# Míry asociace

- obecná definice – síla a směr vztahu
- míry asociace pro nominální data
- míry asociace pro ordinální data
- korelace

# Míry asociace

- míry asociace vyjadřují **těsnost vztahu proměnných** (a případně **směr** vztahu)
- z chí-kvadrátu se dozvíme pouze, **zda nějaký vztah mezi proměnnými existuje** (tj. zda se liší četnosti pozorované a četnosti očekávané za předpokladu, že proměnné jsou nezávislé)

# 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
- **pozitivní vztah** – čím vyšší hodnoty jedné proměnné, tím vyšší hodnoty druhé proměnné
- **negativní vztah** - čím vyšší hodnoty jedné proměnné, tím nižší hodnoty druhé proměnné

# Míry asociace pro nominální data

- míry asociace pro nominální data <u>ukazují pouze sílu vztahu dvou proměnných</u>, nikoli směr či jiné informace o povaze vztahu

# Míry asociace pro nominální data

- 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

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

# Míry založené na chí-kvadrátu

- mezi nejčastěji užívané míry asociace založené na chí-kvadrátu patří koeficienty
  - Fí (Phi)
  - Cramerovo V (Cramer's V)

# 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

# Míry založené na chí-kvadrátu

- **Cramerovo V** – podobný výpočet jako Fí; počet osob se navíc násobí (počtem řádků – 1)
  - (pokud je počet řádků menší než počet sloupců, jinak počtem sloupců – 1)
- používá se pro tabulky větší než 2x2

# Příklad

- příklad z přednášky o Chí-kvadrátu - jak souvisí model manželství s jeho vydařeností
- Chí-kvadrát = 18.71
- počet osob N = 154
- m = (počet řádků – 1) = 3 – 1 = 2

# Kontingenční tabulka

| MODEL_M: model rodic. rodiny - muz | 2-rozměrná tabulka: Pozorované četnosti (family) RODICE_M vydarene | RODICE_M prumerne | RODICE_M nevydarene | Řádk. součty |
|---|---|---|---|---|
| matka dominance | 22 | 29 | 18 | 69 |
| Řádk. četn. | 31,88% | 42,03% | 26,09% | |
| otec dominance | 14 | 19 | 11 | 44 |
| Řádk. četn. | 31,82% | 43,18% | 25,00% | |
| kooperativnost | 29 | 8 | 4 | 41 |
| Řádk. četn. | 70,73% | 19,51% | 9,76% | |
| Celk. | 65 | 56 | 33 | 154 |

# Příklad

- tabulka 3x3 – použijeme Cramerovo V
- V = $\sqrt{\chi^2/(N*m)}$
- V = $\sqrt{18.71/(154*2)}$
- **V = 0,246**

# Příklad

- **interpretace:** hodnota 0,246 je poměrně nízká – vztah mezi modelem manželství a jeho vydařeností není příliš těsný (i když statisticky významný – viz výstup ve Statistice)

# Výstup ve Statistice

| Statist. | Statist. : MODEL_M(3) x RODICE_M(3) (family) | | |
|---|---|---|---|
| | Chí-kvadr. | sv | p |
| Pearsonův chí-kv. | 18,71174 | df=4 | p=,00090 |
| M-V chí-kvadr. | 18,83684 | df=4 | p=,00085 |
| Fí | ,3485754 | | |
| Kontingenční koeficient | ,3291517 | | |
| Cramér. V | ,2464800 | | |

# Další míry asociace

- **Cohenova kappa** – koeficient shody
- většinou používán pro měření shody mezi posuzovateli

# Další testy

- **McNemarův test** – pro závislé výběry (opakovaná měření)
- pro tabulky 2x2 – zachycuje míru změny (kolik osob z určité kategorie při prvním měření přejde při druhém měření do jiné kategorie)
- obdobný test pro více než dvě měření – **Cochranův test**

# Míry asociace pro ordinální data

- u ordinálních dat je výpočet založen na poměru tzv. souhlasných a nesouhlasných párů případů
- **souhlasný** pár případů – hodnota obou proměnných je u jednoho člena páru vyšší (nebo nižší) než u druhého
- **nesouhlasný** pár případů – hodnota jedné proměnné je u jednoho člena páru vyšší a hodnota druhé proměnné je nižší než u druhého člena páru

# Příklad

- souvisí spokojenost v manželství
  - 3 velmi šťastné, 2 spíše šťastné, 1 ne příliš šťastné
- s hodnocením života
  - jako 3 vzrušujícího, 2 stereotypního až 1 nudného?

