

Míry asociace

1. obecná definice – síla a směr vztahu
 2. Pearsonův korelační koeficient
 3. míry asociace pro pořadová data
 4. míry asociace pro nominální data
-

Míry asociace

- míry asociace vyjadřují **těsnost vztahu proměnných** (a případně **směr** vztahu)
-

Míry asociace

- **těsnost (síla) vztahu** – vyjádřena absolutní hodnotou koeficientu
 - není shoda v tom, od jaké hodnoty je vztah považován za těsný (někdy uváděno >0.70 , jindy >0.30), středně těsný či slabý
-

Míry asociace

- **směr vztahu** – pouze u ordinálních a kardinálních proměnných, vyjádřen znaménkem koeficientu
 - **pozitivní vztah** – čím vyšší hodnoty jedné proměnné, tím vyšší hodnoty druhé proměnné
 - **negativní vztah** - čím vyšší hodnoty jedné proměnné, tím nižší hodnoty druhé proměnné
-

Pearsonův korelační koeficient

- u kardinálních dat můžeme jako míru asociace – vztahu mezi proměnnými použít **Pearsonův korelační koeficient**
 - **korelace**
 - ko = s, spolu, vzájemně
 - relace = vztah
 - korelace = vzájemný vztah proměnných
-

Pearsonův korelační koeficient

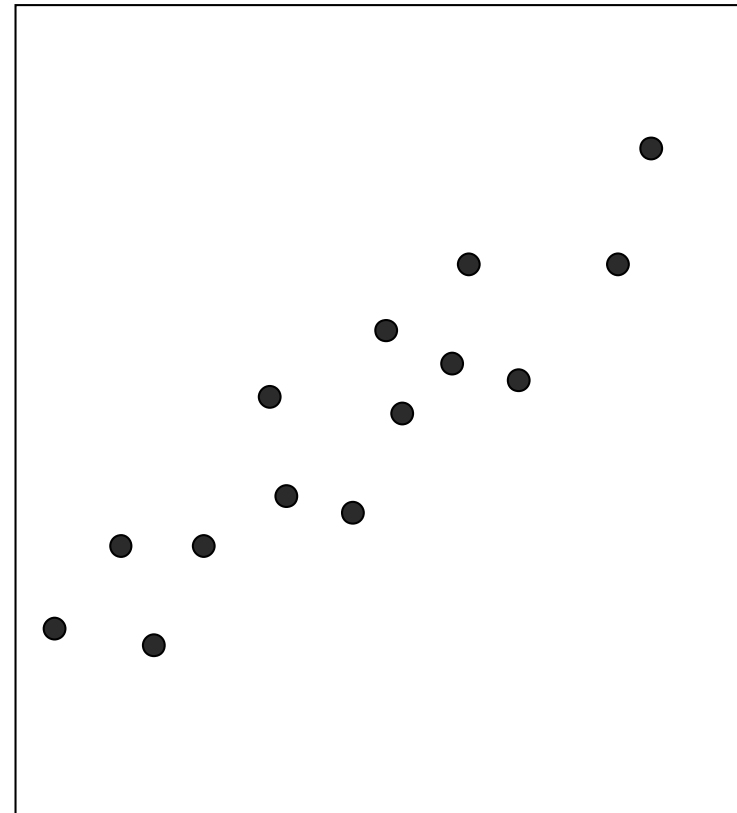
- absolutní hodnota koeficientu vyjadřuje **sílu (těsnotu) vztahu**
 - znaménko (+ nebo -) **směr vztahu**
 - rozsah -1 až +1**
 - označuje se **r**
-

Pearsonův korelační koeficient

- je mírou asociace **pouze pro lineární vztahy**
 - před výpočtem je vhodné zobrazit vztah mezi proměnnými také graficky – tzv. **scatter** (dvourozměrný tečkový diagram)
-

