MUNI FSS Katedra psychologie

Přednáška 12:
Shoda posuzovatelů

28. 11. 2022 | PSYn4790 | **Psychometrika: Měření v psychologii**

Katedra psychologie, Fakulta sociálních studií MU

Hynek Cígler | hynek.cigler@mail.muni.cz

Posuzování/hodnocení v psychologii

Posuzovací škály

- Intenzita prožitků, příznaků nemoci, ...

Pozorování a observační studie

- Bylo / nebylo pozorováno nějaké chování? Do jaké kategorie zařadit to, co jsem pozoroval(a)?

Psychologická diagnostika

- Diagnostický nález, skóry z checklistu, ...

Hodnocení výkonu

- V rámci školní třídy, v testu, pořadí uchazečů při náboru zaměstnanců, ...

Kódování v kvalitativním výzkumu

... napadne vás ještě nějaký příklad?

Posuzování/hodnocení v psychologii

Inter-rater reliability.

Inter-rater agreement.

Vřelost

1 – Výrazný nedostatek lásky

- Takto jsou hodnoceni rodiče respondenta, kteří nejenže nebyli oporou jeden druhému, ale odmítali vzájemně spolupracovat nebo spolu soupeřili, nechovali se k sobě nikterak láskyplně či ohleduplně. Takto se posuzují vztahy charakteristické přítomností hněvu a nepřátelských projevů nebo vztahy, v nichž se rodiče k sobě chovali chladně a nezúčastněně. Toto hodnocení se využívá také v případech, kdy jeden z rodičů druhého psychicky či fyzicky týral či zneužíval. Manželství, která byla ukončena rozvodem, se hodnotí v rozmezí bodů 1–3.

3 – Nedostatek vřelosti

- Vztah se vyznačuje mírnou, nicméně neadekvátní nebo nekonzistentní oporou. Potřeby jednoho nebo obou rodičů bývají občas uspokojeny, většinou jsou však přehlíženy. Tyto páry se vyznačují vzájemnou lhostejností, každý z partnerů žil víceméně vlastním životem, které se prolínaly pouze sporadicky. Toto hodnocení se užívá i pro páry, které spolu sice žily aktivně, ale jejich vzájemná interakce byla charakterizována spíše negativně, jednali spolu například s neúctou a s nedostatečným poskytováním opory.

5 – Ani neláskyplný, ani aktivně láskyplný

- Respondent hodnotí vztah svých rodičů jako „dobrý“ či „láskyplný“, ale neuvádí detaily, které by tento pohled potvrdily či vyvrátily. Pokud je k dispozici více detailů, lze říci, že rodiče poskytovali adekvátní emocionální oporu jeden druhému. Přestože nijak výrazně nerozuměli potřebám toho druhého, snažili se být si ve většině oblastí soužití nápomocni.
- Někteří respondenti se mohou při popisu soustředit na dovednosti rodičů v oblasti výchovy, a výzkumník/tazatel tak získává dojem, že manželství rodičů hrálo sekundární roli oproti výchově dětí, která byla pro pár prvořadá. Toto hodnocení také slouží jako průměrné hodnocení, pokud se manželé v minulosti nechovali k sobě láskyplně, ale tyto negativní epizody byly ve vztahu vystřídány či vynahrazeny věrohodnými láskyplnými či obětavými činy.

7 – Láskyplný

- Přestože se ve vztahu mohly objevovat problémy, rodiče se vůči sobě projevovali láskyplným a chápajícím způsobem. Lze vytušit, že vztah byl plný důvěry a opory. Hodnocení 7 je odpovídající, pokud respondent souvisle a srdečně hovoří o vztahu rodičů a udává, že se k sobě pár choval s láskou, ale současně to dokládá menším množstvím specifických detailů.

9 – Velmi láskyplný

- Tito rodiče se k sobě aktivně chovali láskyplně a s vzájemnou náklonností a očividně se cítili dobře a užívali si vzájemnou společnost. Respondent uvádí konkrétní příklady, jak si byli jeho rodiče oporou sobě navzájem, partnersky, tak svým dětem jako rodiče. Poskytovali si navzájem přátelství a útěchu. Není nutné, aby byl vztah popisován jako absolutně perfektní, pro toto hodnocení se rozhodujeme tehdy, existují-li silné důkazy, že se rodiče navzájem milovali, respektovali a podporovali jeden druhého.

Proč se zabývat shodou?

