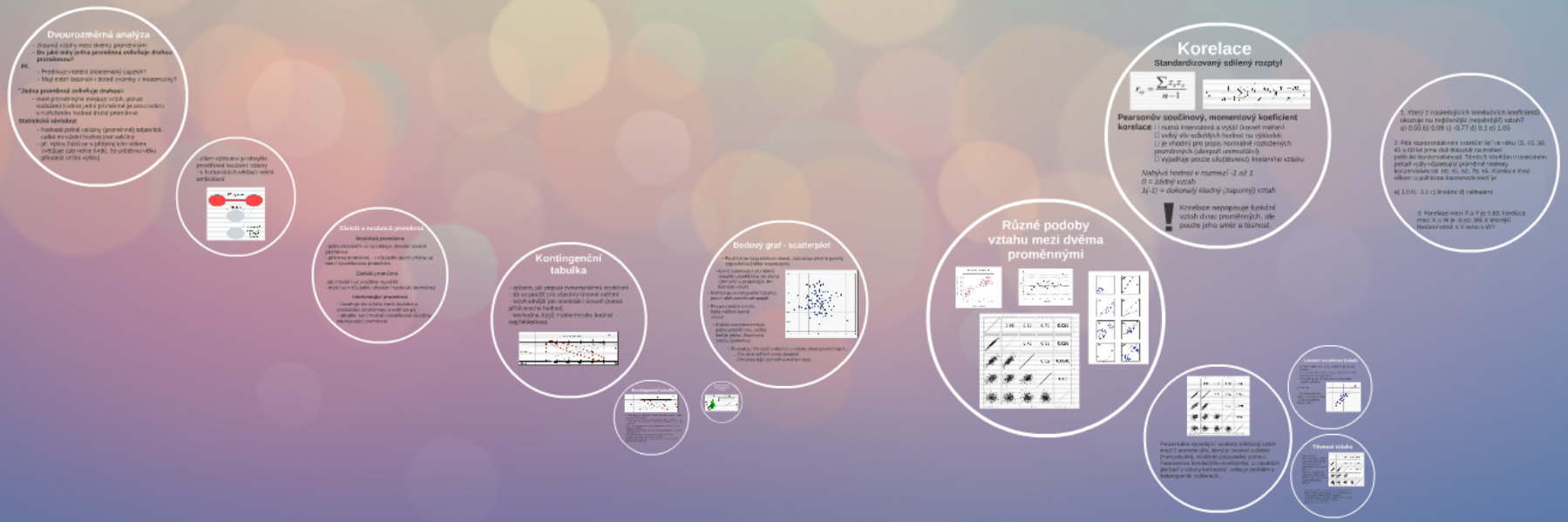
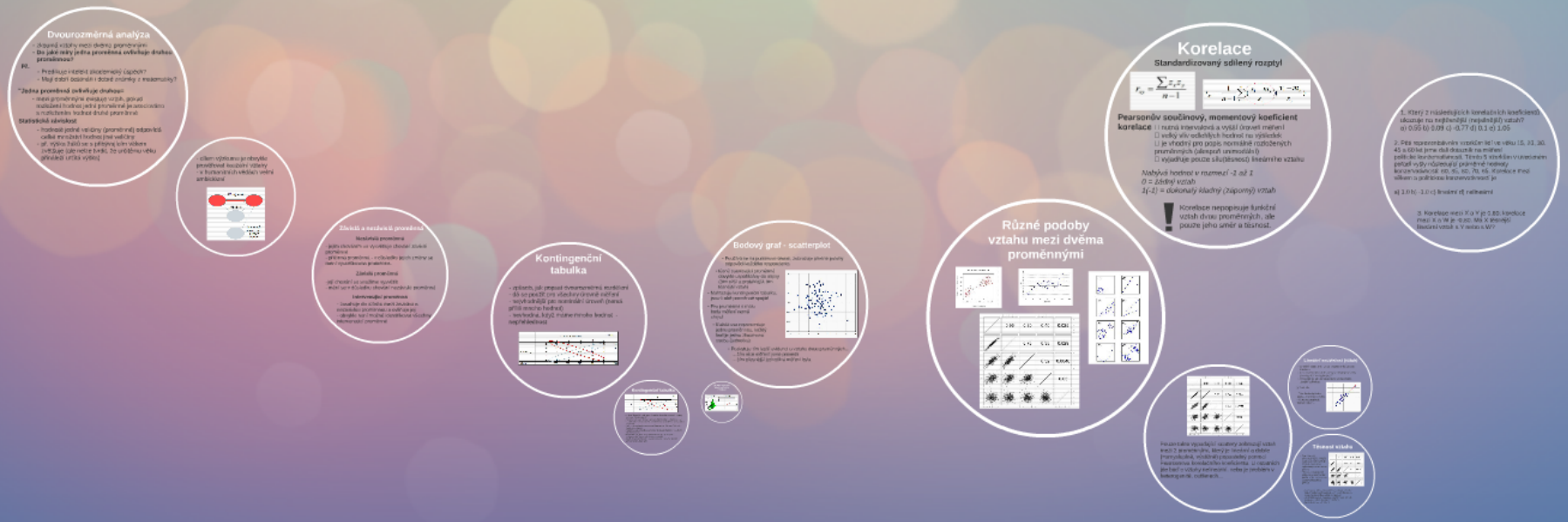
Zobrazení dvourozměrných dat, korelační koeficient