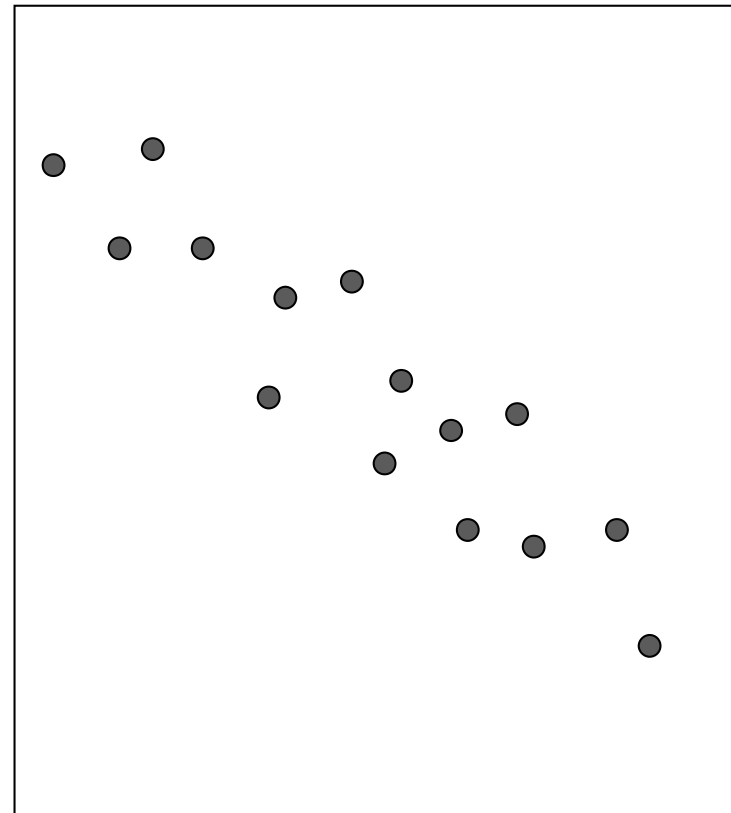
Scatter

- **pozitivní vztah** (přímá úměra) – čím vyšší hodnoty proměnné X , tím vyšší hodnoty proměnné Y
- $r > 0$



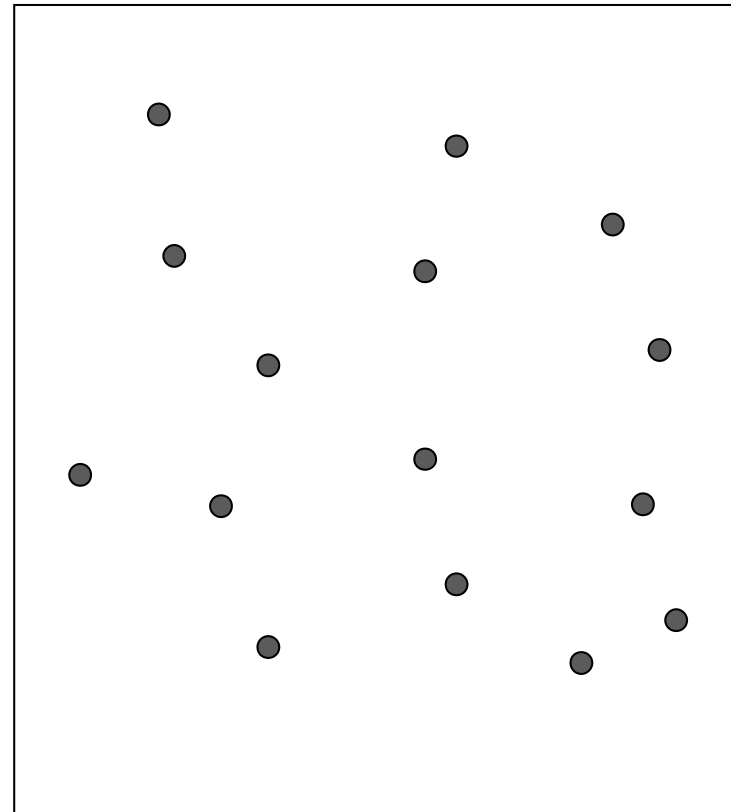
Scatter

- **negativní vztah** (nepřímá úměra) – čím vyšší hodnoty proměnné X, tím nižší hodnoty proměnné Y
- $r < 0$



Scatter

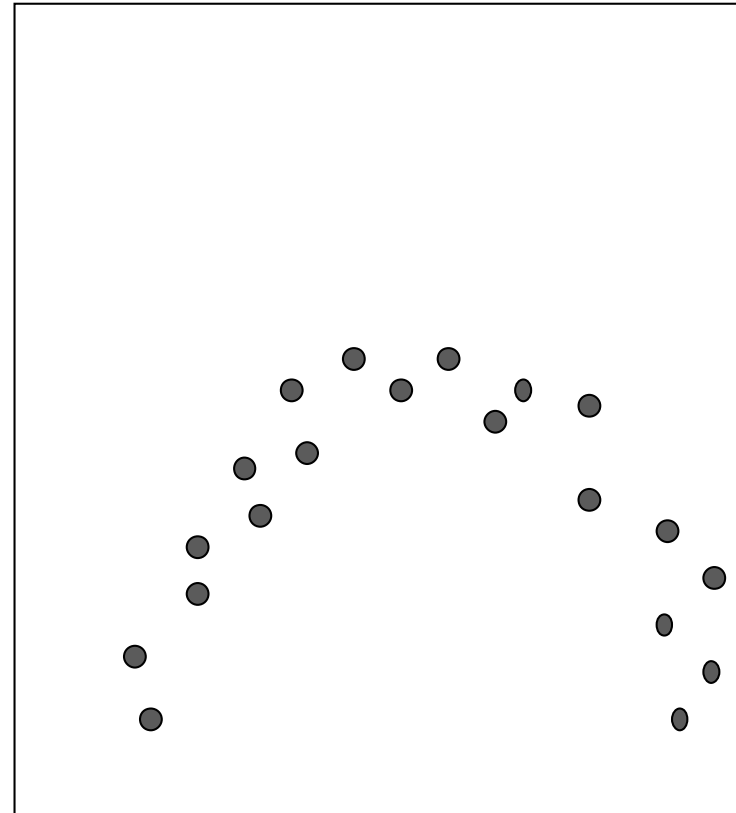
- **žádný vztah** -
hodnoty
proměnné X
nesouvisí s
hodnotami
proměnné Y
- $r = 0$



