MUNI FSS Katedra psychologie

Reliabilita a metody odhadu

PSYb2590: Základy psychometriky | **Přednáška 7**

2. 4. 2024 | Hynek Cígler

Existují žáby?

Jaká je žabovitost dvorku?

Copilot: Žába vydává zvuky u jezírka na tajemném dvorku starého domu. Noc, svítí měsíc a hvězdy. Psycholog vše sleduje u sklenky vína a cigarety. Hyperrealismus.

Opakování poslední přednášky

Klasická testová teorie (CTT), pravé a pozorované skóre.

$$X = T + e$$

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_e^2$$

Minivýlet do algebry:

$$\sigma_{A+B}^2 = \sigma_A^2 + \sigma_B^2 + 2\sigma_{AB} = \sigma_A^2 + \sigma_B^2 + 2r_{ab}\sigma_A\sigma_B$$

Divadlo pana Browna a myšlenkový experiment.

Reliabilita $r_{xx'}$ jako „vysvětlený rozptyl“:

$$r_{xx'} = (R^2) = \frac{\sigma_\tau^2}{\sigma_x^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

$$r_{xx'} = r_{x\tau}^2$$

Reliabilita v pojetí CTT

Reliabilita je vysvětlený rozptyl:

$$r_{xx'} = \frac{\sigma_\tau^2}{\sigma_x^2} = (R^2) = r_{x\tau}^2$$

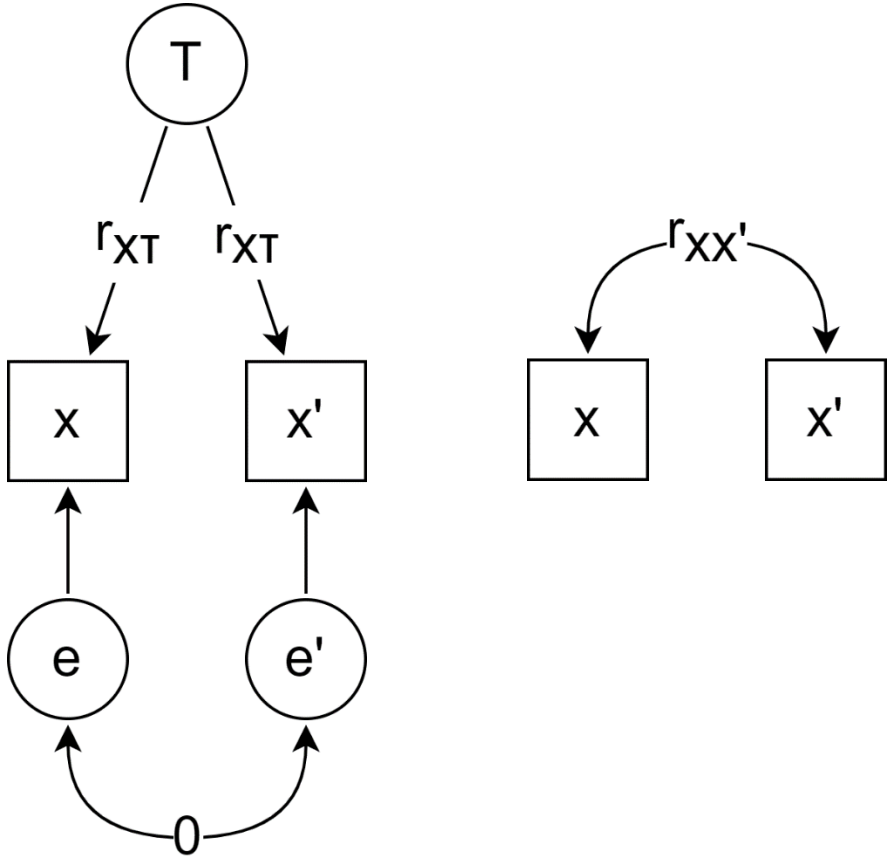
Jak ale zjistit korelaci $r_{x\tau}$?

Lokální nezávislost dvou měření x a x':

$$r(x, x'|\tau) = 0$$

Aplikace Wrightových pravidel:

$$r(x, x') = r_{x\tau} \cdot r_{x\tau} = r_{x\tau}^2 = r_{xx'}$$



Reliabilita v pojetí CTT

Člověka není možné měřit opakovaně a tak imitovat postup přírodních věd.

- Malý počet pozorování.
- Proces testování ovlivňuje testovanou osobu.
- Celkově vysoká míra chyby, která dále zamlžuje všechny odhady.

Spearman (1904) proto přišel s konceptem reliability.

Namísto paralelního testování jedné osoby
pracujeme s paralelními testy na vzorku osob.

Korelace paralelních testů je potom rovna reliabilitě.

- Proto ten symbol $r_{xx'}$.

Attenuation

Spearmanovou (1904) motivací byl odhad korelací pravých skórů nezkreslených chybou měření.