# Příklad

| Případ | manželství | život |
| --- | --- | --- |
| Osoba 1 | 2 | 3 |
| Osoba 2 | 1 | 2 |
| Osoba 3 | 2 | 1 |

# Příklad

- PÁR 1: osoba 1 (2, 3) a osoba 2 (1, 2) - souhlasný
- PÁR 2: osoba 1 (2, 3) a osoba 3 (2, 1) – nerozhodně (tzv. tie)
- PÁR 3: osoba 2 (1, 2) a osoba 3 (2, 1) - nesouhlasný

# Míry asociace pro ordinální data

- koeficient gamma = počet souhlasných mínus počet nesouhlasných párů, tento rozdíl vzhledem k celkovému počtu párů (jen souhlasných a nesouhlasných)
- nerozhodné páry nebere gamma v úvahu

# Míry asociace pro ordinální data

- gamma = $\dfrac{\text{souhlasné - nesouhlasné}}{\text{souhlasné + nesouhlasné}}$

- gamma = (1-1)/2 = **0**

# Míry asociace pro ordinální data

- pokud je většina párů souhlasných, je hodnota gamma kladná – tj. **pozitivní vztah** (až +1)
- pokud je většina párů nesouhlasných, je hodnota gamma záporná – tj. **negativní vztah** (až -1)
- pokud je počet souhlasných a nesouhlasných párů vyrovnán – gamma kolem 0

# Míry asociace pro ordinální data

- gamma je symetrická míra – nedělá rozdíly mezi závislou a nezávislou proměnnou
- asymetrická varianta koeficientu gamma – **Sommerovo D**
- **Kendallovo tau b**– stejný výpočet jako gamma, ale bere v úvahu i nerozhodné páry (tzv. ties); hodnoty v rozsahu -1 až +1 mohou být získány pouze pro čtvercové tabulky (tj. stejný počet kategorií obou proměnných)
- **Kendallovo tau c**– kromě korekce pro ties obsahuje i korekci pro velikost tabulky
- **Spearmanův koeficient korelace** (viz dále)

# Shrnutí

- u nominálních dat hodnota míry asociace proměnných indikuje sílu vztahu – rozsah od 0 do 1
- u ordinálních dat míry asociace indikují jak sílu vztahu (abs. hodnota koeficientu), tak směr vztahu

# Pearsonův korelační koeficient

- u intervalových a poměrových 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 (těsnost) vztahu**
- znaménko (+ nebo -) **směr vztahu**
- **rozsah -1 až +1**
- označuje se **r**

# Pearsonův korelační koeficient

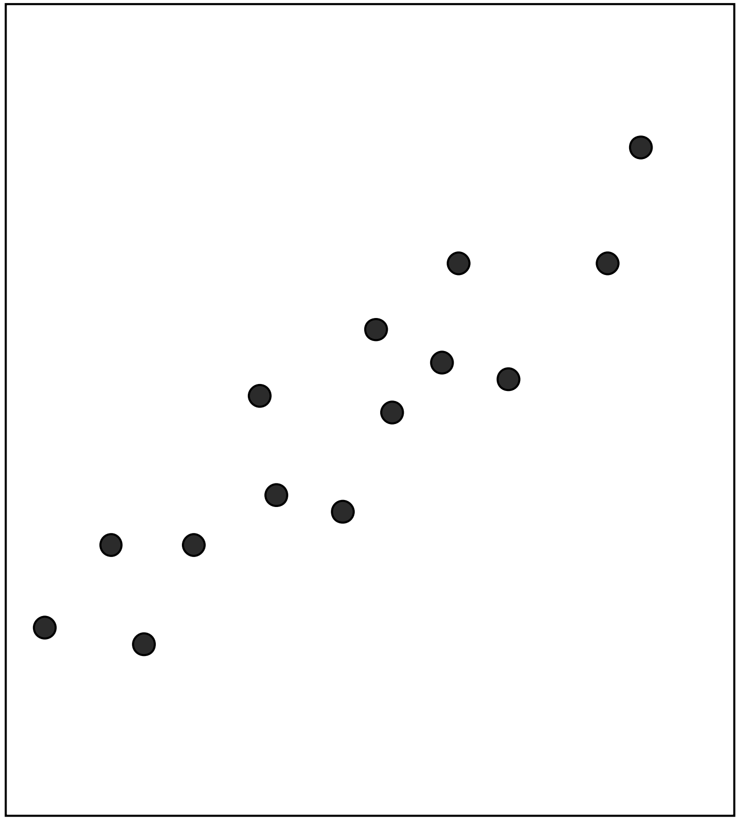
- sám o sobě je deskriptivní statistikou, ale podobně jako u ostatních měr asociace je možno spočíst **statistickou významnost (=zda se se významně liší od nuly**, tj. zda nějaký vztah mezi proměnnými vůbec existuje)
- závisí na velikosti výběru – čím vyšší, tím nižší koeficient vychází průkazný