Scatter

□ **nelineární
vztah**

□ $r = 0$



Pearsonův korelační koeficient

- sám o sobě je deskriptivní statistikou, ale podobně jako u ostatních měř asociace je možno spočít **statistickou významnost**
 - nulovou hypotézou je zde většinou $\rho=0$
-

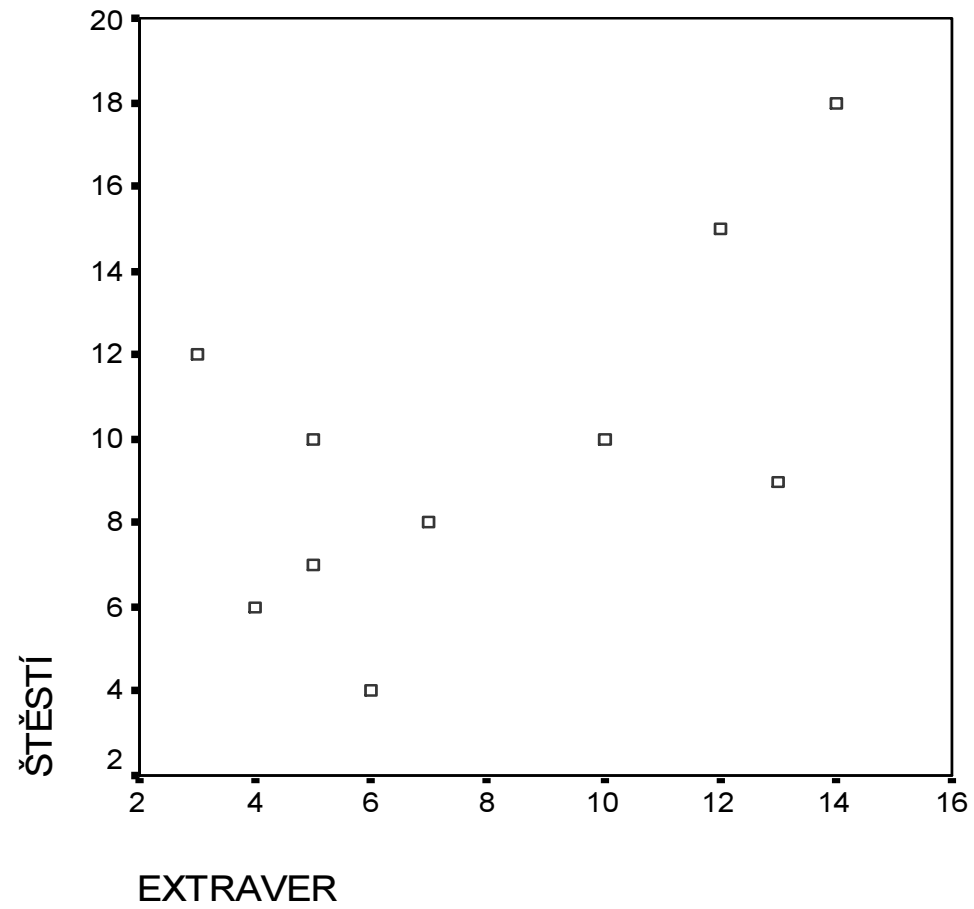
Příklad

- jak spolu souvisí pocit štěstí a míra extraverze?
 - 10 osob, 2 proměnné – skór z dotazníku štěstí a skór ze škály extraverze
-