Kdo může zaručit „objektivitu“ posuzování/hodnocení?

- I pokud jsou hodnotící kritéria jasně definována, jsou stejně chápána a používána?

Ověření reliability výzkumné/diagnostické metody.

- **Hodnocení** na posuzovacích škálách, pozorování chování, hodnocení výkonu
- **Administrace** diagnostických metod – vliv administrátora

Zajištění **interní validity** výzkumných designů.

- Shoda posuzovatelů, pozorovacích schémat atp.

Inter-rater/inter-coder/inter-observer...
... reliability/agreement/concordance.

Co dělat s (ne)shodou?

Shodu můžeme „**vynutit**“ (např. použít průměrné hodnocení)

- Tím se ale připravujeme o informace.

... nebo ji můžeme nějak **kvantifikovat** a vyjádřit její míru.

- Míra (ne)shody je důležitý a interpretovatelný údaj.

Po kvantifikaci můžeme (ne)shodu efektivněji **studovat**

- Jak velké jsou mezi hodnotiteli rozdíly?
- Jsou tyto rozdíly náhodné?
- Jsou tyto rozdíly systematické (např. rozdílně „přísní“ hodnotitelé)?

Není to buď a nebo (např. prvně kvantifikuji, pak shodu „vynutím“).

Dvě hlavní použití míry (ne)shody[1]

Lze několik různých hodnocení „redukovat“ na jediný údaj?

- Kolik spolu mají hodnocení „společného“, jde stále o tu stejnou proměnnou?

Jaká je reliabilita takovéto redukce v případě...

- ... průměrného/výsledného hodnocení několika hodnotiteli?
- ... hodnocení jedním hodnotitelem?

[1] dle Cíglera a Širůčka, nejde o autoritativní zdroj

Proč je o tom samostatná přednáška?

1. Typicky zobecňujeme na všechny potenciální hodnotitele.
   - Tedy „absolutní D-studie“ z pohledu GT, zatímco běžné uvažování o „reliabilitě“ je zpravidla „relativní D-studie“.
2. Velmi často nominální nebo ordinální proměnné.
3. Přítomné i v kvalitativním výzkumu.

Dva hlavní typy neshody

1. Nesystematický rozdíl mezi hodnotiteli.
   - Náhodný rozdíl, neshoda „v pořadí“.
   - Z hlediska GT σ_{pr}^2
2. Systematický rozdíl mezi hodnotiteli.
   - Rozdíl v poměru, průměru... neshoda „v náročnosti/přísnosti“.
   - Z hlediska GT σ_r^2

... zpravidla ale pozorujeme kombinaci obou typů.

- Z hlediska GT $\sigma_r^2 + \sigma_{pr}^2$

Nominální proměnné

Při náboru do armády posuzují dva psychologové, jestli se rekruti hodí spíš na pilota (P) nebo na tankistu (T).

Systematický rozdíl

- Rozdíl v poměru P:T.
- Jeden z psychologů může dávat více závěrů „pilot“ než druhý.

Nesystematický rozdíl

- Oba psychologové mají tento poměr stejný, ale neshodnou se v určitém procentu % případů.

Nominální proměnné

SYSTEMATICKÁ
(κ = 0,6, shoda 80 %)

	B		
A	30	0	**30**
	20	50	**70**
	50	**50**	**100**

NESYSTEMATICKÁ
(κ = 0,6, shoda 80 %)

	B		
A	40	10	**50**
	10	40	**50**
	50	**50**	**100**

SMÍŠENÁ
(κ = 0,4, shoda 70 %)