# Pearsonův korelační koeficient

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

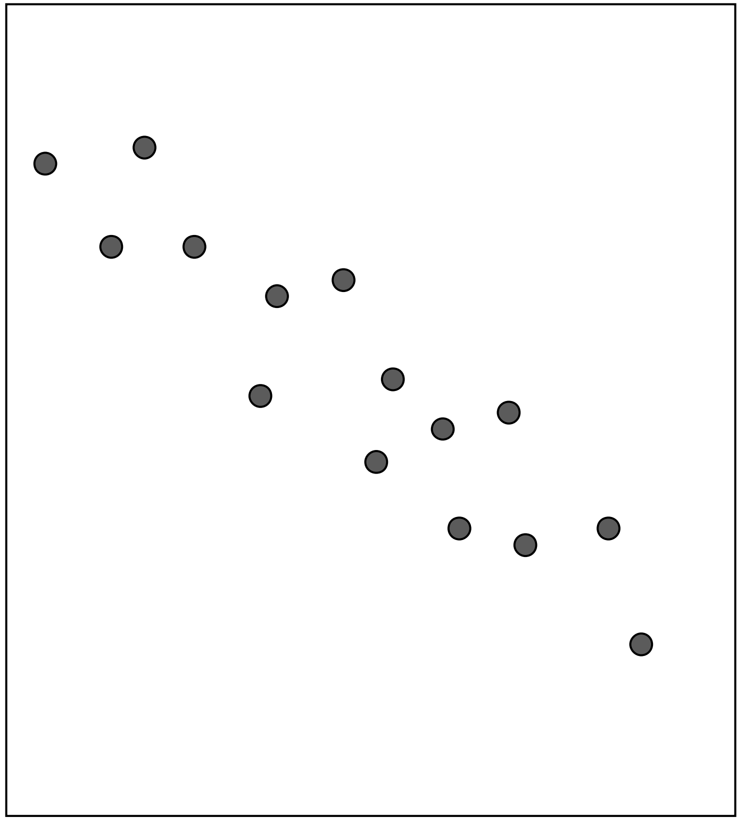
# Scatter

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



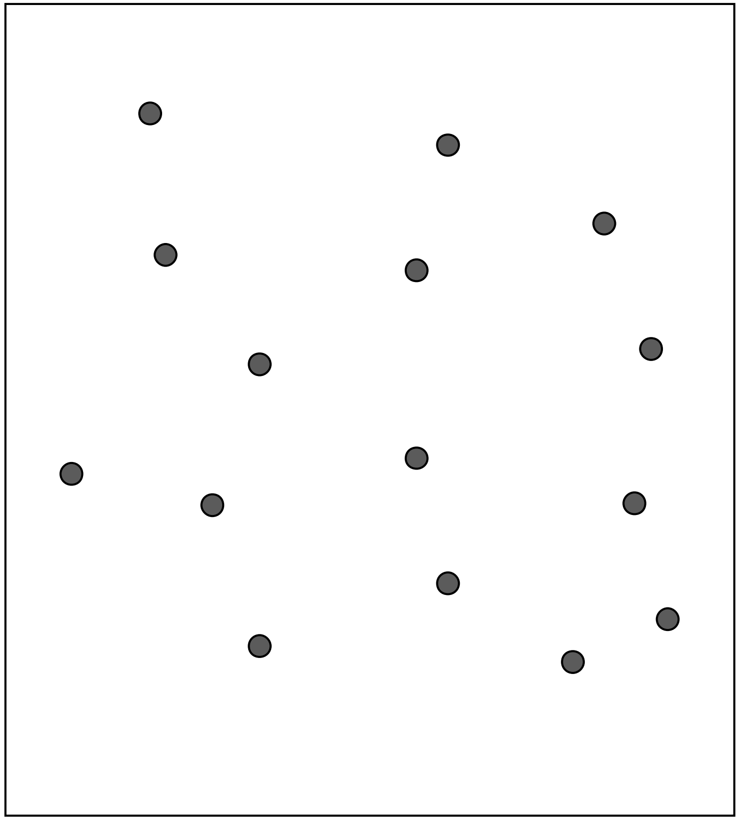
# Scatter

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



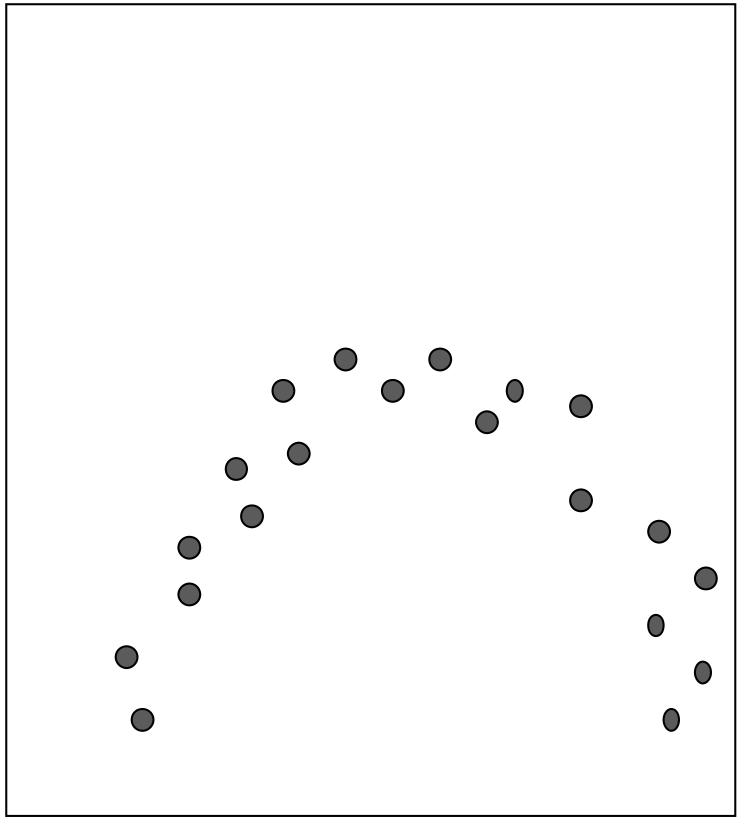
# Scatter

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



# Scatter

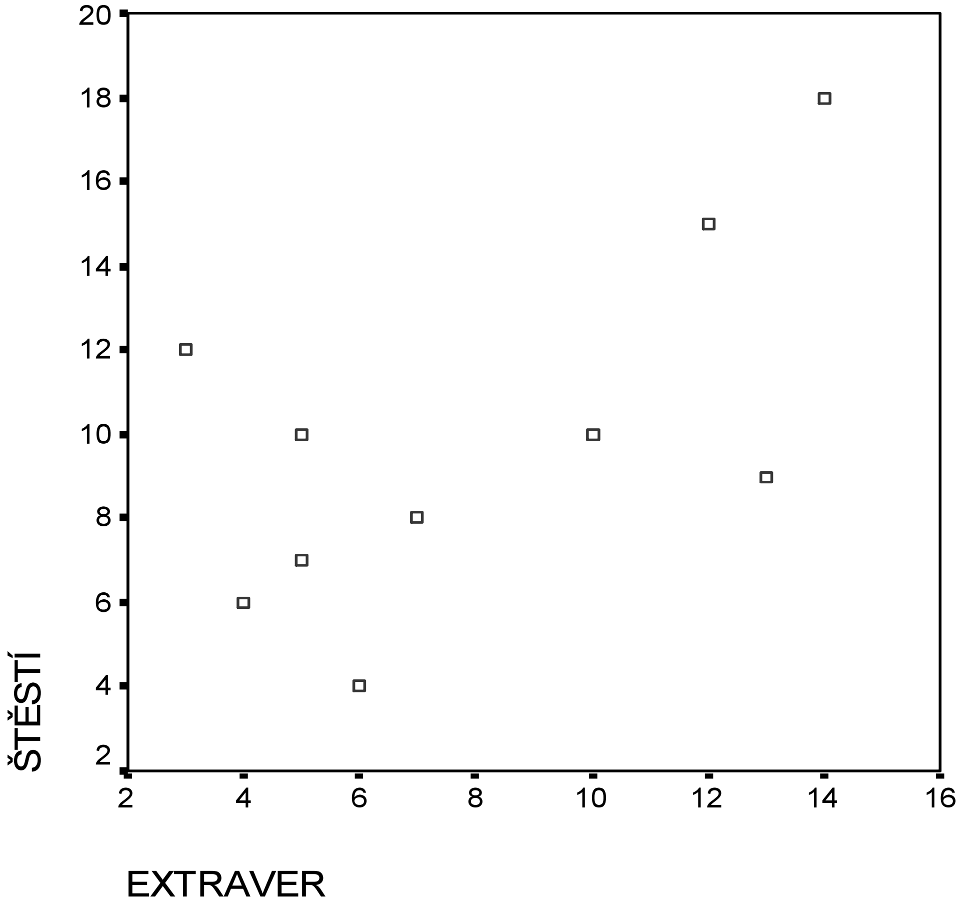
- **nelineární vztah**
- r = 0



# Příklad

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

# Příklad



# Příklad

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **štěstí** | **15** | **8** | **7** | **18** | **4** | **12** | **10** | **10** | **6** | **9** |
| **extraverze** | **12** | **7** | **5** | **14** | **6** | **3** | **5** | **10** | **4** | **13** |