Příklad

extraverze (x)	12	7	5	14	6	3	5	10	4	13
šťěstí (y)	15	8	7	18	4	12	10	10	6	9

Příklad



Příklad

□ $m_x = 7,90$; $s_x = 4,01$

□ $m_y = 9,90$; $s_y = 4,20$

Příklad

□ **výpočet r**

$$r = s_{xy} / s_x * s_y$$

□ $s_{xy} = \Sigma (x_i - \bar{x}) * (y_i - \bar{y}) / (n-1)$

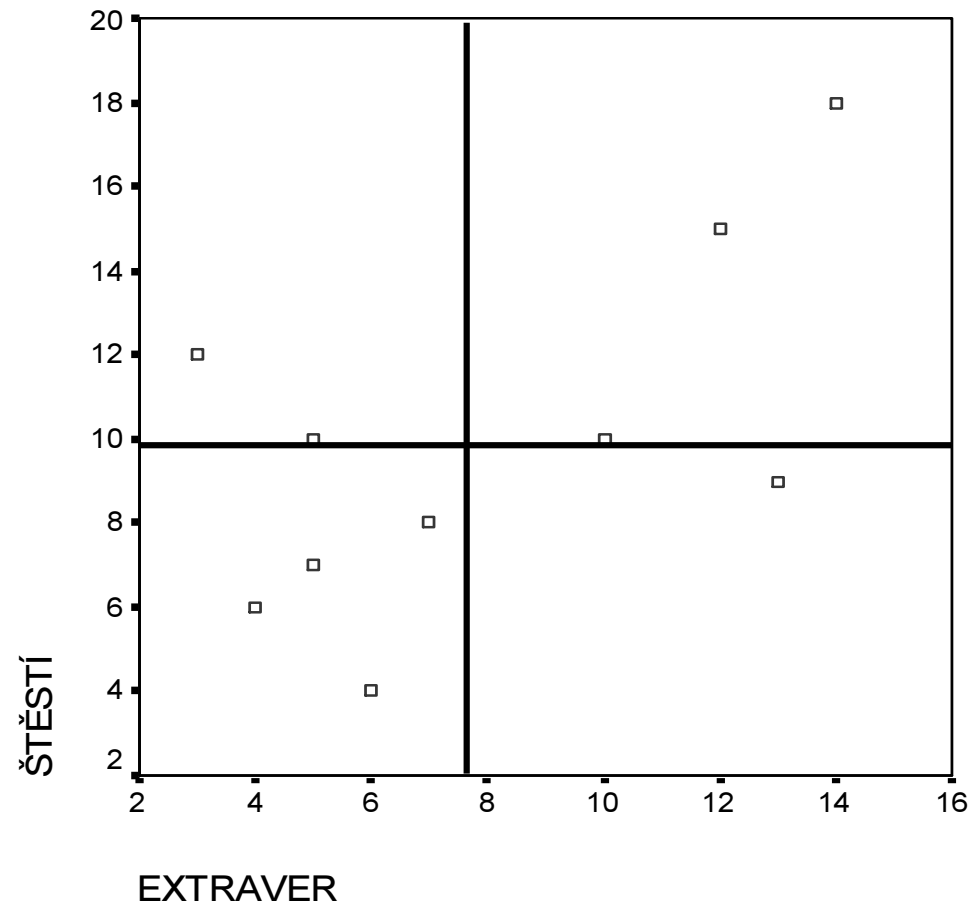
□ s_x, s_y jsou směrodatné odchylky

Příklad

$$\square r_{xy} = \sum x_i' * y_i' / (n-1)$$

$\square x_i'$ a y_i' jsou standardizované hodnoty proměnných x a y

Příklad



Příklad

$$\square \mathbf{s}_{xy} = [(5,1*4,1) + (-1,9*-0,9) + (-2,9*-2,9) + (8,1*6,1) + (-5,9*-1,9) + (2,1*-4,9) + (0,1*-2,9) + (0,1*2,1) + (-3,9*-3,9) + (-0,9*5,1)]/9$$

$$\square \mathbf{s}_{xy} = 91,9/9 = 10,21$$

$$\square \mathbf{r}_{xy} = 10,21 / (4,01*4,20) = 10,21/16,84$$

$$\square \mathbf{r}_{xy} = \mathbf{0,606}$$

Výstup v SPSS

Correlations

		EXTRAVER	ŠTĚSTÍ
EXTRAVER	Pearson Correlation	1	,606
	Sig. (2-tailed)	,	,064
	N	10	10
ŠTĚSTÍ	Pearson Correlation	↗ ,606	1
	Sig. (2-tailed)	,064	,
	N	10	10

Interpretace r

- není shoda v tom, jaká hodnota r je považována za těsný vztah
 - interpretace navržená Guilfordem:
 - <0.20 zanedbatelný vztah
 - $0.20-0.40$ nepříliš těsný vztah
 - $0.40-0.70$ středně těsný vztah
 - $0.70-0.90$ velmi těsný vztah
 - >0.90 extrémně těsný vztah
-

Interpretace r

- pro lepší interpretaci se koeficient korelace někdy převádí na **koeficient determinace (r^2)**
 - interpretuje se jako ukazatel, kolik rozptylu v jedné proměnné může být vysvětleno rozptylem ve druhé proměnné
-

Interpretace r

- v našem příkladu
 - $r = 0,606$
 - $r^2 = 0,367$
 - 36,7% rozdílů v míře štěstí můžeme vysvětlit rozdíly v míře extraverze
-