Mgr. Zuzana Szabó Lenhartová



Zobrazení dvourozměrných dat, korelační koeficient

Mgr. Zuzana Szabó Lenhartová



Dvourozměrná analýza

- zkoumá vztahy mezi dvěma proměnnými
- **Do jaké míry jedna proměnná ovlivňuje druhou proměnnou?**

Př.

- Predikuje intelekt akademický úspěch?
- Mají dobří češtináři i dobré známky z matematiky?

"Jedna proměnná ovlivňuje druhou=

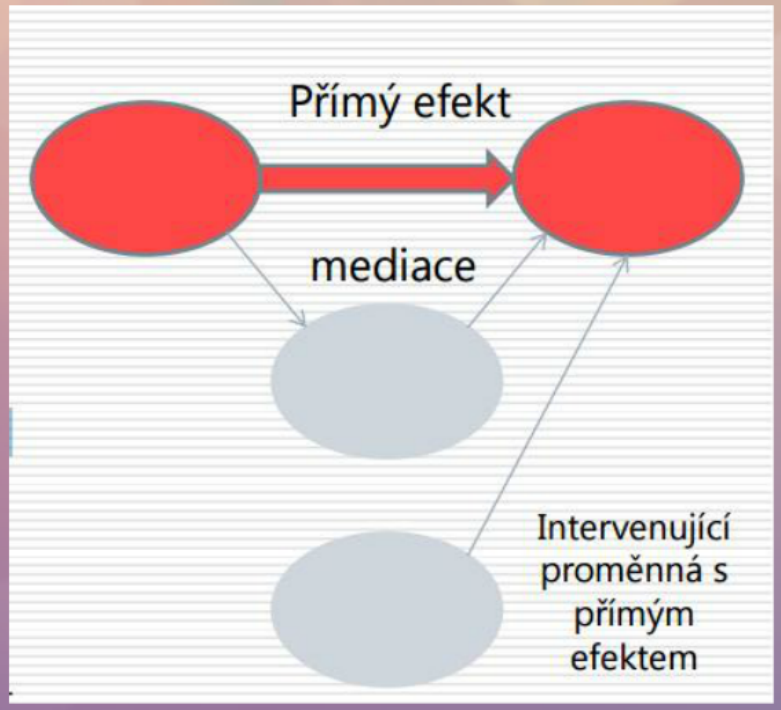
- mezi proměnnými existuje vztah, pokud rozložení hodnot jední proměnné je asociováno s rozložením hodnot druhé proměnné

Statistická závislost

- hodnotě jedné veličiny (proměnné) odpovídá celké množství hodnot jiné veličiny
- př. výška žáků se s přibývajícím věkem zvětšuje (ale nelze tvrdit, že určitému věku přináleží určitá výška)

- cílem v
prověřov
- v huma
ambicióz

- cílem výzkumu je obvykle prověřovat kauzální vztahy
- v humanitních vědách velmi ambiciózní



Závislá a nezávislá proměnná

Nezávislá proměnná

- jejím chováním se vysvětluje chování závislé proměnné
- příčinná proměnná - v důsledku jejich změny se mění vysvětlovaná proměnná.

