

PSY117/454

Statistická analýza dat v psychologii

Přednáška 5

ZOBRAZENÍ DVOUROZMĚRNÝCH DAT KORELAČNÍ KOEFICIENT

Všichni žijeme v matrixu.

V minulých dílech jsme viděli/y:

- Frekvence = četnosti
 - Procenta = relativní četnosti
 - Údaje o střední hodnotě:
 - Modus, median, průměr
 - Údaje o variabilitě:
 - minimum, maximum, rozpětí
 - rozptyl, směrodatná odchylka
 - A mnohé další...
 - ...popisné statistiky.
 - Popisné statistiky jsou tzv. jednorozměrné
 - Cílem výzkumu je často odhalit vzájemné vztahy dvou proměnných. Metodou je analýza dvourozměrného rozložení.
-

Analýza vztahů mezi dvěma proměnnými

Souvisí nějak...?

- Výška a váha
- Znamky u jednotlivých předmětů
- Znamky a intelekt
- Úzkost a depresivita

Terminologická pozn.
„Úroveň“ (level) proměnné je zde použita ve významu „hodnota“. Např. proměnná „pohlaví“ má 2 úrovně – mužské a ženské. ...

□ Roste úroveň proměnné x s proměnnou y ?

- Je intelekt dobrým prediktorem školního úspěchu? (=Jak dobře můžeme ze znalosti IQ odhadnout známky?)
- Čím je x vyšší/nížší, tím má y tendenci být vyšší/nížší...

□ Na pořadové a vyšší úrovni

... a zde je termín „úroveň“ použit ve významu „úroveň měření“

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...

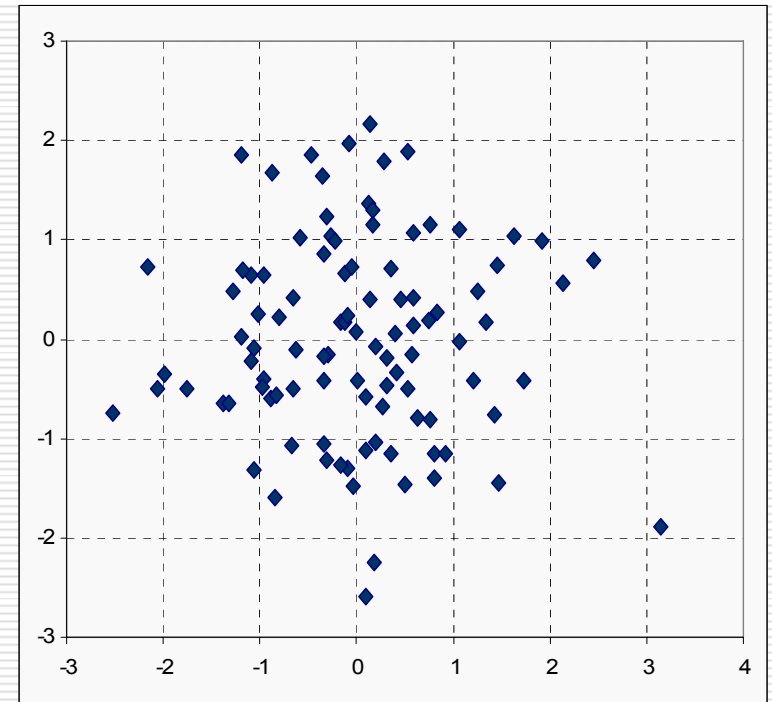
- Hodnoty je třeba přehledně uspořádat (stejně jako u tabulky četnosti)
- Pro data všech úrovní měření, nejvhodnější pro diskrétní prom. s málo hodnotami
- Buňky mohou obsahovat absolutní četnosti, rel. četnosti (řádkové, sloupcové, celkové)
- Poslední sloupec/řádek obsahuje tzv. sloupcové/řádkové marginální (relativní) četnosti
- Její grafickou podobou je trojrozměrný sloupcový diagram či histogram
- Lineární vztah se projevuje vysokými četnostmi na jedné z diagonál (zde červená elipsa)

-AJ: contingency table, crosstabulation, cells, row/column marginal frequencies, linear relationship (vs. curvilinear (non-linear) relationship), 3D barchart, 3D histogram

