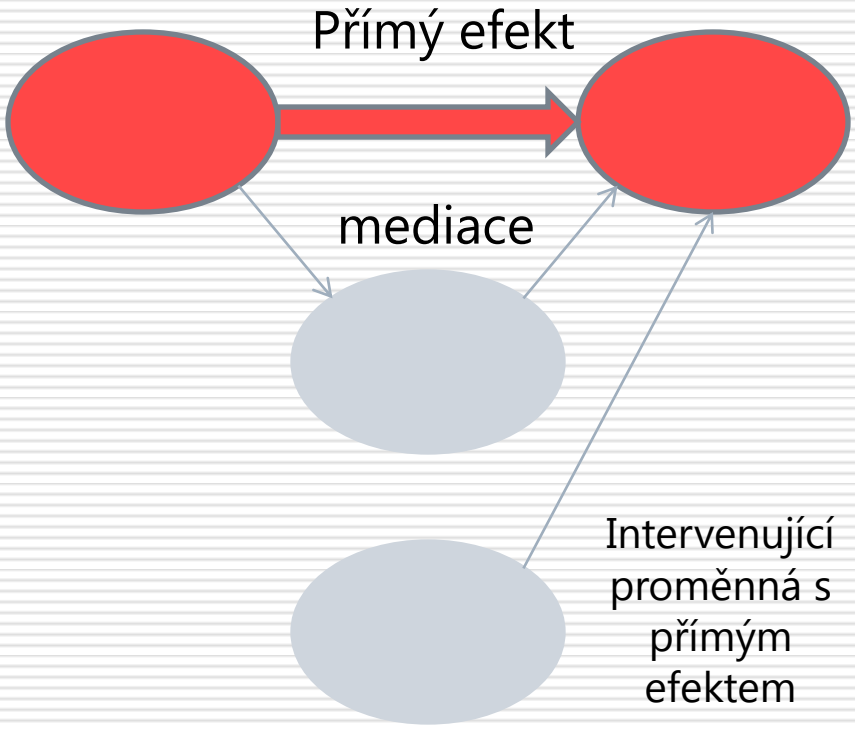
Všichni žijeme v matrixu.

Výzkumné otázky…

- Hypotézy o vzájemné souvislosti jevů:
  - Predikuje intelekt akademický úspěch?
  - Mají dobří češtináři i dobré známky z matematiky?
  - Existuje souvislost mezi mírou depresivní a anxiózní symptomatiky?
  - Jsou měsíční příjem a délka pracovní doby dobrými prediktory životní spokojenosti?
  - Jsou různá umělecká nadání specifická, nebo vycházejí ze stejného „všeobecného“ talentu?

Klasifikace proměnných z hlediska funkce v problému

- Cílem výzkumu je obvykle prověřovat kauzální vztahy
  - ...na úrovni humanitních věd velmi ambiciózní ☺
  - Statistická analýza nemá potenciál ke zjištění nebo testování kauzality. To je úlohou designu výzkumu a teoretického zpracování.
  - Špatně sebraná data (nevhodný design) nelze zachránit sebelepší analýzou.
- Klasifikace proměnných:
  - Závislé, nezávislé, intervenující
  - Exogenní, endogenní, moderátory, mediátory
  - Obvykle není možné identifikovat všechny intervenující proměnné...