Závislá proměnná

-její chování se snažíme vysvětlit
- mění se v důsledku chování nezávislé proměnné

Intervenující proměnná

- zasahuje do vztahu mezi závislou a nezávislou proměnnou a ovliňuje jej
- obvykle není možné identifikovat všechny intervenující proměnné

Kontingenční tabulka

- způsob, jak popsat dvourozměrná rozdělení
- dá se použít pro všechny úrovně měření
- nejvhodnější pro nominální úroveň (nemá příliš mnoho hodnot)
- nevhodná, když máme mnoho hodnot - nepřehlednost

		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

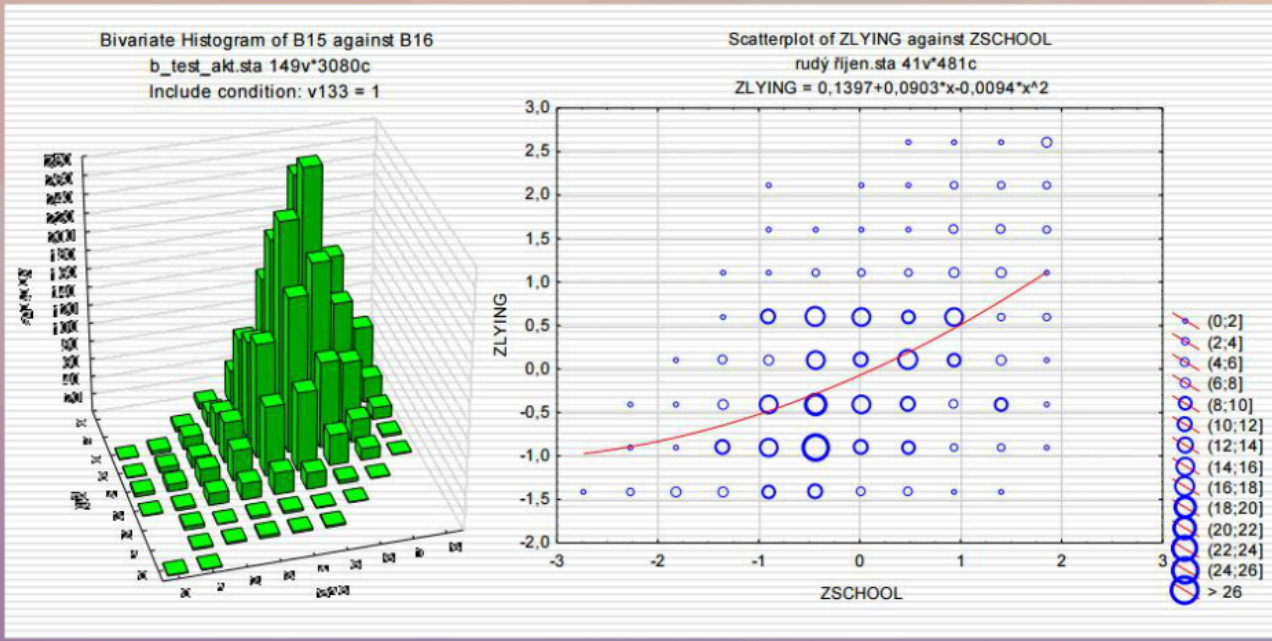
Kontingenční tabulka

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

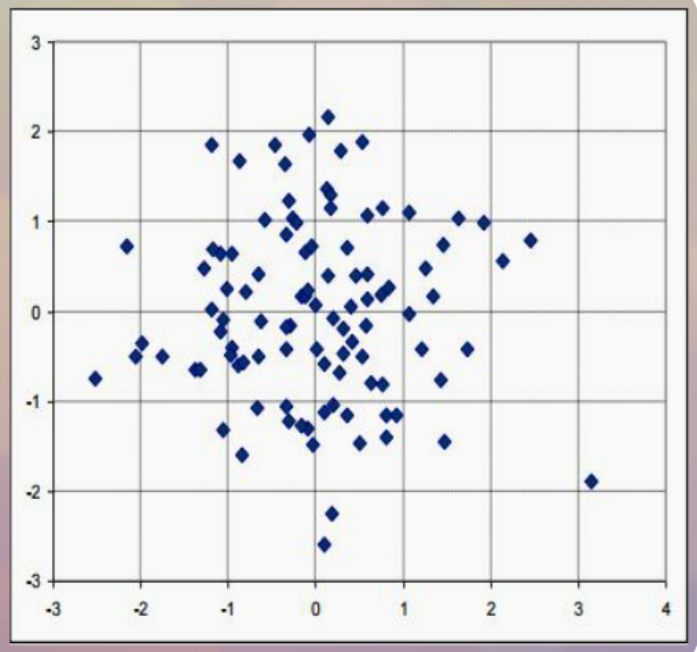
- v hlavní diagonále kontingenční tabulky více nakumulované hodnoty než jinde - lineární trend