# Příklad

- **výpočet r**
- $r_{xy} = s_{xy}/s_x s_y$

- $s_{xy}$ **= kovariance**
- $s_{xy} = (\Sigma^{n}{}_{i=1}\,(x_i - \bar{x})(y_i - \bar{y}))/(n-1)$

- $s_x$, $s_y$ **= směrodatné odchylky**

# Příklad

- $\overline{x} = m_x = 9{,}90$
- $s_x = 4{,}20$
- $\overline{y} = m_y = 7{,}90$
- $s_y = 4{,}01$
- $n = 10$

# Příklad

- $\mathbf{s_{xy} = (\Sigma^{n}_{i=1} (x_i-\overline{x})(y_i-\overline{y}))/(n-1)}$

- $s_{xy}$= ($\Sigma^{n}_{i=1}$ ($x_i$-9,9)($y_i$-7,9)/(10-1)
- $s_{xy}$= ( (15-9,9)(12-7,9) + (8-9,9)(7-7,9) + (7-9,9)(5-7,9) + (18-9,9)(14-7,9) + (4-9,9)(6-7,9) + (12-9,9)(3-7,9) + (10-9,9)(5-7,9) + (10-9,9)(10-7,9) + (6-9,9)(4-7,9) + (9-9,9)(13-7,9) )/9
- $s_{xy}$= ((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
- $s_{xy}$= (20,91+1,71+8,41+49,41+11,21+(-10,29)+(-0,29)+0,21+15,21+(-4,59))/9
- $s_{xy}$= 91,9/9
- **$s_{xy}$= 10,21**

# Příklad

- $r_{xy} = s_{xy}/s_x s_y$

- $r_{xy} = 10,21/4,20*4,01$
- $r_{xy} = 10,21/16,84$
- $r_{xy} = 0,606$

# Výstup ve Statistice

| Proměnná | Korelace (data příklad přednáška 2)<br>Označ. korelace jsou významné na hlad. p < ,0500(<br>N=10 (Celé případy vynechány u ChD) | |
|---|---|---|
| | štěstí | extraverze |
| štěstí | 1,00 | 0,61 |
| extraverze | 0,61 | 1,00 |

# 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 je možné převést koeficient korelace na **koeficient determinace ($r^2$)**
- ukazuje, 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 např. experimentem, kdy jsou všechny ostatní proměnné udržovány konstantní, proměnná X předchází Y v čase atd.