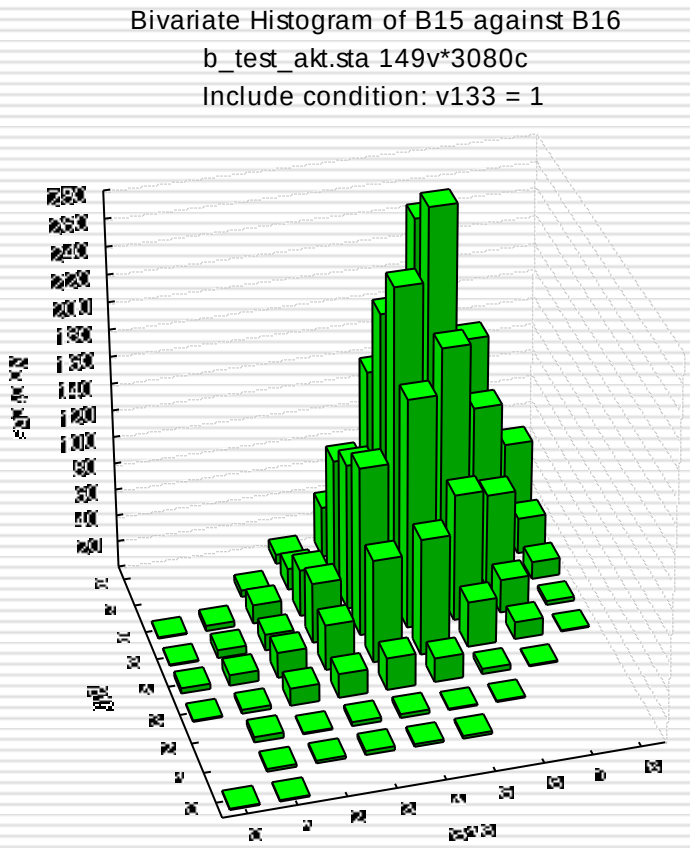
Kontingenční tabulka

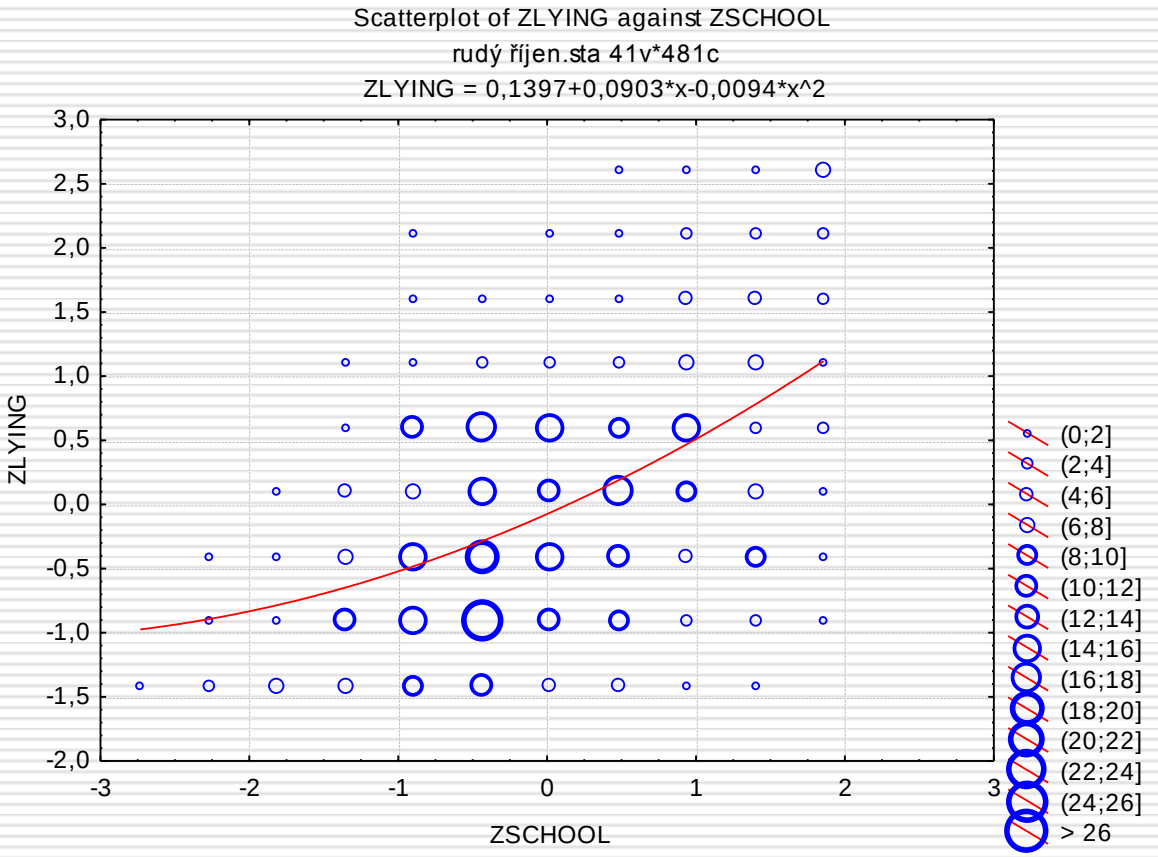
		známka z matematiky					celkem
		1	2	3	4	5	
známka z čj	1	82	40	8	1	0	131
	2	71	200	73	17	0	361
	3	4	75	109	25	0	213
	4	1	7	23	24	1	56
	5	0	0	2	1	2	5
celkem		158	322	215	68	3	766