	B		
A	25	5	**30**
	25	45	**70**
	50	**50**	**100**

(Alespoň) ordinální proměnné

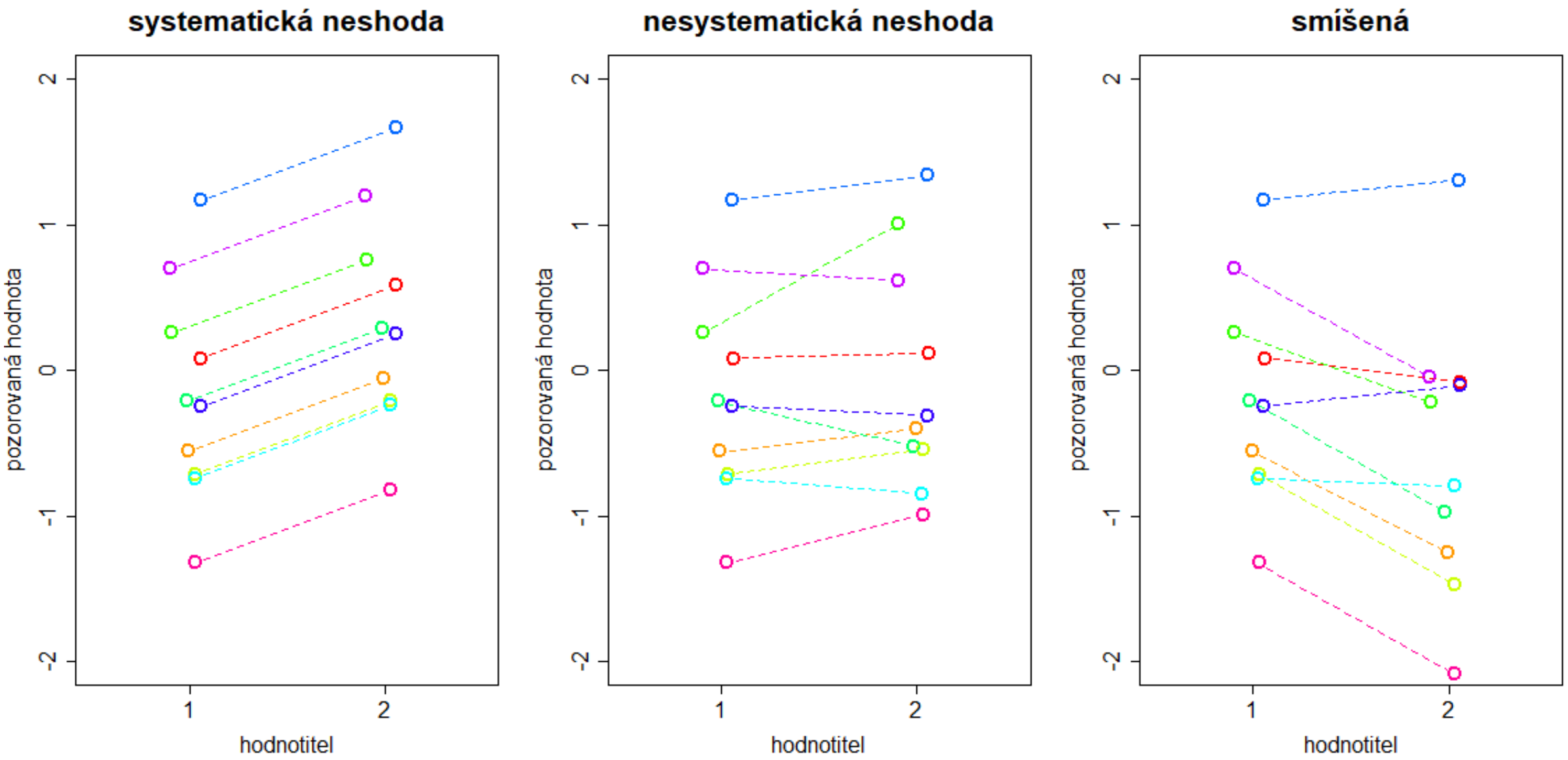
Během náboru zaměstnanců mají dva psychologové za úkol obodovat každého uchazeče na stupnici 1–3 (přijat, náhradník, nepřijat).

Systematický rozdíl

- Rozdíl v průměru.
- Jeden z psychologů je „přísnější“ a hodnotí každého méně body.

Nesystematický rozdíl

- Oba psychologové se neshodnou na tom, kdo je nejlepší, kdo druhý nejlepší, třetí nejlepší, atd.


systematická neshoda
nesystematická neshoda
smíšená
pozorovaná hodnota
hodnotitel
1
2
-2
-1
0
1
2

Jaké otázky si klást?

Kdo se má shodovat s kým?

- **Shoda administrátorů:** Vede individuální vyšetření různými administrátory ke stejným výsledkům? (WISC...)
- **Shoda hodnotitelů:** Ohodnotí již získaný protokol různí lidé stejně? (ROR; kvalitativní výzkum).
- **Intra-rater reliabilita:** Obdobné otázky, ale pro jednoho administrátora/hodnotitele v různých časech.

Kolik bylo hodnotitelů?

- Dva (a nebo jeden dvakrát).
- Tři a více (nebo jeden alespoň třikrát).