Interpretace r

□ **korelace neznamená příčinný vztah mezi proměnnými!!**

- ten můžeme ověřovat pouze experimentem, kdy jsou všechny ostatní proměnné udržovány konstantní, proměnná X předchází Y v čase atd.
-

Faktory ovlivňující r

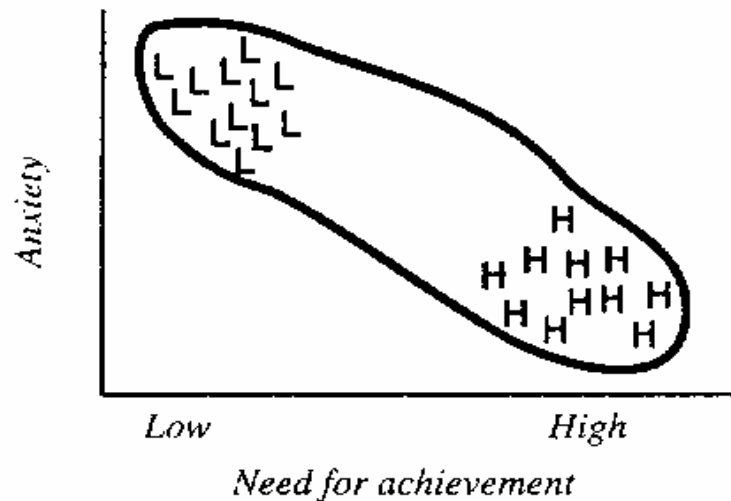
- omezený rozsah hodnot proměnné
 - použití extrémních skupin
 - nehomogenní soubor
 - odlehlé hodnoty
 - nelineární vztahy
 - reliabilita použitých nástrojů
-

Omezený rozsah hodnot

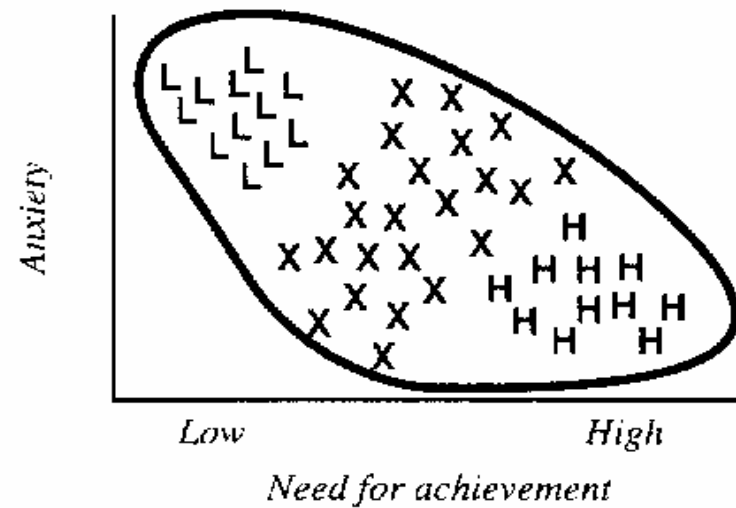
- omezený rozsah hodnot jedné nebo obou proměnných snižuje hodnotu r
 - stejně tak nízká variabilita (extrémní případ: pokud by všechny hodnoty jedné proměnné byly stejné, zákonitě $r=0$)
-

Použití extrémních skupin

- použití extrémních skupin (např. jen osob s vysokým IQ) vede k vyššímu r



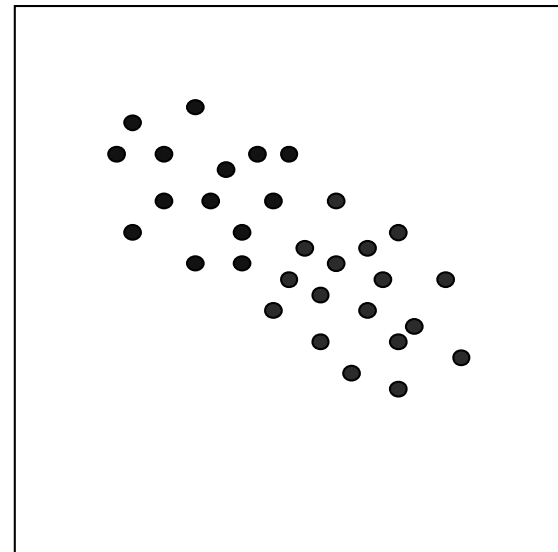
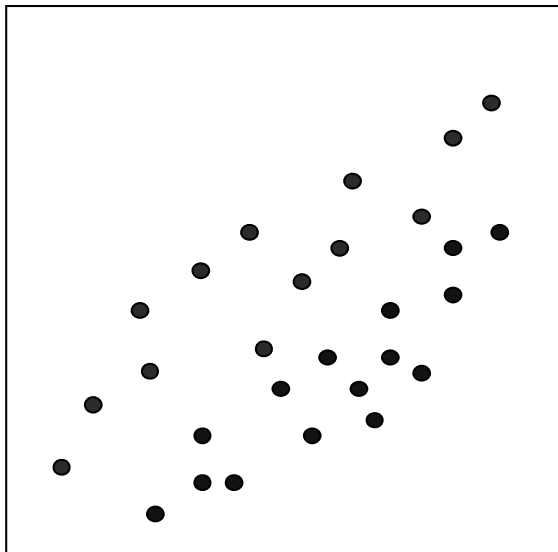
(a)