- Kontingenční tabulka…
  - Hodnoty je třeba přehledně uspořádat (stejně jako u tabulky četností)
  - Pro data všech úrovní měření, nejvhodnější pro diskrétní prom. s málo hodnotami
  - Buňky mohou obsahovat absolutní četnosti, rel. četnosti (řádkové, sloupcové, celkové)
  - Poslední sloupec/řádek obsahuje tzv. sloupcové/řádkové marginální (relativní) četnosti
  - Je grafickou podobou je trojrozměrného sloupcový diagramu či histogramu (může obsahovat i intervaly)
  - Relativně vysoké četnosti v jedné z diagonál naznačují lineární provázanost proměnných

AJ: contingency table, crosstabulation, cells, row/column marginal frequencies, linear realtionship (vs. curvilinear (non-linear) realtionship), 3D barchart, 3D histogram
Fuj: Tab.7.2(s239) je správně kontingenční tabulka, korelační tabulka je něco jiného

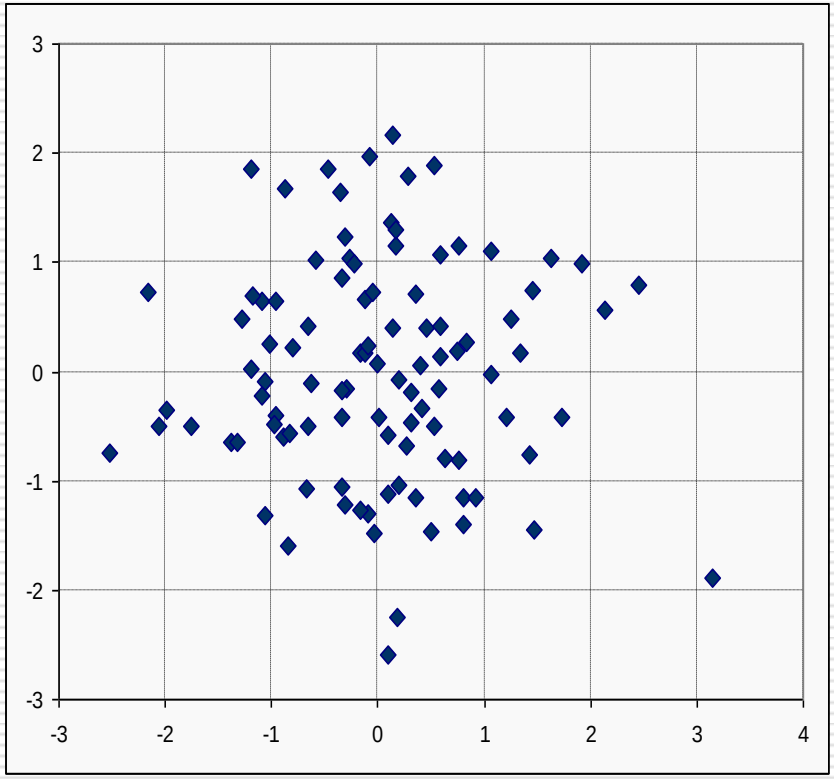
Grafická zobrazení dvourozměrného rozdělení