# Faktory ovlivňující r

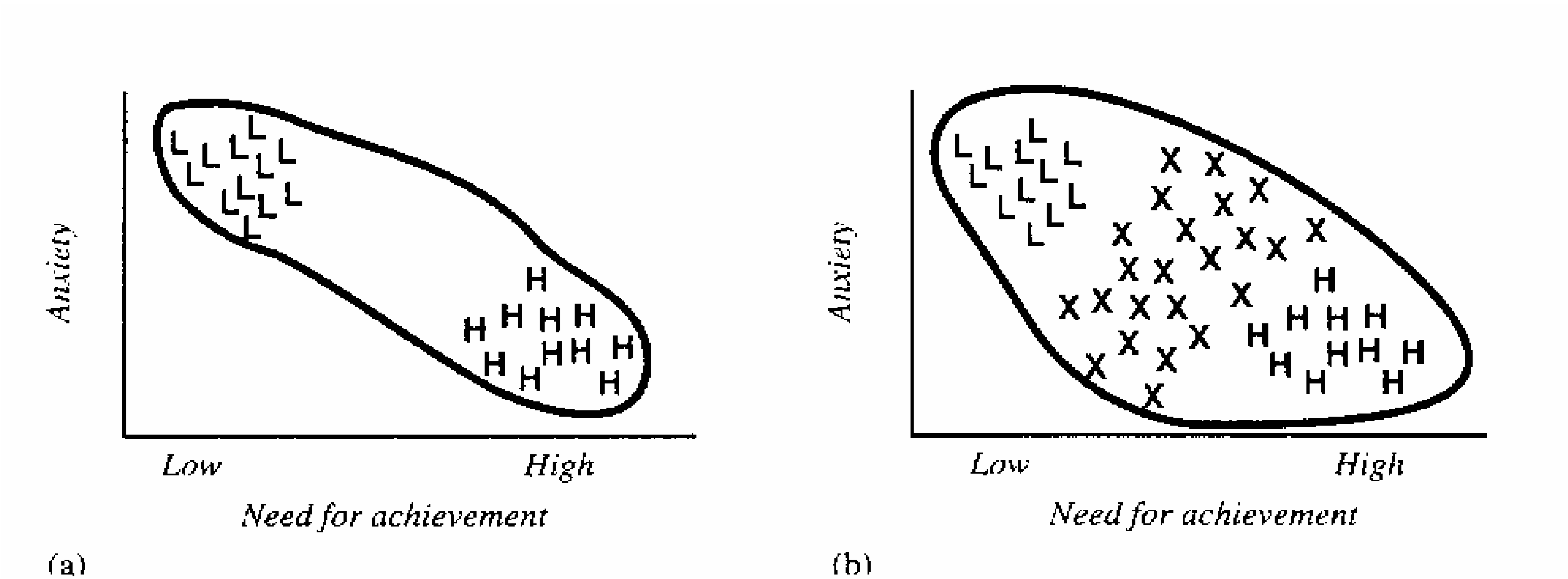
- omezený rozsah hodnot proměnné
- použití extrémních skupin
- nehomogenní soubor
- extrémní hodnoty (outliers)
- nelineární vztahy
- reliabilita použitých nástrojů

# Omezený rozsah hodnot

- omezený rozsah hodnot jedné nebo obou proměnných snižuje hodnotu r
- stejně tak nízká variabilita (extrémní případ:pokud by všechny hodnoty 1 proměnné byly stejné, zákonitě r=0)