Typy proměnných a související hypotézy

Nominální nebo ordinální

- Jaká je absolutní/relativní míra shody 2 nebo více osob?

Ordinální

- Jaká je míra shody v pořadí hodnocených osob?
- Jaká je míra shody ve střední hodnotě?
- Celková míra shody (pořadí i střední hodnota dohromady).
- Absolutní míra shody (jako by šlo o nominální proměnnou).

Intervalová/poměrová

- Jaká je míra shody v pořadí hodnocených osob?
- Jaká je míra shody ve střední hodnotě?
- Celková shoda (pořadí i střední hodnoty dohromady).

- V psychologické diagnostice je typickým postupem ověření shody v případě položek nominálními či ordinálními statistikami (analogie korigovaných korelací se škálou) a pro celkové skóry intervalovými statistikami.

Statistiky pro odhad shody posuzovatelů

Nominální proměnné

Jakým nejjednodušším způsobem lze vyjádřit shodu nominálních proměnných?

Ve které z tabulek je shoda vyšší?

Srovnání tabulek:

- shoda nahoře: 92 %;
- shoda dole: 92 %.

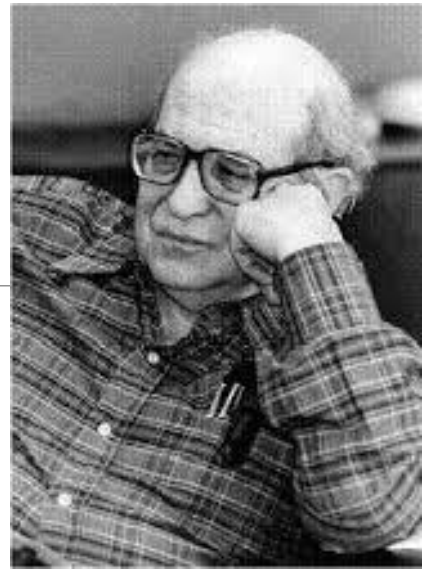
Procenta nejsou vypovídajícím ukazatelem shody hodnotitelů!

- Masivní vliv prevalence daného jevu.

	0	1	SUM
0	42	4	46
1	4	50	54
SUM	46	54	100

	0	1	SUM
0	0	3	3
1	6	92	98
SUM	6	95	100

Nominální proměnné ($n = 2$)



Cohenovo kappa

- Kolikrát je shoda hodnotitelů vyšší než náhodná shoda?

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

- P_o = pozorovaná shoda hodnocení
- P_e = shoda hodnocení očekávaná na základě prosté náhody

Nominální proměnné (n = 2)

	0	1	SUM
0	35	13	48
1	3	49	52
SUM	38	62	100

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Pozorovaná shoda hodnocení:

- $P_o = \frac{35+49}{100} = 0{,}84$

Očekávaná shoda hodnocení na základě náhody:

- $P_e = \left(\frac{35+3}{100} \cdot \frac{35+13}{100}\right) + \left(\frac{13+49}{100} \cdot \frac{3+49}{100}\right) = 0{,}505$
- V případě, že by oba odpovídali zcela nezávisle na sobě, shodli by se v 50,5 % případů.

Kohenovo kappa: $\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{0{,}84 - 0{,}505}{1 - 0{,}505} = 0{,}677$

Kritika za příliš silnou penalizaci P_e (Grant et al., 2017).

Nominální proměnné (n = 2)

Cohenovo kappa

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{0{,}84 - 0{,}505}{1 - 0{,}505} = 0{,}677$$

Interpretace: Podíl nárůstu shody oproti náhodné shodě je 0,68 % maximálního možného nárůstu.

Cohenovo kappa nabývá hodnot mezi -1 a 1.

- Interpretace vzdáleně podobná korelaci, ale měřítko je zcela jiné.
- Více k interpretaci: Warrens, M. J. (2015). Five Ways to Look at Cohen's Kappa. *Journal of Psychology & Psychotherapy, 5*(4). https://doi.org/10.4172/2161-0487.1000197

Proč není dobré používat procentuální shodu?

	0	1	SUM
0	42	4	46
1	4	50	54
SUM	46	54	100

$$P_o = 0{,}920$$
$$P_e = 0{,}503$$
$$\kappa = 0{,}839$$

	0	1	SUM
0	1	4	5
1	5	91	95
SUM	5	95	100

$$P_o = 0{,}920$$
$$P_e = 0{,}905$$
$$\kappa = 0{,}158$$

Použití % shody je téměř vždy špatně. Zpravidla **nadhodnocuje** skutečnou míru shody!