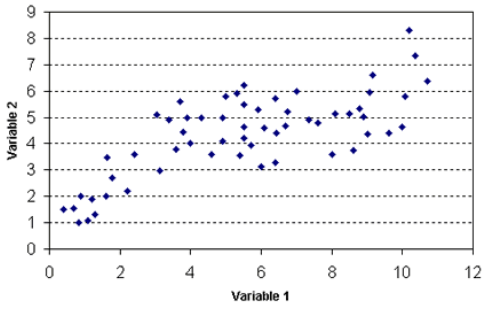


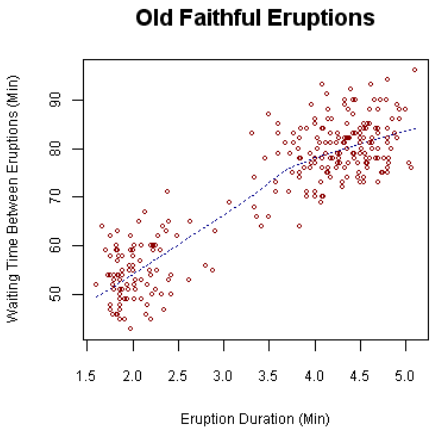
Bodový graf - scatterplot

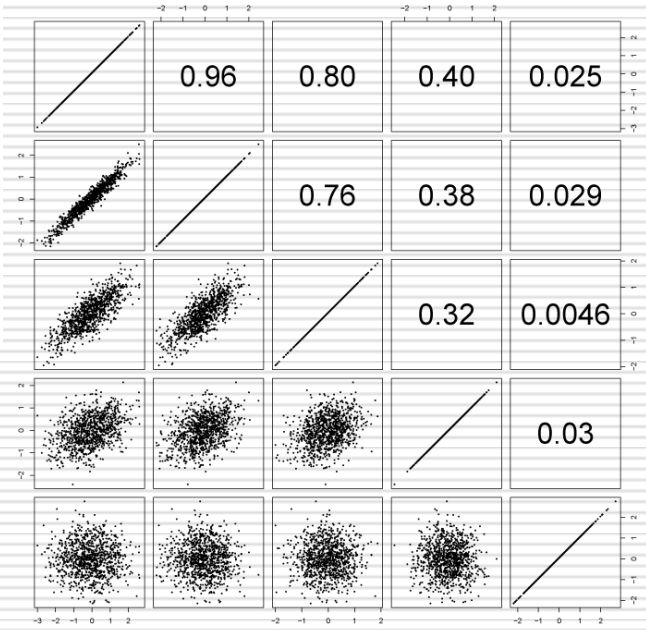
- Bodový graf – scatterplot
- Nahrazuje kontingenční tabulku, jsou-li obě proměnné spojité; pro proměnné s málo body měření nemá smysl
- Každá osa reprezentuje jednu proměnnou, každý bod je jedna zkoumaná osoba (jednotka)
- Poskytuje tím lepší evidenci o vztahu dvou proměnných…
  - …čím více měření jsme provedli
  - …čím přesnější jednotlivá měření byla
- Parametrem počtu měření může být např. velikost bodu
  - Frequency scatterplot