- Hodnoty je třeba přehledně uspořádat (stejně jako u tabulky četností)
- Pro data všech úrovní měření, nejvhodnější pro diskrétní prom. s málo hodnotami
- Buňky mohou obsahovat absolutní četnosti, rel. četnosti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

Grafická zobrazení dvourozměrného rozdělení



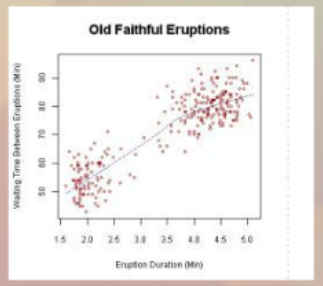
Bodový graf - scatterplot

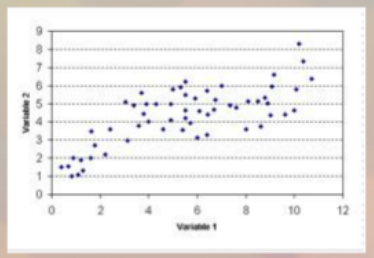
- Používá se na poměrové úrovni, zobrazuje přesné polohy odpovědi každého respondenta
  - těsně související proměnné obvykle uspořádány do elipsy (čím užší a protáhlejší, tím těsnější vztah)
- Nahrazuje kontingenční tabulku, jsou-li obě proměnné spojité
- Pro proměnné s málo body měření nemá smysl
  - Každá osa reprezentuje jednu proměnnou, každý bod je jedna zkoumaná osoba (jednotka)