(b)

Nehomogenní soubor

- může zkreslit r jak směrem nahoru, tak dolů



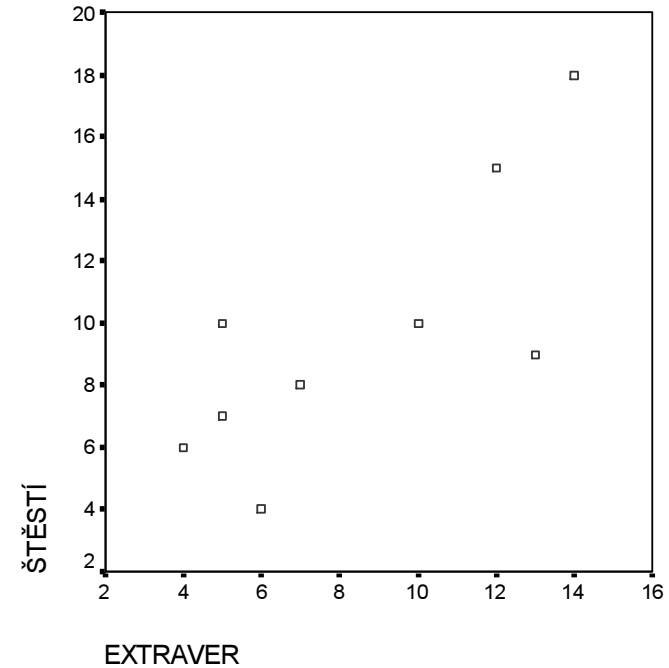
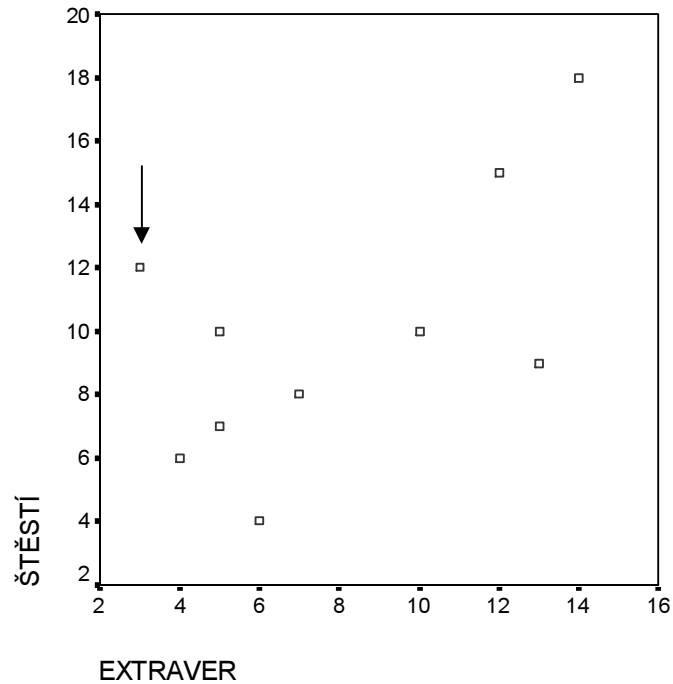
Odlehlé hodnoty

- extrémní hodnoty v jedné nebo obou proměnných mohou r výrazně zkreslit (nejen hodnotu, ale i směr), zvláště když je počet osob v souboru nízký
-

Extrémní hodnoty

□ $r = 0,606$

□ $r = 0,766$



Spearmanův koeficient

- pro pořadová data je možno spočítat **Spearmanův koeficient pořadové korelace** (ρ_s)
 - počítá se tak, že
 - hodnoty obou proměnných se seřadí od nejnižší po nejvyšší a přidělí se jim pořadové číslo
 - z těchto pořadí se pak počítá Pearsonův koeficient korelace
-

Spearmanův koeficient

extraverze (x)	12	7	5	14	6	3	5	10	4	13
pořadí (x)	8	6	3,5	10	5	1	3,5	7	2	9
šťěstí (y)	15	8	7	18	4	12	10	10	6	9
pořadí (y)	9	4	3	10	1	8	6,5	7	2	5

Spearmanův koeficient

- používá se i u kardinálních dat, pokud jsou přítomny odlehlé hodnoty
-

Kendallův koeficient

- používá se rovněž pro pořadová data
 - označuje se τ_k (Kendalovo tau)
 - princip výpočtu je jiný než u Spearmanova koeficientu
-