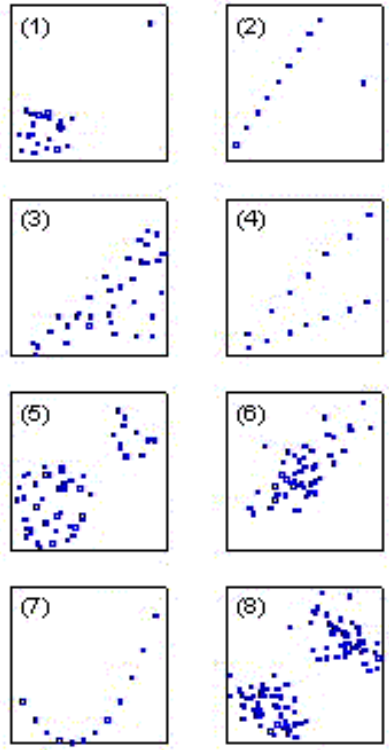


Různé podoby/druhy vztahu









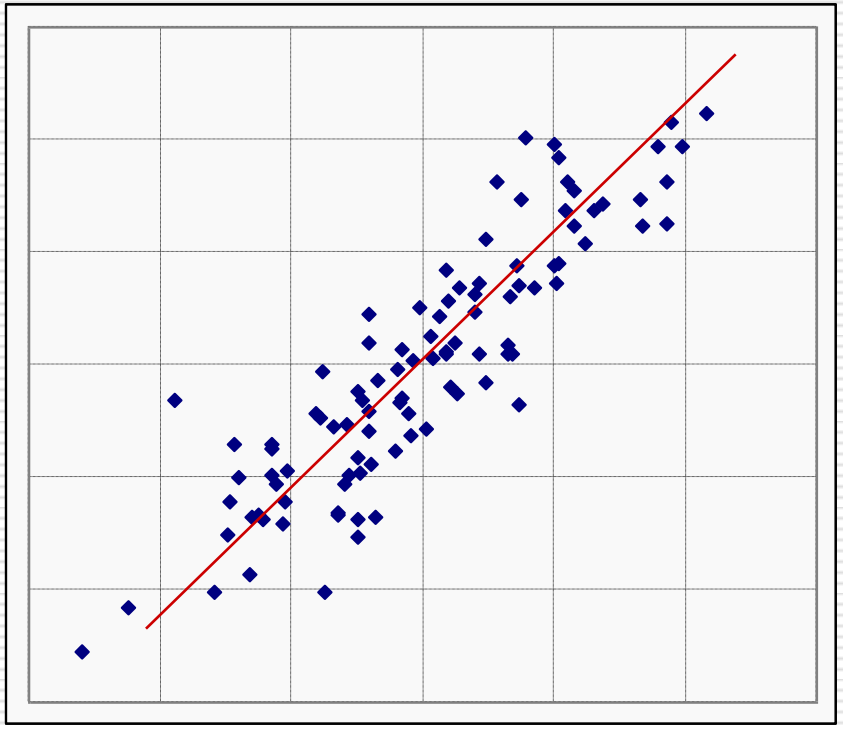
Pouze takto vypadající scattery zobrazují vztah mezi 2 proměnnými, který je lineární a dobře (=smysluplně, výstižně) popsatelný pomocí Pearsonova korelačního koeficientu. U ostatních jde buď o vztahy nelineární, nebo je problém v heterogenitě, outlierech…

Lineární souvislost, vztah

- Lineární vztah je to, co se obvykle míní slovem korelace.
- Je to monotónní vztah, který se dá popsat slovy čím více X, tím více/méně Y.
- Projevuje se tak, že scatterplot se dá proložit „ideální“ přímkou
  - $y = ax + b$

  Tato funkce/přímka popisuje strmost vztahu.