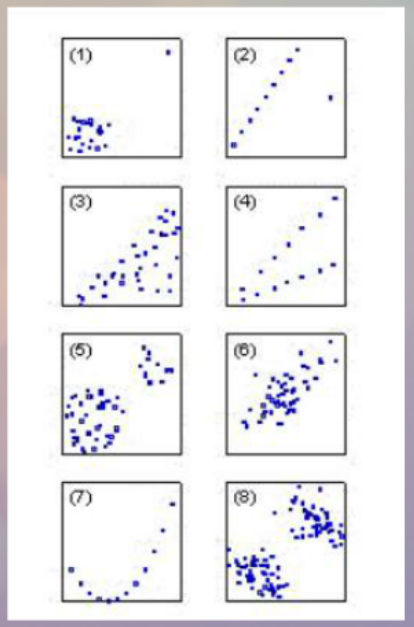


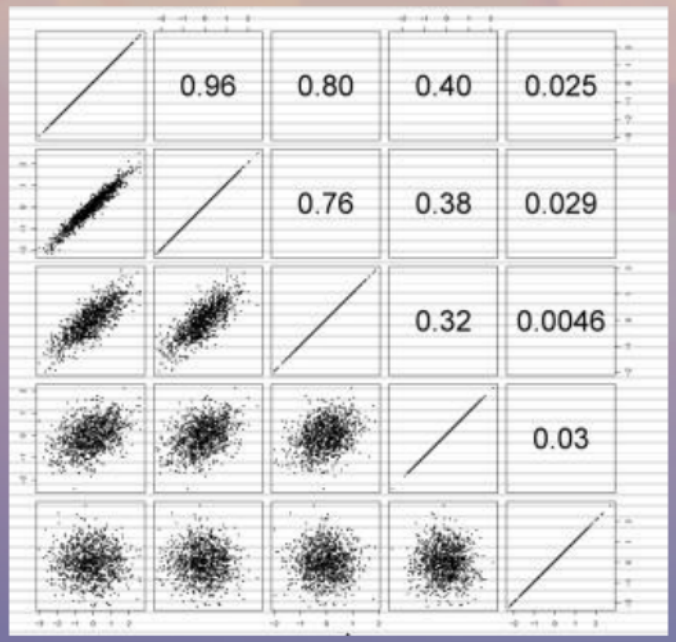
- Poskytuje tím lepší evidenci o vztahu dvou proměnných…
  …čím více měření jsme provedli
  …čím přesnější jednotlivá měření byla

Různé podoby vztahu mezi dvěma proměnnými

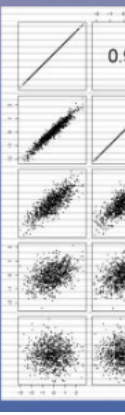




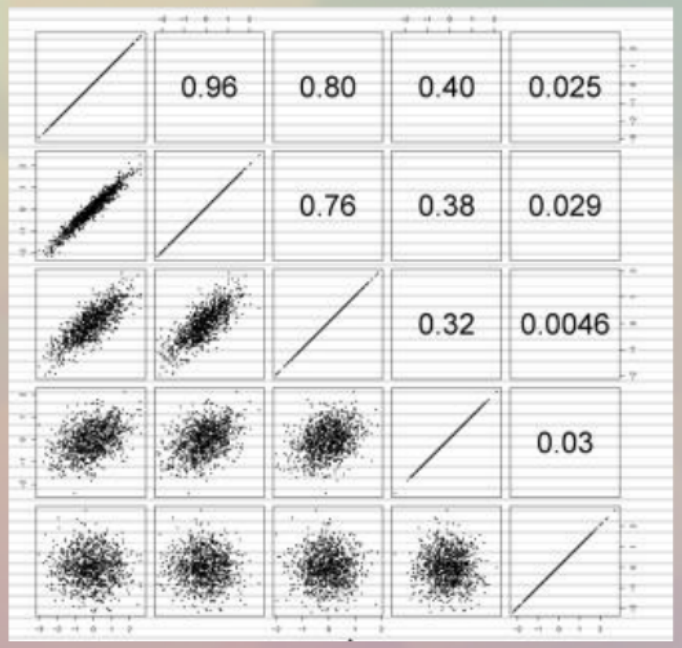




Korelac
vztah dv
pouze je



Prezi



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 vztah
  korelace.
- Je to monotónn
  čím více X, tím
- Projevuje se ta
  „ideální“ přímk

y = ax + b

- Tato funkce/přím
popisuje strmost
- Korelace popisu
těsnost vztahu.

Tě

- Čím těsnější
  (=intenzivnější
  vztah 2 proměn
  tím jsou body v
  nahuštěny oko
  přímky
- Těsnost nesou
  sklonem té přím
  pouze s tím, ja
  scatterplot pod
  přímce.