Tzv. *„attenuation coefficient“*, *„korekce proti oslabení“*, *„korekce proti nereliabilitě“*. Odhad korelace pravých skórů:

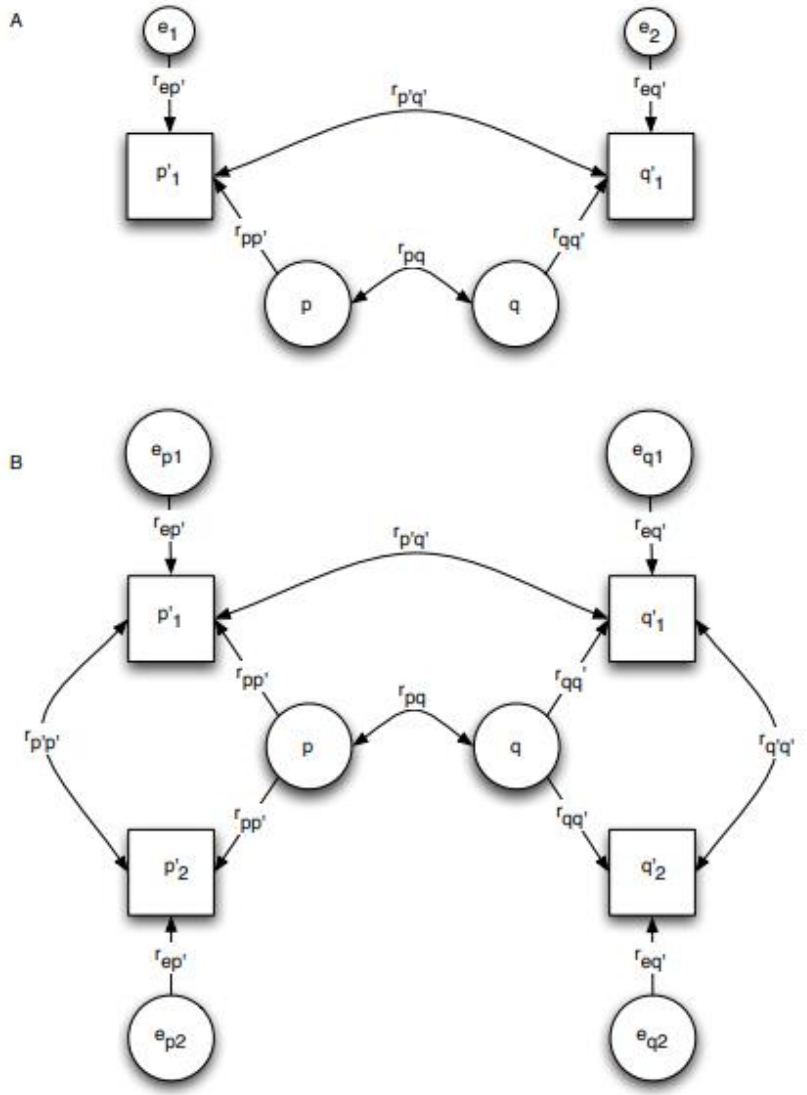
$$r_{pq}^{*} = \frac{r_{pq}}{\sqrt{r_{pp'} r_{qq'}}}$$

- Kde r_{pq}^{*} je odhad korelace pravých skórů *p*, *q*, r_{pq} je pozorovaná korelace testů p a q a $r_{pp'}$, $r_{qq'}$ jsou jejich reliability.

- Protože korelace pravých skórů $r_{pq}^{*} \leq 1$, lze odhadnout maximální možnou pozorovanou korelaci 2 testů jako:

$$r_{pq} \leq \sqrt{r_{pp'} r_{qq'}}$$

- **Korelace nemůže být vyšší než odmocnina součinu reliabilit!**



(Pozor, notace na diagramu je atypická a neodpovídá rovnicím.)

Fig. 7.1 Spearman's model of attenuation and reliability. Panel A: The true relationship between p and q is attenuated by the error in p' and q'. Panel B: the correlation between the latent variable p and the observed variable p' may be estimated from the correlation of p' with a parallel test.

https://www.personality-project.org/r/book/Chapter7.pdf

Paralelní testy

Aby celý postup fungoval, je nutné zavést několik realistických předpokladů.

Paralelní testy jsou takové, pro které platí:

- A. Pravý skór je ve všech testech a pro každý měřený subjekt stejný.

$$T = \mathrm{E}(X) = \lim_{n \to \infty} \frac{\sum_{i=1}^{n} X_i}{n}$$

- B. Rozptyl pravých skórů je v obou testech stejný (důsledek A): $\sigma_\tau = \sigma_{\tau'}$.
- C. Chybový rozptyl je v obou testech a pro každý subjekt stejný: $\sigma_e = \sigma_{e'}$.
- D. Shodný rozptyl pozorovaných skórů obou testů (důsledek A a C): $\sigma_x = \sigma_{x'}$.