Nominální proměnné (n > 2)

Cohenovo kappa je určeno jen pro dva hodnotitele.

Pro *n* hodnotitelů je zobecněním Fleissovo kappa.

Stejná logika a interpretace, pouze složitější výpočet.

- Jednoduše jen multidimenzionální kontingenční tabulka.
- Může být výpočetně náročnější; důležitá je volba efektivního algoritmu.

Ordinální proměnné

Lze do jisté míry použít běžné statistiky, které už znáte:

Shoda středních hodnot (přísnost hodnotitelů):

- 2 hodnotitelé: Mann-Whitney („neparametrický t-test“).
- N hodnotitelů: Kruskal-Wallis („neparametrická ANOVA“).

Shoda pořadí:

- 2 hodnotitelé: Běžná pořadová korelace (Spearman, Kendall) pro shodu pořadí.
- N hodnotitelů: Kendallův koeficient konkordance (W) – viz dále

...ale máme k dispozici lepší nástroje ☺

Ordinální proměnné (n=2)

Můžeme k nim přistupovat jako k nominálním proměnným, ale výsledkem je obvykle podhodnocení shody

Řešením je **vážená Cohenova kappa (weighted kappa)**.

Neshody jsou váženy různým způsobem – čím dále od diagonály, tím jde o větší neshodu

- Jak vážit? Více možností
- lineární váhy: vzdálenost od diagonály
- kvadratické váhy: (vzdálenost od diagonály)^2
- vlastní váhy dle účelu

shoda		hodnotitel A		
		1	2	3
hodnotitel B	1	**15**	**12**	**1**
	2	**9**	**23**	**5**
	3	**0**	**8**	**17**

Ordinální proměnné (n=2)

Běžná (kategorická) kappa: $\kappa = 0{,}401$.

Ordinální kappa (lineární váhy): $\kappa_{wlin} = 0{,}502$.

Ordinální kappa (kvadratické váhy): $\kappa_{wquad} = 0{,}620$.

- Asi nejčastější případ.
- Vzdálenost je v řádku i sloupci... proto na druhou.

Matice vah ale může být libovolná.

- Např. i stejné váhy pro různá pole.

shoda		hodnotitel A		
		1	2	3
hodnotitel B	1	**15**	**12**	**1**
	2	**9**	**23**	**5**
	3	**0**	**8**	**17**

lineární váhy		hodnotitel A		
		1	2	3
hodnotitel B	1	**0**	**1**	**2**
	2	**1**	**0**	**1**
	3	**2**	**1**	**0**

kvadr. váhy		hodnotitel A		
		1	2	3
hodnotitel B	1	**0**	**1**	**4**
	2	**1**	**0**	**1**
	3	**4**	**1**	**0**

Ordinální proměnné (n>2)

Vážená Fleissova kappa

- Kombinace Fleissovy kappy a vážené Cohenovy kappy
- Bere v potaz shodu pořadí i středních hodnot

Shoda pořadí: **Kendallův koeficient konkordance (W)**

- Odpovídá na otázku, nakolik hodnotitelé udávají stejné pořadí.
- Analogie Spearmanovy pořadové korelace pro více hodnotitelů.
  - $W = \frac{\bar{\rho}(k-1)+1}{k}$, kde $\bar{\rho}$ je průměrná Spearmanova korelace napříč všemi páry hodnotitelů a k je počet hodnotitelů.
  - Reálně se používá trochu jiný, efektivnější výpočet.

Intervalové proměnné

Opět lze do jisté míry použít běžné statistiky.

Shoda průměrů (přísnost hodnotitelů):

- 2 hodnotitelé: t-test
- N hodnotitelů: one-way ANOVA

Shoda pořadí:

- 2 hodnotitelé: Pearsonova korelace
- N hodnotitelů: Cronbachova alfa

... ale máme k dispozici lepší nástroje ☺ (ano, už zase...)

Intervalové proměnné

Teorie zobecnitelnosti ☺

Pro zjednodušení jsou definovány 2×3 základní typy intra-class korelací, které jsou konkrétními speciálními případy teorie zobecnitelnosti.

- Historicky ale starší přístup předcházející GT (Fisher, zřejmě 1925).