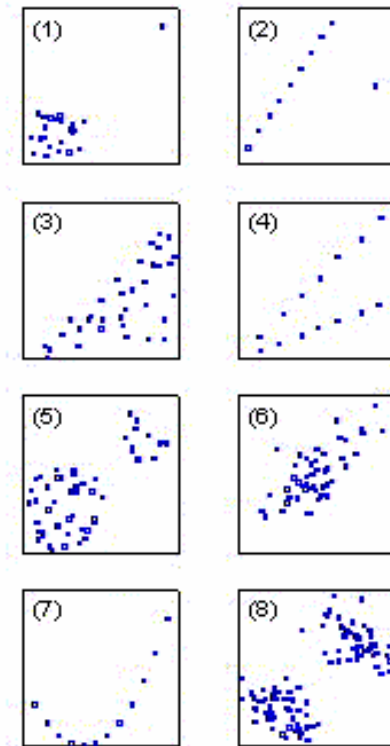
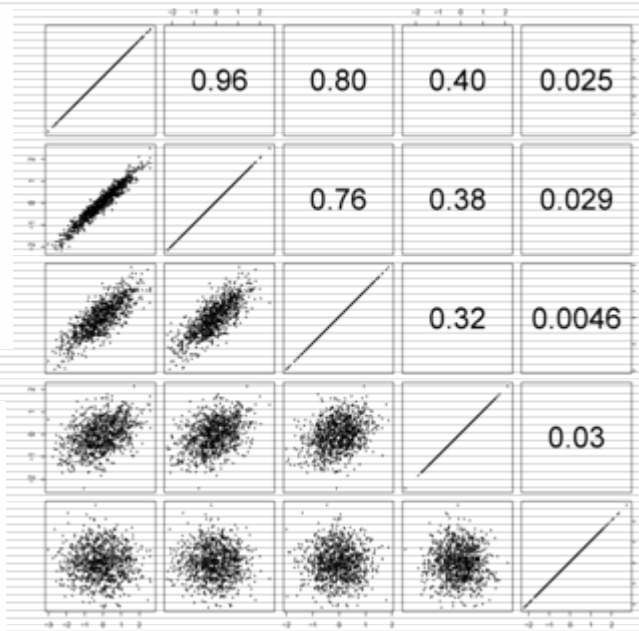
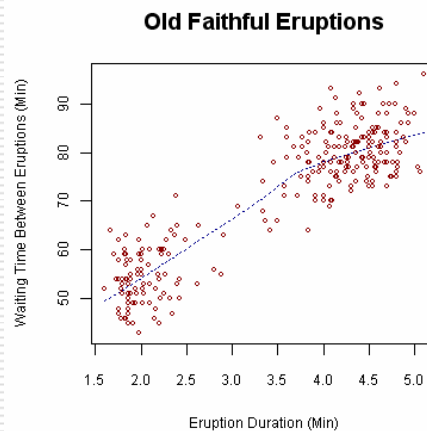
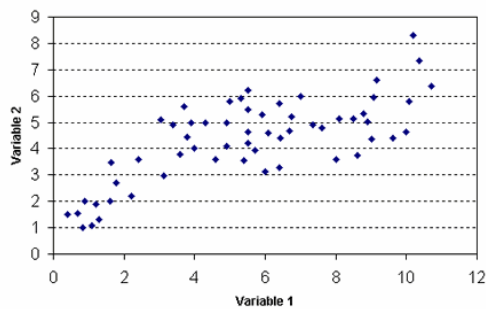
Fuj: Tab.7.2(s239) je správně kontingenční tabulka, korelační tabulka je něco jiného

Bodový graf - scatterplot

- Bodový graf – scatterplot
- Nahrazuje kontingenční tabulku, jsou-li obě proměnné spojité
- 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



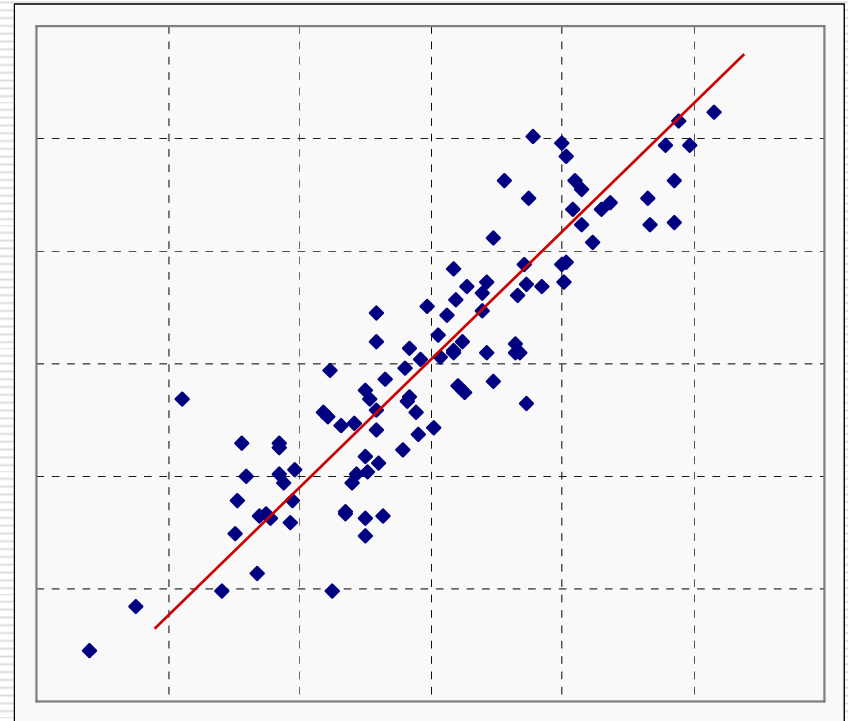
Různé podoby/druhy vztahu



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...