Kendallův koeficient

- seřadíme dvojice hodnot proměnných x a y tak, aby hodnoty proměnné x byly v pořadí od nejmenší po největší
 - pokud je mezi proměnnými x a y kladný vztah, pak by i hodnoty proměnné y měly být ve vzestupném pořadí
 - každou hodnotu proměnné y porovnáme se všemi následujícími hodnotami proměnné y
-

Kendallův koeficient

- pokud je $y_j > y_i$ (kdy $j > i$), pak nastává tzv. **konkordance** (P) – značí pozitivní vztah
 - pokud je $y_j < y_i$ (kdy $j > i$), pak nastává tzv. **diskordance** (Q) – naznačuje negativní vztah
 - Kendalovo **$S = P - Q$**
 - dělí se počtem možných konkordancí a diskordancí **$D = n * (n - 1) / 2$**
-

Kendallův koeficient

extraverze (x)	3	4	5	5	6	7	10	12	13	14
šťěstí (y)	12	6	7	10	4	8	10	15	9	18
konkordance	2	7	5	2	5	4	2	1	1	0
diskordance	7	1	1	3	0	0	1	1	0	0

29

14

Kendallův koeficient

□ $t_k = S/D$

□ $t_k = (P-Q)/[n*(n-1)/2]$

□ $t_k = (29-14)/[10*(10-1)/2]$

□ $t_k = 15/45$

□ **$t_k = 0,333$**

Kendallův koeficient

- pokud v datech existuje větší počet shod ($y_j = y_i$ nebo $x_j = x_i$), upravuje se hodnota D
 - tento modifikovaný koeficient se označuje jako Kendallovo tau-b
-

Míry asociace pro nominální data

- míry asociace pro nominální data ukazují pouze sílu vztahu dvou proměnných, nikoli směr či jiné informace o povaze vztahu
 - rozlišujeme míry založené na chí-kvadrátu a míry PRE
-

Míry založené na chí-kvadrátu

- velikost hodnoty chí-kvadrát je ovlivněna velikostí výběru a počtem kategorií tabulky
 - účelem koeficientů založených na chí-kvadrátu je eliminovat tyto vlivy
-

Míry založené na chí-kvadrátu

- rozsah koeficientů je obvykle mezi 0 a 1
 - čím vyšší hodnota, tím těsnější vztah
 - 0 – žádný vztah
 - 1 – absolutní vztah (z hodnot jedné proměnné můžeme předpovědět hodnoty druhé proměnné)
 - pro koeficienty je možno spočítat statistickou významnost
-

Míry založené na chí-kvadrátu

- mezi nejčastěji užívané míry asociace založené na chí-kvadrátu patří koeficienty
 - Fí (Phi)
 - Cramerovo V (Cramer's V)
 - někdy je užíván i koeficient kontingence (Contingency Coefficient)
-

Míry založené na chí-kvadrátu

- **Fí koeficient** - užívá se pro tabulky 2x2 (tj. pro dichotomické proměnné, např. pohlaví)
 - vypočte se tak, že se hodnota chí-kvadrátu vydělí počtem osob a výsledek se odmocní
 - $\Phi^2 = \chi^2/n$
-

Míry založené na chí-kvadrátu

- Cramerovo V – podobný výpočet jako F_i ; počet osob se navíc násobí (počtem řádků - 1)
 - (pokud je počet řádků menší než počet sloupců, jinak počtem sloupců - 1)
 - $V = \chi^2 / (n * m)$
 - používá se pro tabulky větší než 2x2
-

Příklad

- příklad z minulé přednášky - jak souvisí model manželství s jeho vydařeností
 - Chí-kvadrát = 18.71
 - počet osob $N = 154$
 - $m = (\text{počet řádků} - 1) = (3 - 1) = 2$
-

Kontingenční tabulka (SPSS)

model rodic. rodiny - muz * hodnoceni manzelstvi rodicu - muz Crosstabulation

Count

		hodnoceni manzelstvi rodicu - muz			Total
		vydarene	prumerne	nevydarene	
model rodic. rodiny - muz	matka dominance	22	29	18	69
	otec dominance	14	19	11	44
	kooperativnost	29	8	4	41
Total		65	56	33	154

Příklad

□ tabulka 3x3 – použijeme Cramerovo V

$$\square V = \sqrt{\chi^2 / (n * m)}$$

$$\square V = \sqrt{18.71 / (154 * 2)}$$

$$\square \mathbf{V = 0,246}$$

Příklad

- **interpretace:** hodnota 0,246 je poměrně nízká – vztah mezi modelem manželství a jeho vydařeností není příliš těsný
 - v SPSS jsou uvedeny oba koeficienty (F i V), je třeba zvolit ten správný pro každou tabulku
-