Intra-class korelace: Jak moc se podobají hodnoty v rámci stejných tříd?

- Vnitrotřídní korelace.

Inter-class korelace: Jak moc se podobají hodnoty napříč třídami?

- Příkladem je Pearsonova korelace.

Třídou je myšlen subjekt pozorování (typicky respondent).

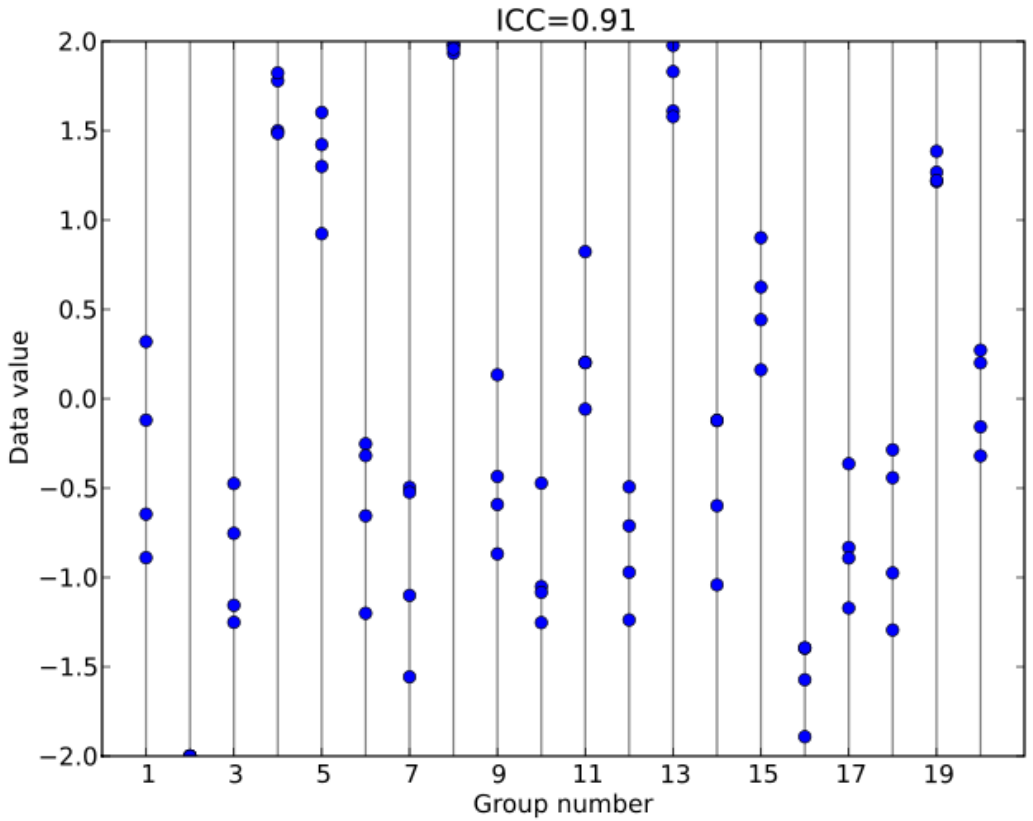
Intra-class / vnitrotřídní korelace (ICC)