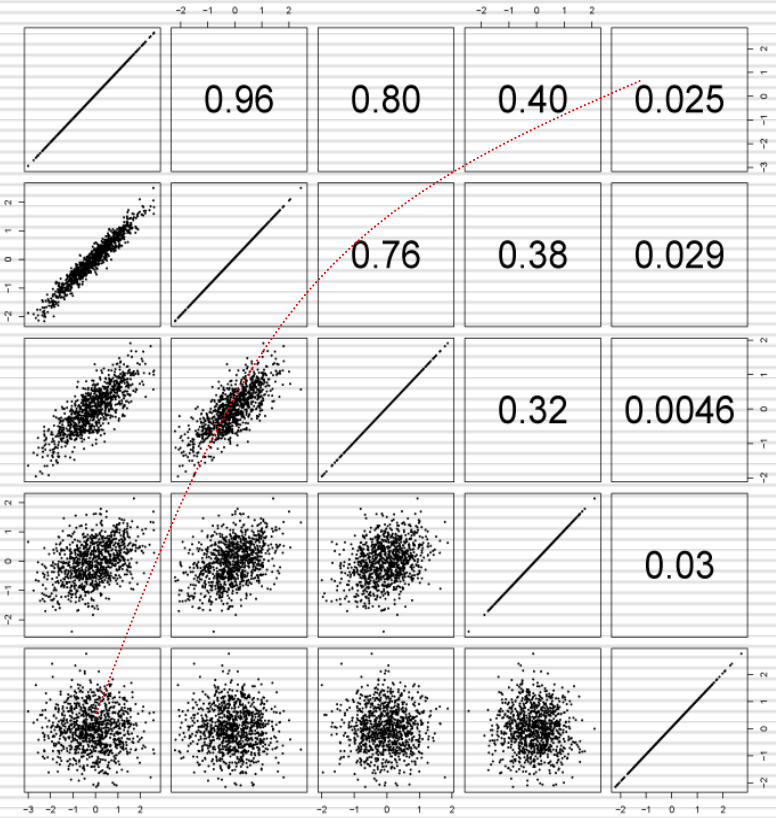
  Korelace popisuje **těsnost** vztahu.



AJ: linear association, correlation, monotonous relationship

Těsnost vztahu

- Čím těsnější (=intenzivnější, silnější) vztah 2 proměnných je, tím jsou body více nahuštěny okolo nějaké přímky
- Těsnost nesouvisí se sklonem té přímky, ale pouze s tím, jak moc se scatterplot podobá přímce.
- Těsnost se udává bezrozměrným číslem od 0 do 1, kde 0=žádný vztah(těsnost) a 1= maximální vztah
  (data na diagonále v obrázku napravo)
- Znaménko udává, zda jde o vztah čím víc, tím víc (+) nebo o vztah čím víc, tím míň (-)
- Rozsah je tedy od -1 do 1
- Těsnost -> kovariance



AJ: strength of association/relationship/correlation, positive relationship, negative(inverse) relationship

Kovariance (=sdílený rozptyl)

- Míru těsnosti lineárního vztahu dvou proměnných lze vyjádřit číselně
- Kovariance vypovídá o míře „sdíleného rozptylu“

$$c_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} x_i y_i$$

Vzpomeňte si na výpočet rozptylu. Ten byl $\Sigma x^2 / (n - 1)$. Tohle je $\Sigma xy / (n - 1)$. Místo x*x je tu x*y, proto je to ko-variance

Tato suma je tím vyšší čím máme v sadě dat více dvojic xy, u nichž je hodnota x i y nadprůměrná nebo podprůměrná. Sumu naopak snižují dvojice, kde je jedna hodnota nadprůměrná a druhá podprůměrná.

- kde *x, y* jsou deviační skóry, tj. odchylky od průměru
- Kovariance je stejně jako rozptyl nepraktická – výsledek je v jakýchsi „jednotkách na druhou“

AJ: covariance, shared variance

Korelace (=standardizovaný sdílený rozptyl)

- Chceme-li se zbavit obtížně interpretovatelných jednotek u kovariance, dosáhneme toho podobně jako při výrobě z-skórů – podělením deviačního skóru příslušnou směrodatnou odchylkou (=standardizace)

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_1 - m_x}{s_x}\right)\left(\frac{y_1 - m_y}{s_y}\right)$$