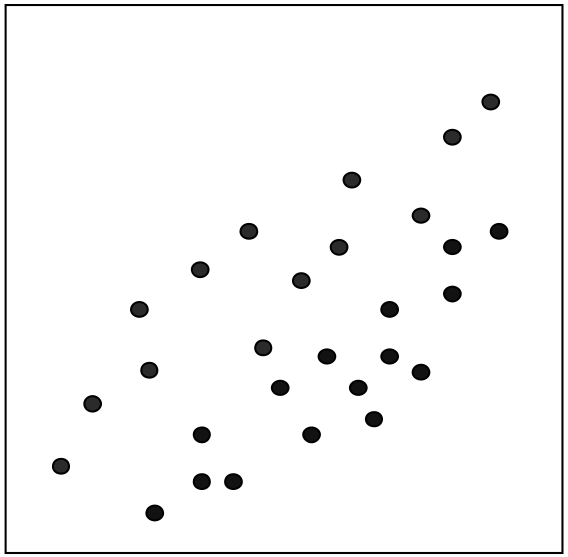
# Použití extrémních skupin

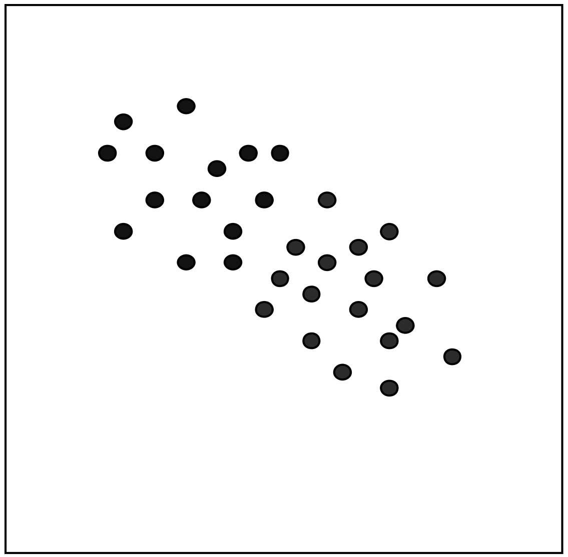
- použití extrémních skupin (např. jen osob s vysokým IQ) vede k vyššímu r



# Nehomogenní soubor

- může zkreslit r jak směrem nahoru, tak dolů



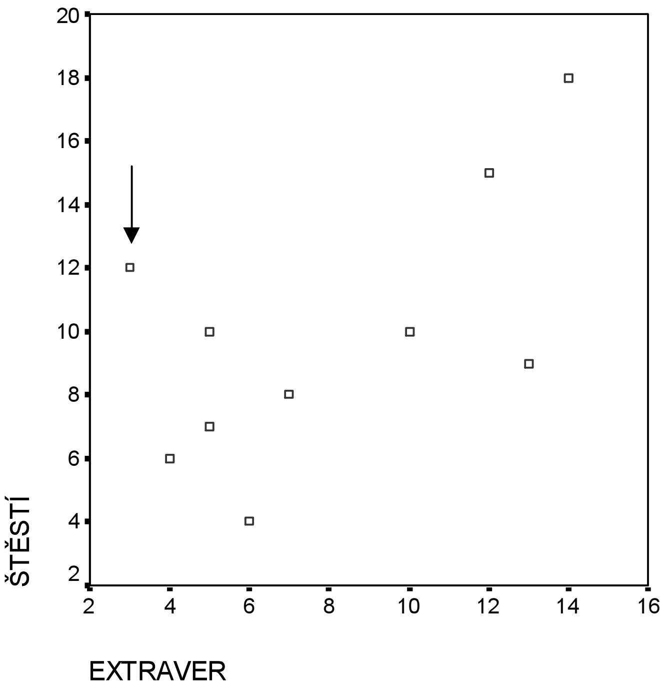


# Extrémní hodnoty

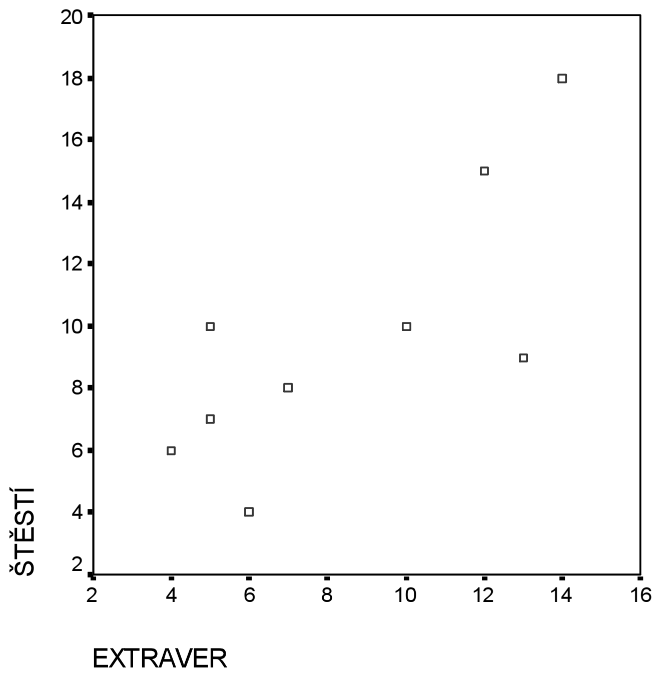
- extrémní hodnoty v jedné nebo obou proměnných mohou r výrazně zkreslit (nejen hodnotu, ale i směr), zvláště když je počet osob v souboru nízký

# Extrémní hodnoty

- r= 0,606




- r= 0,766



# Neparametrický koeficient

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

# Literatura

- **Hendl: kapitoly 8.3, 7.1 a 7.2**