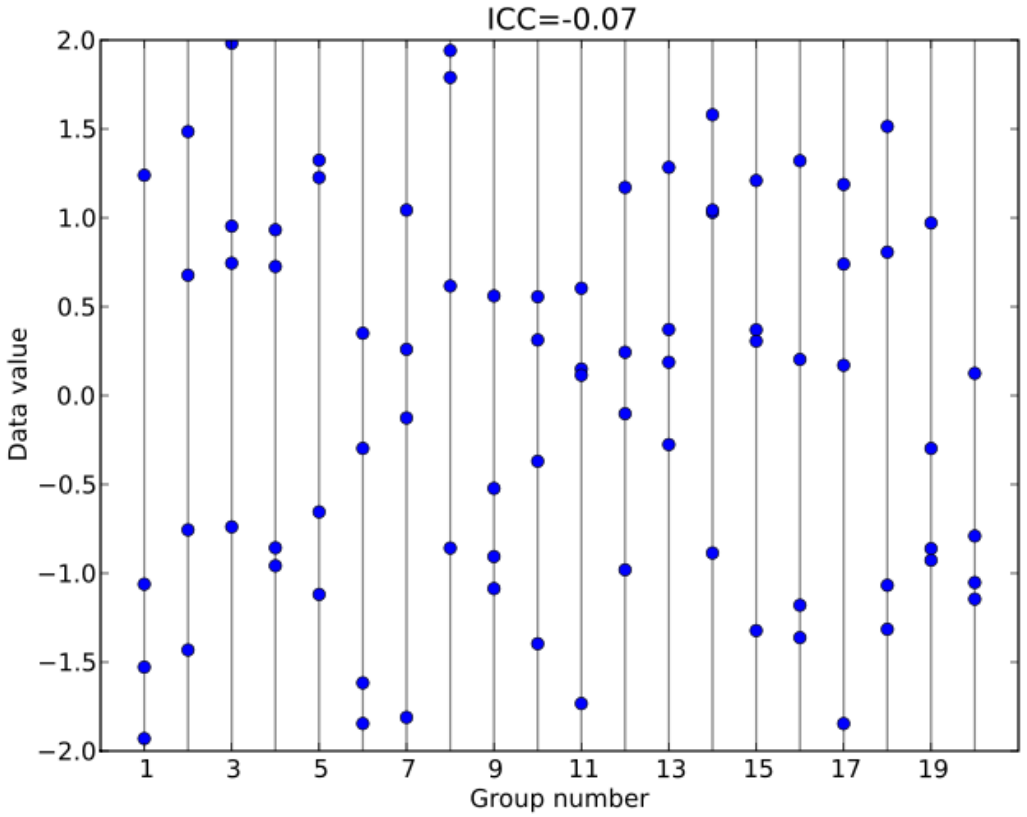
One key difference between the two statistics is that ***in the ICC, the data are centered and scaled using a pooled mean and standard deviation, whereas in the Pearson correlation, each variable is centered and scaled by its own mean and standard deviation.*** *This pooled scaling for the ICC makes sense because all measurements are of the same quantity (albeit on units in different groups).*

For example, in a paired data set where each "pair" is ***a single measurement made for each of two units*** *(e.g., weighing each twin in a pair of identical twins) rather than two different measurements for a single unit (e.g., measuring height and weight for each individual), the ICC is a more natural measure of association than Pearson's correlation.*

Popis originální definice ICC podle Fishera (Wikipedie)

Intra-class / vnitrotřídní korelace





https://en.wikipedia.org/wiki/Intraclass_correlation

Intra-class / vnitrotřídní korelace (ICC)

PEARSONOVA KORELACE

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad \text{(Eq.2)}$$

where:

σ_Y and σ_X are defined as above

μ_X is the mean of X

μ_Y is the mean of Y

$\mathbb{E}$ is the expectation.

VNITROTŘÍDNÍ KORELACE

$$r = \frac{1}{Ns^2} \sum_{n=1}^{N} (x_{n,1} - \bar{x})(x_{n,2} - \bar{x}),$$

where

$$\bar{x} = \frac{1}{2N} \sum_{n=1}^{N} (x_{n,1} + x_{n,2}),$$

$$s^2 = \frac{1}{2N} \left\{ \sum_{n=1}^{N} (x_{n,1} - \bar{x})^2 + \sum_{n=1}^{N} (x_{n,2} - \bar{x})^2 \right\}$$

Intra-class / vnitrotřídní korelace

Dva krát tři typy / modely (proč modely?) podle Shrouta a Fleisse (1979):

ICC1: každý subjekt je hodnocen stejným počtem **různých náhodných** hodnotitelů, kteří jsou ale **pokaždé jiní**.

- Hodnotitelé jsou striktně paralelními a pro každé měření znovu a náhodně losovanými testy.

ICC2: každý subjekt je hodnocena **stejnými náhodnými** hodnotiteli, ti jsou **pokaždé stejní**.

- Zobecňujeme na všechny hodnotitele, absolutní D-studie.
- Typicky je tohle to, co chcete.

ICC3: každý subjekt je hodnocen **stejnými nenáhodných** hodnotiteli.

- Zobecňujeme pouze na daný vzorek hodnotitelů, relativní D-studie.

Doporučuji (včetně SPSS notace): https://en.wikipedia.org/wiki/Intraclass_correlation

Intra-class / vnitrotřídní korelace

Tyto tři modely se dále dělí podle toho, jestli reálně dochází k:

- Udělení jednoho hodnocení jedním hodnotitelem: ICC(x, 1)
  - Reliabilita jednoho posuzovatele.
- Udělení průměrného hodnocení od všech hodnotitelů: ICC(x, k).
  - Kde k je počet hodnotitelů; například ICC(2, 3) pro ICC II. typu a 3 hodnotitele.
  - Reliabilita průměru posuzovatelů.