Výstup v SPSS

Symmetric Measures

		Value	Approx. Sig.
Nominal by	Phi	,349	,001
Nominal	Cramer's V	↗ ,246	,001
N of Valid Cases		154	

- a. Not assuming the null hypothesis.
 - b. Using the asymptotic standard error assuming the null hypothesis.
-

Míry PRE

- **PRE** je zkratka pro **Proportional Reduction in Error** (poměrná redukce chyby odhadu)
 - princip PRE: porovnání odhadu hodnot závislé proměnné bez znalosti hodnot nezávislé proměnné a s její znalostí (o kolik se sníží chyba odhadu?)
-

Míry PRE

- **příklad** – jaký je vztah mezi pohlavím a užíváním rtěnky?*
 - vypočítáme koeficient **lambda**
 - pokud bychom měli odhadnout, zda náhodně vybraný respondent používá rtěnku: jaká je pravděpodobnost chybného odhadu?
- *příklad převzat z Disman: Jak se vyrábí sociologická znalost
-

Míry PRE

- můžeme očekávat, že více lidí rtěnku nepoužívá než používá (naprostá většina mužů + některé ženy)
 - takže bude výhodnější odhadnout, že náhodně vybraný respondent rtěnku nepoužívá
 - pravděpodobnost chyby závisí na podílu lidí užívajících rtěnku
-

Míry PRE

RTĚNKA

	Frequency	Percent
Valid nepoužívá	97	60,6
používá	63	39,4
Total	160	100,0

Míry PRE

- při tomto podílu osob je pravděpodobnost chyby asi 40% (když budeme odhadovat, že náhodný respondent rtěnku neužívá)
 - ze 160 případů bychom se zmýlili 63x
-

Míry PRE

- o kolik by se chyba zmenšila, pokud bychom znali pohlaví respondenta?**
 - pro muže bychom odhadovali, že rtěnku nepoužívá, pro ženu naopak - že ji používá
-

Míry PRE

POHLAVÍ * RTĚNKA Crosstabulation

Count

	RTĚNKA		Total
	nepoužívá	používá	
POHLAVÍ muži	78	2	80
ženy	19	61	80
Total	97	63	160

Míry PRE

- pokud bychom znali pohlaví respondenta, zmýlili bychom se ve svém odhadu 21x (2 x u muže a 19x u ženy)
 - **o kolik by se náš odhad zlepšil? tj. o kolik by se zmenšila naše chybovost, oproti původní chybovosti?**
-

Míry PRE

- chyby předtím – chyby teď
= $63 - 21 = 42$
 - poměrná redukce chyby (tj. vzhledem k předchozím chybám) = **lambda** = $42/63 = \mathbf{0,667}$
 - **chyba v odhadu užívání rtěnky se sníží asi o 67%, pokud známe pohlaví respondenta**
-

Míry PRE

- rozsah koeficientu λ je od 0 do 1
 - **0** znamená, že znalost hodnoty nezávislé proměnné vůbec nesníží chybu v odhadu hodnot závislé proměnné; **proměnné jsou vzájemně nezávislé**
 - čím blíže **1**, tím lépe můžeme z hodnot nezávislé proměnné předpovědět hodnoty závislé proměnné
-


Míry PRE

- v SPSS jsou počítány 3 varianty koeficientu lambda
 - symetrická – není určeno, co je závislá a co nezávislá proměnná
 - 2 asymetrické – pro proměnnou 1 jako závislou a pro proměnnou 2 jako závislou
-

Výstup v SPSS

Directional Measures

			Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Nominal	Lambda	Symmetric	,706	,063	7,120	,000
by		POHLAVÍ Dependence	,738	,051	9,187	,000
Nominal		RTĚNKA Dependence	,667	,082	5,057	,000



Míry PRE pro nominální data

- kromě koeficientu lambda se užívají také
 - Goodmanovo a Kruskalovo **tau**
(nevyužívá při predikci nejčastější kategorii závislé proměnné jako lambda, ale rozdělení ve všech kategoriích závisle proměnné)
 - Cohenova **Kappa** – pro měření **shody dvou posuzovatelů**
-

Kontrolní otázky

- co vyjadřuje absolutní hodnota Pearsonova koeficientu korelace? a co jeho znaménko (+ nebo -)?
 - co je to koeficient determinace?
 - čím může být zkreslen korelační koeficient?
-

Kontrolní otázky

- rozdíl mezi mírami založenými na chí-kvadrátu a mírami PRE
 - nejužívanější míry pro nominální data
 - nejužívanější míry pro ordinální data
-

Literatura

- Hendl: kapitola 7
 - ukázka výsledků korelační analýzy (v IS):
 - Parker, J. D. A., Austin, E. J., Hogan, M. J., Wood, L. M., & Bond, B. J. (2005). Alexithymia and academic success: Examining the transition from high school to university. *Personality and Individual Differences, 38*, 1257-1267.
-