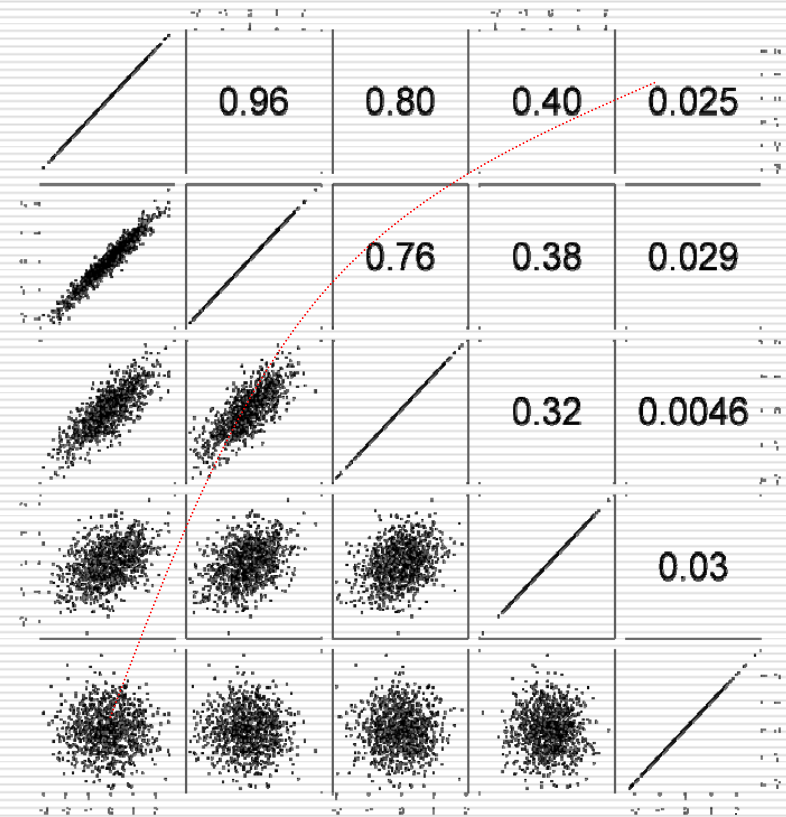
Lineární souvislost, vztah

- 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.



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



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“

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)

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

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

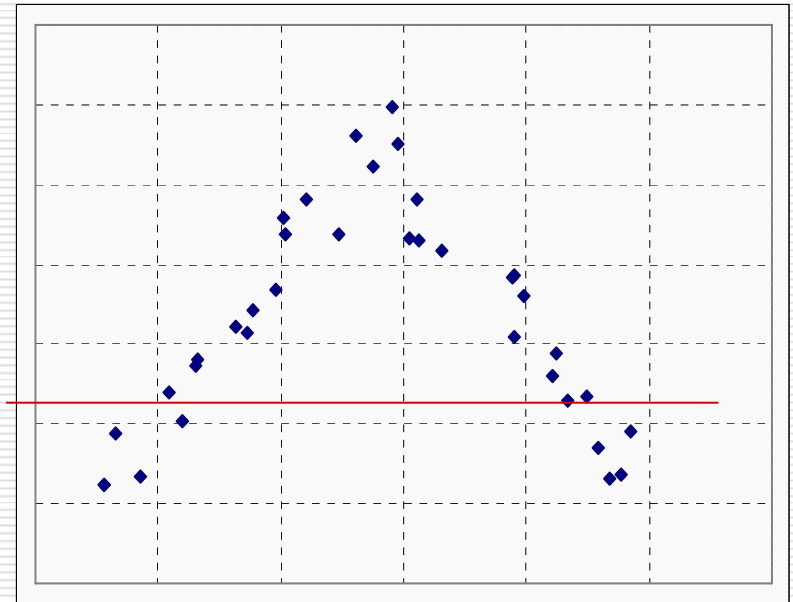
$$r_{xy} = \frac{\sum z_x z_y}{n-1}$$

Vlastnosti popsaného koeficientu korelace I.

- Jde o tzv. Pearsonův součinný, 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

Vlastnosti Pearsonova koeficientu korelace II.

- Je vázán na homogenitu souboru
- Není aditivní
- $r^2 = R^2 =$ koeficient determinace (někdy D)
 - = proporce sdíleného rozptylu
- $r = 0$ neznamená, že mezi proměnnými není žádný vztah, znamená, že mezi nimi není *lineární* vztah (viz obr.)



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 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 mimo pořadí
 - vyjadřuje spíše pravděpodobnost, že se podle obou proměnných uspořádají do stejného pořadí

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 !!