ICC(3, k) je shodná s Cronbachovou alfou.

- Relativní D-studie napříč všemi položkami, které jsou „fixed“.

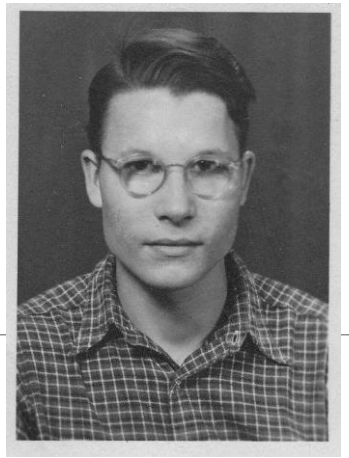
Odhad s pomocí ANOVA nebo smíšeného (mixed) lineárního modelu.

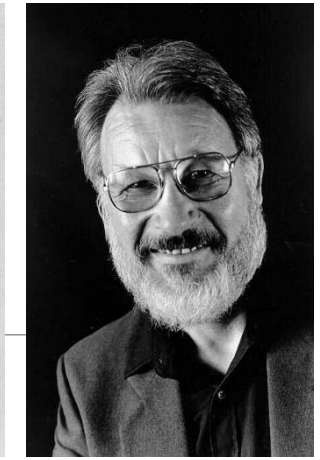
Vnitrotřídní korelace pro P×I design

Shrout a Fleiss (nejběžněji používané)	McGraw a Wong (občasně používané)	GT design
ICC(1,1)	One-way random, single score ICC(1)	I:p (jediná faseta plus error, N_e=1)
ICC(2,1)	Two-way random, single score ICC(A,1)	p×I (absolutní, N_i = 1)
ICC(3,1)	Two-way mixed, single score ICC(C,1)	p×I (relativní, N_i = 1)
ICC(1,k)	One-way random, average score ICC(k)	I:p (jediná faseta plus error, N_e=k)
ICC(2,k)	Two-way random, average score ICC(A,k)	p×I (absolutní, N_i = k)
ICC(3,k) = α	Two-way mixed, average score ICC(C,k)	p×I (relativní, N_i = k)

A=agreement, C=consistency

Krippendorfova alfa





Zobecnění konceptu klasického koeficientu alfa (např. Cronbachovy).

Cronbachova alfa: $\alpha = 1 - \frac{\text{chybový rozptyl}}{\text{celkový rozptyl}}$

- (plus nějaké korekce na počet stupňů volnosti)

Krippendrofova alfa:

$$\alpha = 1 - \frac{\text{pozorovaná neshoda}}{\text{očekávaná neshoda}} \sim 1 - \frac{\text{rozdílnost v hodnocení subjektů}}{\text{rozdílnost subjektů} + \text{rozdílnost v hodnocení subjektů}}$$

Použitelné pro nominální, ordinální i intervalové proměnné a libovolný počet hodnotitelů.

- Jen se různým způsobem vyjádří pozorovaná a očekávaná neshoda.
- **Díky tomu stejný význam napříč různými typy proměnných, koeficienty lze částečně srovnávat.**
- Použitelné i v případě chybějících dat.

Kde začít? Software

SPSS: scale (ICC), crosstabs (kappa) a pluginy.

R: zejm. balíčky `irr`, `raters`, `concord`, `psych`.

JASP: modul reliability (ICC, kappa)

JAMOVI: modul SimplyAgree

Reálně existuje mnohem větší množství dalších koeficientů.

- Je v tom celkově zmatek.
- Pokusil jsem se představit ty hlavní a nejčastěji používané.

Kazuistika: Přijímací zkoušky do NMGR psychologie FSS během COVIDu

Cígler, H., Ježek, S., Širůček, J., & Lacinová, L. (2022). Hodnocení bakalářských prací jako přijímací kritérium do navazujícího magisterského studia: Psychometrická kazuistika. *Studia Paedagogica, 1*(93–124). https://doi.org/10.5817/SP2022-1-4

Odkazy:

- Popularizace: https://psych.fss.muni.cz/cosedeje/aktuality/prijimaci-zkouska-hodnoceni-bp
- Data: https://doi.org/10.17605/OSF.IO/QX5U7