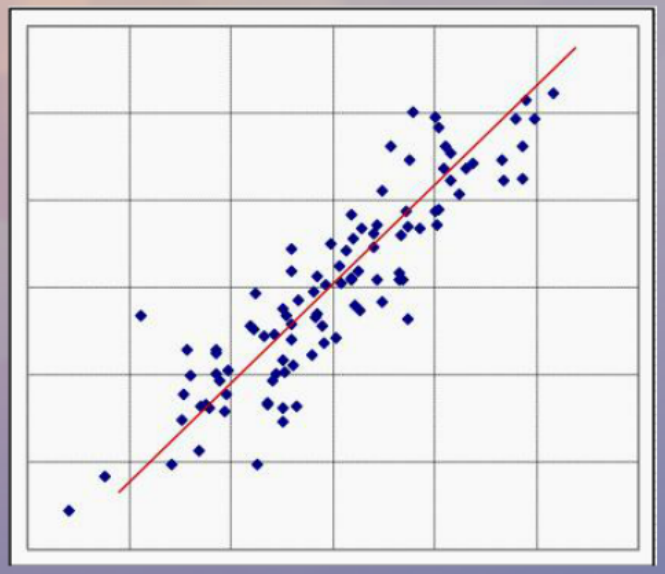
- Těsnost se u
  kde 0=žádný
  (data na diag
- Znaménko u
  (+) nebo o vz
- Rozsah je te

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

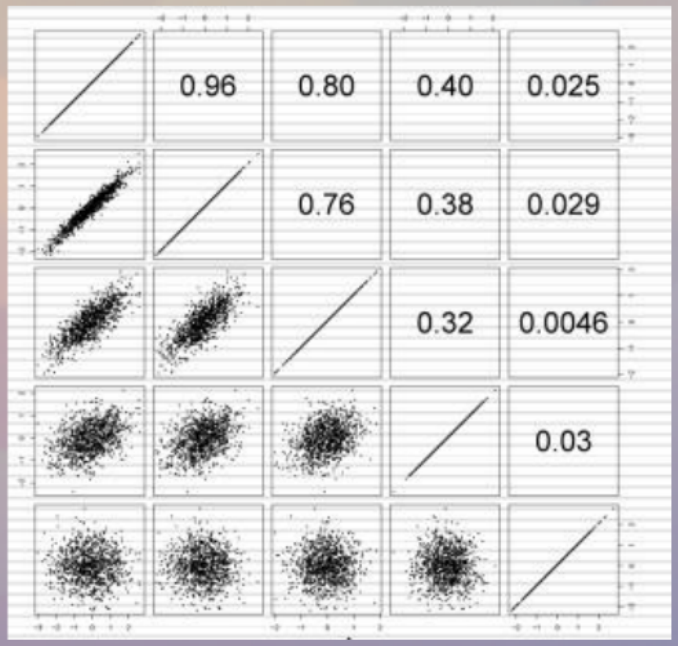
y = ax + b

- 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

Korelace

Standardizovaný sdílený rozptyl

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

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

Pearsonův součinový, momentový koeficient korelace

- nutná intervalová a vyšší úroveň měření
- velký vliv odlehlých hodnot na výsledek
- je vhodný pro popis normálně rozložených proměnných (alespoň unimodální)
- vyjadřuje pouze sílu(těsnost) lineárního vztahu

Nabývá hodnot v rozmezí -1 až 1
0 = žádný vztah
1(-1) = dokonalý kladný (záporný) vztah

! Korelace nepopisuje funkční vztah dvou proměnných, ale pouze jeho směr a těsnost.

oby

1. Který z následujících korelačních koeficientů ukazuje na nejtěsnější (nejsilnější) vztah?
a) 0,55 b) 0,09 c) -0,77 d) 0,1 e) 1,05

2. Pěti reprezentativním vzorkům lidí ve věku 15, 20, 30, 45 a 60 let jsme dali dotazník na měření politické konzervativnosti. Těmto 5 vzorkům v uvedeném pořadí vyšly následující průměrné hodnoty konzervativnosti: 60, 85, 80, 70, 65. Korelace mezi věkem a politickou konzervativností je

a) 1.0 b) -1.0 c) lineární d) nelineární

3. Korelace mezi X a Y je 0,60; korelace mezi X a W je -0,80. Má X těsnější lineární vztah s Y nebo s W?

Zobrazení dvourozměrných dat, korelační koeficient

Mgr. Zuzana Szabó Lenhartová