Jinými slovy: „Lidé se **nemění** a test měří pořád ‚**stejně**‘.“

Tyto předpoklady jsou v psychologii příliš silné.

- Proto častěji uvažujeme o míře paralelnosti.

CTT: Paralelní testy

Úrovně paralelnosti položek (podobné modelu faktorové analýzy):

$$X_{ip} = i_i + a_i\tau_p + e_{ip}, \qquad e_{ip} \sim \mathrm{N}\left(0, \mathrm{var}(e_i)\right)$$

Kongenerické: Vybrané ze stejné domény. Stejná struktura rovnice pro všechny položky.

- Měří stejný rys (trs rysů), ale jiným způsobem.

Tau-ekvivalentní: Stejná lineární souvislost s měřeným atributem.

- + Shodné nestandardizované faktorové náboje („měřítko“ položky).

Paralelní: Položky měří se stejnou velikostí chyby.

- + Shodné reziduální rozptyly.

Striktně paralelní: Stejná obtížnost všech položek.

- + Shodné intercepty/průměry položek.
- U binárních položek paralelní = striktně paralelní, protože $\mathrm{var}(X_i) = P_i(1 - P_i)$.

- X_{ip} – pozorované skóre osoby p na pol. i
- i_i, a_i – intercept a faktorový náboj pol. i
- τ_p – pravé skóre osoby p
- e_{ip} – náhodná chyba osoby p na pol. i (reziduum)
- $e_{ip} \sim \mathrm{N}\left(0, \mathrm{var}(e_i)\right)$ – tato chyba pochází z normálního rozložení s průměrem 0 a rozptylem $\mathrm{var}(e_i)$

Někdy též kongenerické → esenciálně tau-ekvivalentní → tau-ekvivalentní → paralelní

CTT: Paralelní testy

Úrovně paralelnosti položek (podobné modelu faktorové analýzy):

$$X_{ip} = i_i + a_i\tau_p + e_{ip}, \qquad e_{ip} \sim \mathrm{N}\left(0, \mathrm{var}(e_i)\right)$$

Kongenerické: Vybrané ze stejné domény. Stejná struktura rovnice pro všechny položky.

- Měří stejný rys (trs rysů), ale jiným způsobem.

Tau-ekvivalentní: Stejná lineární souvislost s měřeným atributem.

- + Shodné nestandardizované faktorové náboje („měřítko“ položky).

Paralelní: Položky měří se stejnou velikostí chyby.

- + Shodné reziduální rozptyly.

Striktně paralelní: Stejná obtížnost všech položek.

- + Shodné intercepty/průměry položek.
- U binárních položek paralelní = striktně paralelní, protože $\mathrm{var}(X_i) = P_i(1 - P_i)$.

Někdy též kongenerické → esenciálně tau-ekvivalentní → tau-ekvivalentní → paralelní

CTT: Paralelní testy

Úrovně paralelnosti položek (podobné modelu faktorové analýzy):

$$X_{ip} = i_i + a\tau_p + e_{ip}, \qquad e_{ip} \sim \mathrm{N}\left(0, \mathrm{var}(e_i)\right)$$

Kongenerické: Vybrané ze stejné domény. Stejná struktura rovnice pro všechny položky.

- Měří stejný rys (trs rysů), ale jiným způsobem.

Tau-ekvivalentní: Stejná lineární souvislost s měřeným atributem. $a_i = a$

- + Shodné nestandardizované faktorové náboje („měřítko“ položky).

Paralelní: Položky měří se stejnou velikostí chyby.

- + Shodné reziduální rozptyly.

Striktně paralelní: Stejná obtížnost všech položek.

- + Shodné intercepty/průměry položek.
- U binárních položek paralelní = striktně paralelní, protože $\mathrm{var}(X_i) = P_i(1 - P_i)$.

Někdy též kongenerické → esenciálně tau-ekvivalentní → tau-ekvivalentní → paralelní

CTT: Paralelní testy

Úrovně paralelnosti položek (podobné modelu faktorové analýzy):

$$X_{ip} = i_i + a\tau_p + e_{ip}, \qquad e_{ip} \sim \mathrm{N}\left(0, \mathrm{var}(e)\right)$$

Kongenerické: Vybrané ze stejné domény. Stejná struktura rovnice pro všechny položky.

- Měří stejný rys (trs rysů), ale jiným způsobem.

Tau-ekvivalentní: Stejná lineární souvislost s měřeným atributem.

- + Shodné nestandardizované faktorové náboje („měřítko“ položky).

Paralelní: Položky měří se stejnou velikostí chyby. $a_i = a, \mathrm{var}\left(e_{ip}\right) = \mathrm{var}(e)$

- + Shodné reziduální rozptyly.

Striktně paralelní: Stejná obtížnost všech položek.

- + Shodné intercepty/průměry položek.
- U binárních položek paralelní = striktně paralelní, protože $\mathrm{var}(X_i) = P_i(1 - P_i)$.