- Zakroužkovanou část vzorce už ale známe – to je transformace na z-skór. Korelace jednodušeji je tedy:

$$r_{xy} = \frac{\sum z_x z_y}{n-1}$$

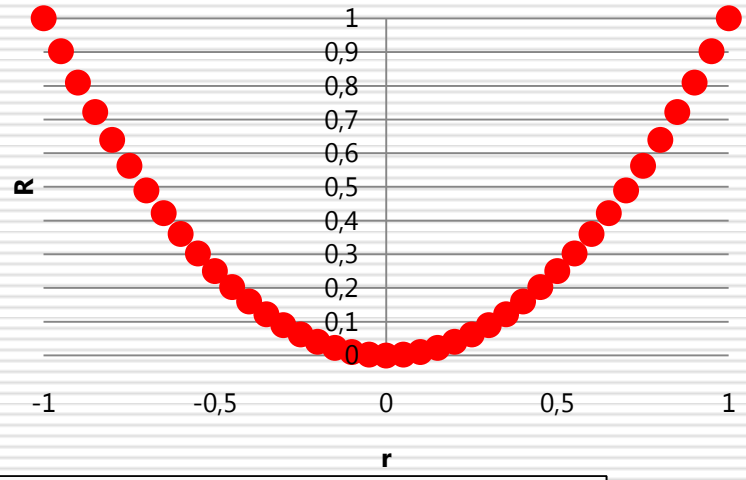
Vlastnosti popsaného koeficientu korelace I.

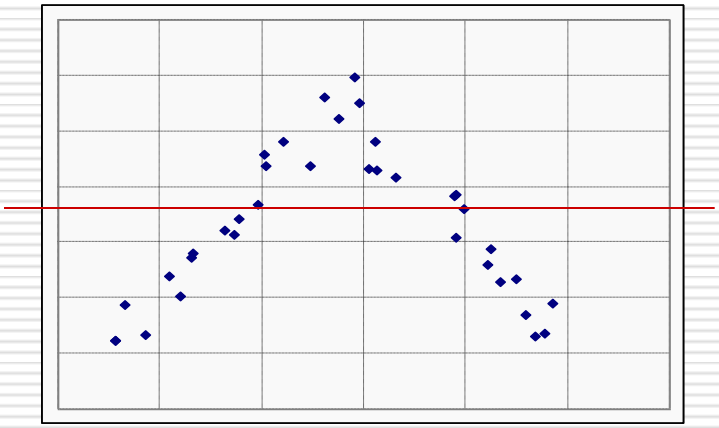
- Jde o tzv. Pearsonův součinový, momentový koeficient korelace
  - patří tedy do kategorie momentových ukazatelů (viz předchozí přednáška) a platí pro něj podobné věci:
    - nutná intervalová a vyšší úroveň měření
    - velký vliv odlehlých hodnot na výsledek
    - je vhodný pro popis normálně rozložených proměnných
    - vyjadřuje pouze sílu(těsnost) lineárního vztahu
- Nabývá hodnot v rozmezí -1 až 1
  - 0 = žádný vztah
  - 1(-1) = dokonalý kladný (záporný) vztah; identita proměnných
- Korelace nepopisuje funkční vztah dvou proměnných, ale pouze jeho směr a těsnost.

AJ: Pearson's product-moment correlation

Vlastnosti Pearsonova koeficientu korelace II.

- Je vázán na homogenitu souboru
- Není aditivní
- $r^2 = R$ = koeficient determinace (někdy D)
  - = proporce sdíleného rozptylu
  - V důsledku toho:
    0,3-0,1 ≠ 0,7-0,5
- $r = 0$ neznamená, že mezi rozděleními proměnných není žádná souvislost, znamená pouze, že mezi nimi není *lineární* vztah.