Někdy též kongenerické → esenciálně tau-ekvivalentní → tau-ekvivalentní → paralelní

CTT: Paralelní testy

Úrovně paralelnosti položek (založené na faktorové analýze):

$$X_{ip} = i + a\tau_p + e_{ip}, \qquad e_{ip} \sim \mathrm{N}\left(0, \mathrm{var}(e)\right)$$

Kongenerické: Vybrané ze stejné domény. Stejná struktura rovnice pro všechny položky.

- Měří stejný rys (trs rysů), ale jiným způsobem.

Tau-ekvivalentní: Stejná lineární souvislost s měřeným atributem.

- + Shodné nestandardizované faktorové náboje („měřítko“ položky).

Paralelní: Položky měří se stejnou velikostí chyby.

- + Shodné reziduální rozptyly.

Striktně paralelní: Stejná obtížnost všech položek. $a_i = a, \mathrm{var}\left(e_{ip}\right) = \mathrm{var}(e), i_i = i$

- + Shodné intercepty/průměry položek.
- U binárních položek paralelní = striktně paralelní, protože $\mathrm{var}(X_i) = P_i(1 - P_i)$.

Někdy též kongenerické → esenciálně tau-ekvivalentní → tau-ekvivalentní → paralelní

Odhad reliability

Ukázali jsme si, že reliabilita je korelací dvou paralelních testů.

Proto postup odhadu v rámci CTT:

- 1. Identifikace (alespoň) dvou paralelních testů.
- 2. Stanovení předpokladů.
- 3. Sběr dat s využitím vzorku populace.
- 4. Odhad korelace těchto testů.

Odhady reliability

Tradiční způsoby odhadu:

1. Stabilita v čase (test-retest)

2. Shoda posuzovatelů

3. Paralelní formy testu

4. Vnitřní konzistence

A další netradiční postupy
(zejm. model-based reliabilita)

Lee Cronbach (1916–2001)
autor koeficientu alfa

Reliabilita: typické postupy ověření v CTT

Stabilita v čase, reliabilita typu test-retest

- Měří test stále stejně? Paralelním testem (PT) je ten samý test administrovaný jindy.

Shoda posuzovatelů, inter-rater reliabilita.

- Docházejí administrátoři ke stejným závěrům? PT je stejný test administrovaný někým jiným.

Reliabilita paralelních forem.

- Měří obě/všechny formy testu to stejné? PT je jiný test vytvořený tak, aby „byl stejný“.

Vnitřní konzistence a split-half

- Měří položky to stejné? PT jsou jednotlivé položky/půlky testu.
- Cronbachovo alfa, split-half a další.

Lze čekat, že všechny koeficienty/odhady reliability budou stejné?

Metoda test-retest

ODHAD RELIABILITY

Stabilita v čase, test-retest reliabilita

Poskytuje test při opakovaném měření shodné odhady atributu?

Metoda: Korelace dvou měření (rank-order stability).

Předpoklady:

- Rys je (dostatečně) stabilní v čase.
- Měření jsou na sobě nezávislá. Zapamatování položek? Únava?

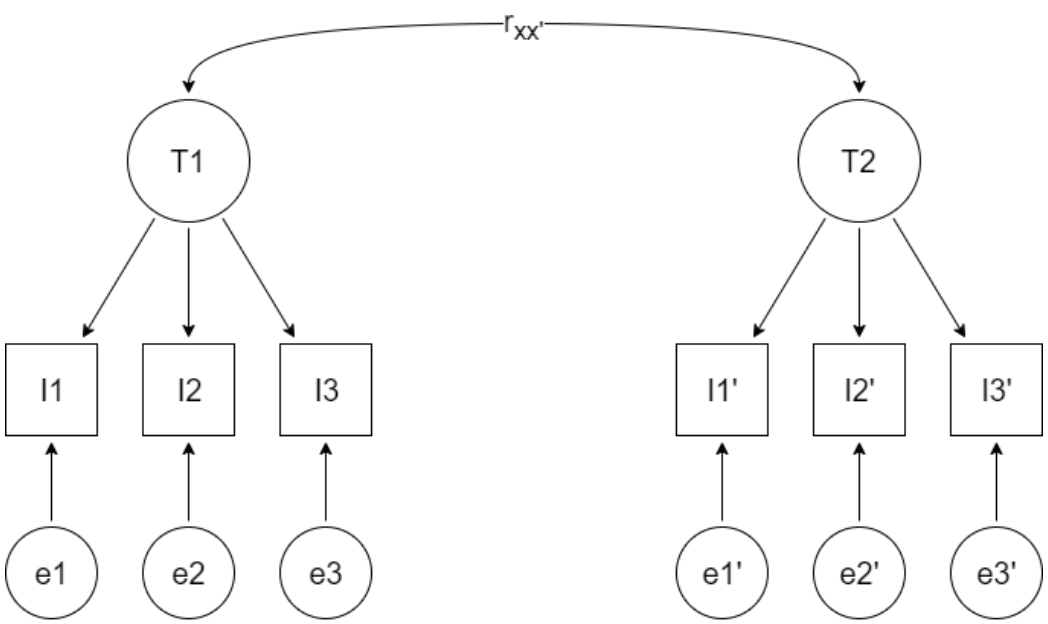
Problém: reálná fluktuace rysu v čase je považována za chybu měření.

Stabilita rysu (korelace T) vs. stabilita metody (korelace X|T).

Někdy se rozlišuje:

- **Dependabilita měření** – krátký interval, nepředpokládá se změna úrovně rysu.
- **Stabilita měření** – dlouhý interval, zahrnuje přirozené rysu fluktuace rysu v čase.

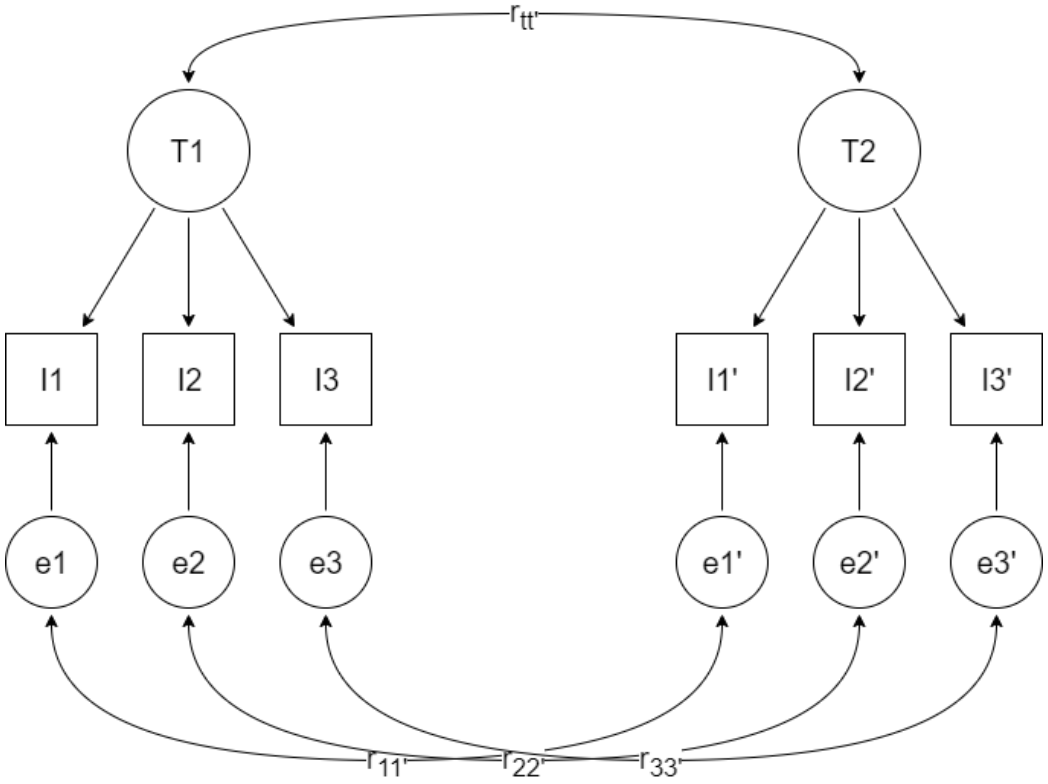
Nezávislost chyb měření



Chyby měření nebývají zcela náhodné, ale obsahují systematickou složku stabilní v čase.

- U výkonových testů méně, u dotazníků více.

Nezávislost chyb měření



Chyby měření nebývají zcela náhodné, ale obsahují systematickou složku stabilní v čase.

- U výkonových testů méně, u dotazníků více.

Co to udělá s korelací celého testu?

- Tedy $r(\sum I_i , \sum I_i')$?

Jakou informaci ponese tato korelace?

Jaký bude vztah reliability a korelace?

Dvě pojetí reliability:

- Reliabilita jako vysvětlený rozptyl ($r_{x\tau}^2$)
- Reliabilita jako korelace paralelních testů (r_{xx}')
- **Nejsou-li chyby měření nezávislé, $r_{x\tau}^2 \neq r_{xx}'$**

Reliabilita paralelních forem

ODHAD RELIABILITY

Reliabilita paralelních forem

Poskytují dva testy shodné odhady atributu?

Metoda: Korelace paralelních forem testu.

Účel používání paralelních testů:

- Zabránit opisování při hromadné administraci.
- Zabránit zapamatování položek při opakované administraci a retestování (PPP).
- Umožnit sběr data ve více nezávislých termínech (SCIO, TSP...).

Problém: I když jsou testy vytvořené stejným způsobem, málokdy měří zcela ten samý pravý rys.