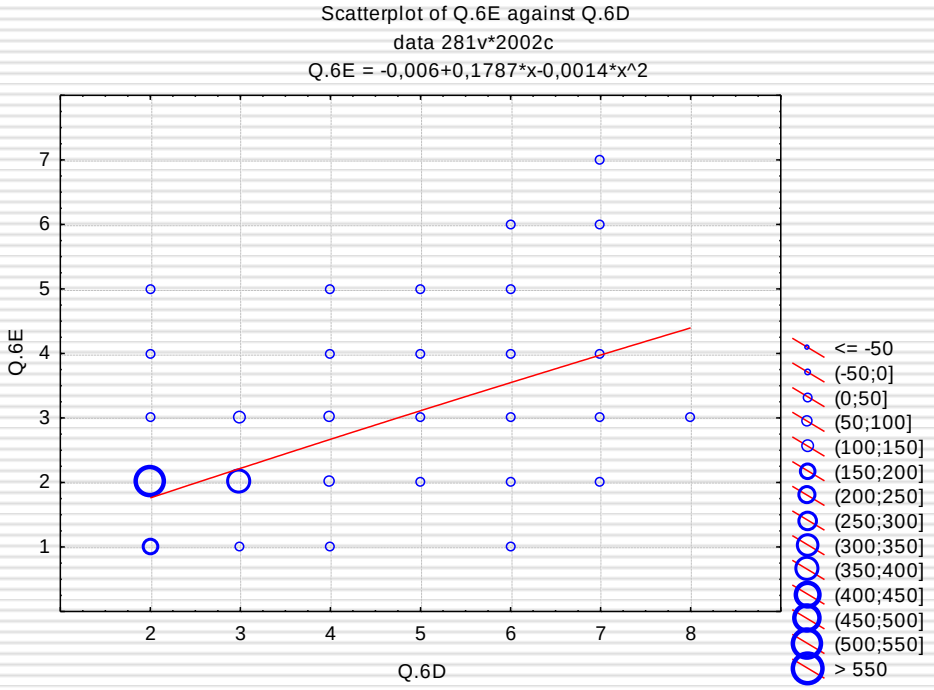




AJ: sample/population homogeneity, additivity, coefficient of determination

Vlastnosti Pearsonova koeficientu korelace III.

- Kdy nemá korelace smysl?
  - V1: Kolik hodin denně sledujete televizi?
  - V2: Kolik hodin denně sledujete televizní zpravodajství?
  - Proč? ☺
- Korelace proměnných se společnou příčinnou:
  - Swoboda: platy kněžích a ceny vodky v průběhu doby korelují!
  - IQ dětí a velikost a jejich výška prý také…
  - … kovariance proměnných se společnou příčinnou je základem dalších metod analýzy dat v psychologii: analýzy reliability a faktorové analýzy.



Korelační koeficienty pro pořadová data

- vhodné nejen pro pořadová data, ale i pro intervalová, která mají rozložení výrazně odlišné od normálního
- zachycují i nelineární monotónní vztahy (viz Hendl, s260)
- ukazatele toho, nakolik jsou pořadí podle korelovaných dvou proměnných stejná
- Spearmanův koeficient rhó – ρ, r_s
  - založený na velikosti rozdílů v pořadí
  - ekvivalentem Pearsonova koeficientu na pořadových datech
  - lze interpretovat r^2
- Kendallův koeficient tau – τ (s variantami „b“ nebo „c“)
  - založený na počtu hodnot (prvků výběrového souboru) mimo pořadí
  - vyjadřuje spíše pravděpodobnost, že se prvky výběrového souboru uspořádají podle obou proměnných do stejného pořadí

AJ: Spearman (rank correlation) rho, Kendall tau (-b,-c), rank

Korelační koeficienty další

- korelačních koeficientů existuje velké množství
- specifická užití – např. ϕ
- zjednodušení ručních výpočtů – např. r_{pb}
- ještě budeme mluvit o vztazích mezi nominálními proměnnými…

!! Korelace neznamená kauzalitu, jde spíše o koincidenci !!

AJ: phi, point-biserial correlation