Je nutné odlišit reliabilitu paralelních forem od existence paralelních forem jako takových.

- Vyvažování paralelních forem je celkově velmi náročné.

Více stupňů ekvivalence dvou testů:

- Alternativní: pouze podobné.
- Srovnatelné: srovnatelné standardní skóry.
- Ekvivalentní: srovnatelné hrubé skóry.
- (Striktně) paralelní: shodné pravé skóry.

Podobné stupňům paralelnosti.

Paralelní formy prakticky

Pokud neaspirujeme na „vyvážené“, striktně paralelní testy…
… postupujeme stejně, jako v případě test-retest.

Převedeme na stejné jednotky (T-skóry atp.) pro každou formu zvlášť a ověříme:

- shodu pořadí (korelace);
- shoda průměrů v případě práce s hrubými skóry (t-test);
- shodu rozptylů = homoskedascitu v případě práce s hrubými skóry (Levenův test);
- případně i linearitu skórů (scatter-plot, kvadratická/polynomická regrese).

(Vnitrotřídní) korelace je potom koeficientem reliability.

Co vše může způsobovat rozdíl průměrů obou forem?

- Jak se vyhnout těm vlivům, které „nechceme“?

Vyvažování paralelních forem

Spíš otázka norem a pedagogického testování (nikoli reliability).

Linking (skóry dvou forem testu jsou srovnatelné) vs. **equating** (testy měří to samé)

Jedno z typických využití *teorie odpovědi na položku* (IRT).

Samostatné obsáhlé publikace, specifická expertíza.

- Kolen, M. J., & Brennan, R. l. (2014). *Test equating, scaling and linking: methods and practices*. Springer.

Raw-score equating, ekvipercentilové vyvažování, linking functions, mapping functions.

Vyvažují se nejen formy, ale i jazykové mutace (PISA, TIMSS, TALIS).

Shoda posuzovatelů

ODHAD RELIABILITY

Shoda posuzovatelů

Docházejí dva hodnotitelé/administrátoři ke shodným závěrům?

Druhy neshody:

- Shoda administrátorů (např. WISC).
- Shoda posuzovatelů (např. ROR) – inter-rater, intra-rater reliabilita.
- V diagnostické praxi obtížně odlišitelné.

Korelace napravo: $r_{AB} = 0{,}93$. Opravdu se hodnotitelé shodují?

Komplikace 1: rozdílná „přísnost“ hodnotitelů.

- Je nutné vzít v úvahu i rozdílnou přísnost (zde Cohenovo $d = 1{,}3$).
- Používá se proto tzv. vnitrotřídní korelace (intra-class correlation), která bere v úvahu shodu pořadí, průměrů a lze použít pro libovolný počet hodnotitelů. Existuje 2×(3+2) variant ICC.
- V tomto případě $ICC(2{,}1) = 0{,}51$.
  - Pozn.: *ICC*(3,k) pro průměrné hodnocení je ekvivalentní s pojetím reliability podle Hoyta [URB, s. 112-114] a tedy s Cronbachovým α, v tomto případě $ICC(3{,}2) = 0{,}96$.

	rater A	rater B
ID1	4	7
ID2	2	4
ID3	6	7
ID4	1	3
ID5	3	5
ID6	5	6
M	3,00	5,67
SD	2,19	1,97
r_{AB}	0,93	

Shoda posuzovatelů: **komplikace 2**

Až příliš často nás zajímá shoda jednotlivých kritérií: **Úroveň měření**.

Reliabilita kódování na úrovni položky.

- Používá se i jako ukazatel interní validity v kvalitativním výzkumu.

Položky bývají nominální nebo ordinální, nelze proto použít *ICC* a korelace.

- A nelze použít podíl shody (např. „shodli se v 90 % případů“) kvůli nahodilé shodě.

Proto velké množství různých statistik:

- Cohenovo kappa – absolutní shoda 2 hodnotitelů vážená proti nahodilé shodě.
  - $\kappa = \frac{P_o - P_e}{1 - P_e}$, kde P_o je pozorovaná shoda a P_e zcela náhodná shoda (očekávaná)
- Vážené kappa – shoda 2 hodnotitelů v případě ordinálních položek.
- Fleissovo (vážené) kappa – shoda N hodnotitelů u nominálních (ordinálních) položek.
- Kendallův koeficient konkordance – analogie Spearmanovy korelace pro N hodnotitelů (jen pořadí).

Shoda posuzovatelů: Co si pamatovat?

V nouzi: shoda průměrů (např. t-test, ANOVA) plus pořadí (alfa, korelace)

- Nebo ordinální ekvivalenty (Mann-Whitney, Kruskal-Wallis, Spearmanova korelace).

V případě nominálních proměnných **za žádných okolností nepoužívat % shody!**

Zpravidla o dost jiná informace, než zbylé koeficienty.

Specifické koeficienty. Některé stojí pamatovat si podle jména:

- (Cohenova) kappa; vnitrotřídní korelace; Kendallův koeficient konkordance; Krippendorfova alfa.

Další zdroje:

- Hallgren, K. A. (2012). Computing Inter-Rater reliability for observational data: An overview and Tutorial. Tutorials in Quantitative Methods for Psychology, 8(1), 23–34. doi:10.20982/tqmp.08.1.p023
- Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., … Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. International Journal of Nursing Studies, 48(6), 661–671. doi:10.1016/j.ijnurstu.2011.01.016

Vnitřní konzistence

ODHAD RELIABILITY

Vnitřní konzistence

Často máme ale jedinou formu testu bez vlivu posuzovatele (dotazník) a nezajímá nás stabilita v čase nebo nemáme prostředky na dvě administrace (nebo to není možné).

Prostě je k dispozici jediné měření jednou metodou.

Dva hlavní postupy:

- Split-half.
- Vnitřní konzistence.

Split-half

Postup: Test rozdělíme na dvě půlky a pracujeme jako s reliabilitou paralelních forem.

Problém 1: Jak test rozdělit?

- Poloviny by měly být paralelní.
- Zpravidla tedy nějaké pseudo-náhodné rozdělení (sudá–lichá).
- Existuje velmi mnoho různých rozdělení a každé poskytne poněkud jiný odhad split-half reliability.

Problém 2: Odhad založen jen na jediné korelaci.

- Při srovnání s jinými koeficienty vnitřní konzistence (alfa, omega) menší přesnost odhadu (širší *CI*).

Problém 3: Zkrácení testu.

- Reliabilita je závislá na délce testu. Delší testy → vyšší reliabilita.
- Rozpůlením testu zjistíme reliabilitu jedné poloviny, reliabilita celého testu je nutně vyšší.

Problém 4: Lichý počet položek. Podstatný není počet položek, ale rozptyl půlek testu.

- U delších testů proto nehraje roli.

Split-half: Spearmanův-Brownův postup

„**Spearmanův-Brownův věštecký vzorec**“ (Spearman-Brown prophecy formula):

$$r^*_{xx'} = \frac{N r_{xx'}}{1 + (N-1) r_{xx'}}$$

- N – poměr délek testů; $r_{xx'}$ – původní reliabilita; $r^*_{xx'}$ odhad reliability po změně délky.
- *„Jaká bude reliabilita $r^*_{xx'}$ při N-násobné změně délky testu?“*

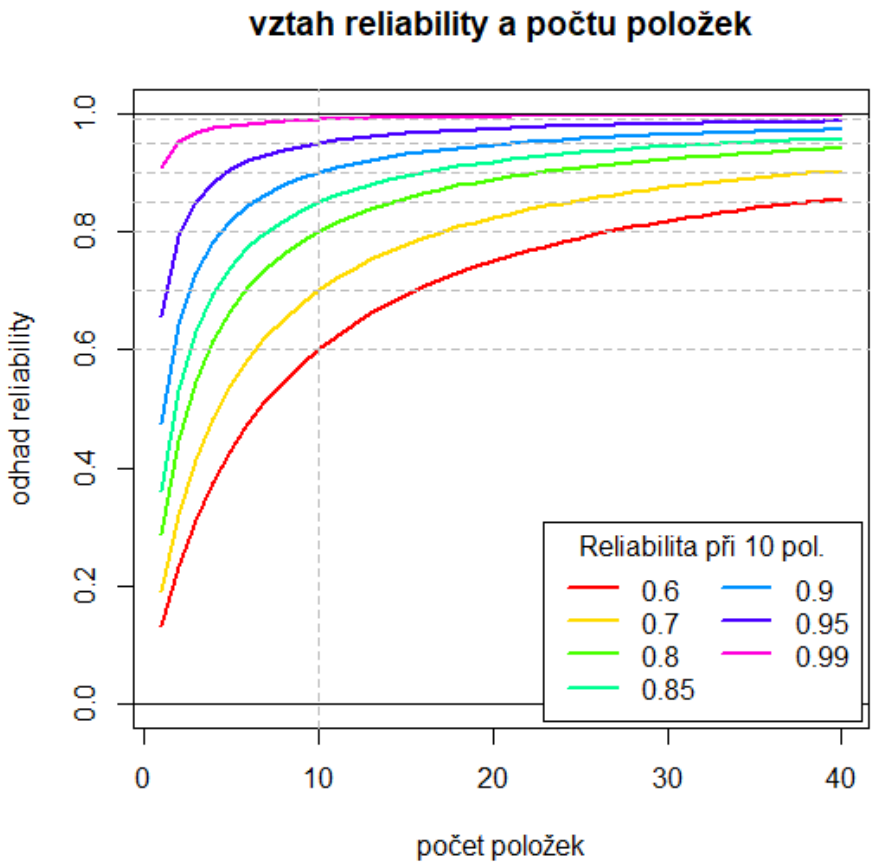
V případě split-half reliability $N = 2$ (test je dvakrát delší než polovina):

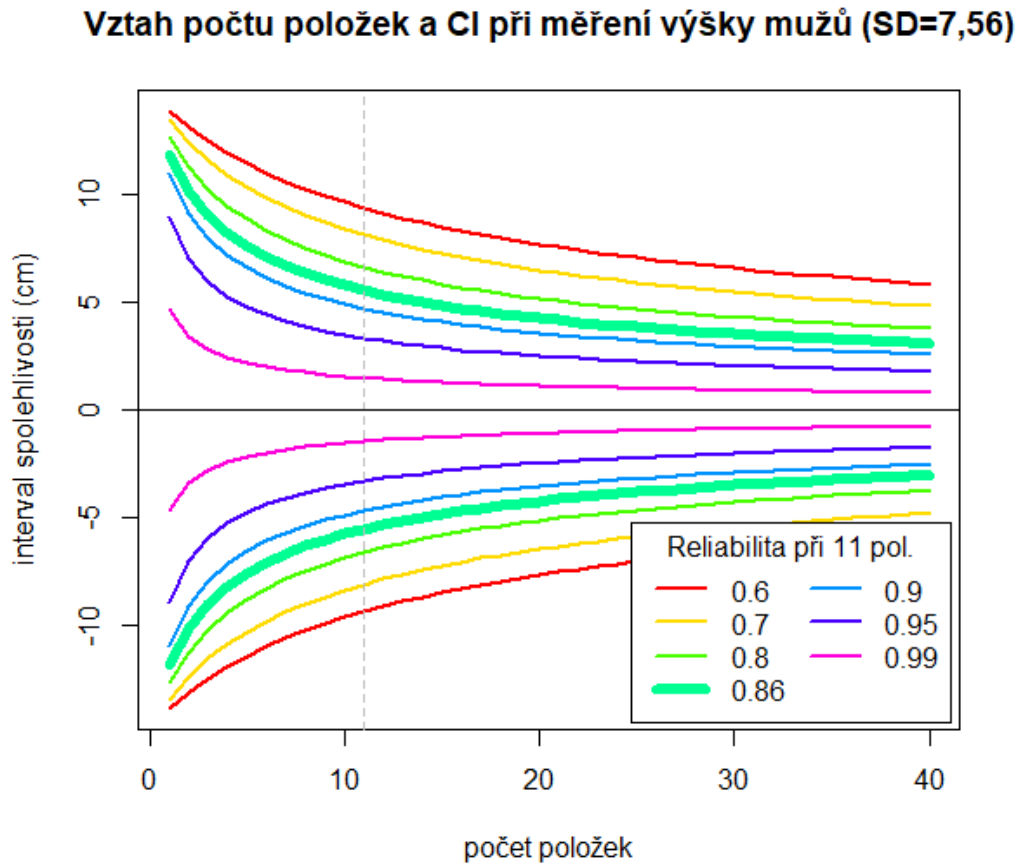
$$r_{SB} = r^*_{xx'} = \frac{2 r_{xx'}}{1 + r_{xx'}}$$

Slouží i k odhadu požadovaného počtu položek pro dosažení určité reliability.

- Předpokladem jsou striktně-paralelní položky.

Vztah reliability testu a jeho délky





V případě 11položkového dotazníku výšky ze začátku semestru r = 0,86.

Split-half: Guttmanova lambda 4

Guttman (1945) publikoval 6 různých odhadů reliability λ_{1-6}. Podstatné jsou dva z nich:

$$\lambda_4 = \frac{4\sigma_{pq}^2}{\sigma_x^2}$$

- kde σ_{pq}^2 je kovariance polovin testu a $\sigma_x^2 = \sigma_p^2 + \sigma_q^2 + 2\sigma_{pq}^2$ je rozptyl celého testu.
- λ_4 je shodná s Cronbachovou alfou u dvoupoložkových testů.
- λ_3 je určena pro vícepoložkové testy a je shodná s Cronbachovou alfou (viz dále).

Spearman-Brown vs. lambda 4:

- SB může při porušení předpokladů reliabilitu nadhodnotit, λ_4 je vždy nižší než skutečná reliabilita.
- Pokud se poloviny testu výrazně liší svou délkou či rozptylem, λ_4 může výrazně podhodnotit.
- Jsou-li poloviny standardizovány, pak platí $\lambda_4 = r_{SB} = \alpha$.
- U dlouhých testů oba postupy vedou k podobným odhadům.

Poloviny testů by při jakémkoli split-half přístupu měly být „stejně dlouhé“.

- Pokud nejsou, lze využít jiné postupy (Cígler a Chvojková, preprint; Warrens, 2016).

Split-half: specifické použití

Greatest-Lower Bound of reliability.

- Řada rozdílných postupů a algoritmů.
- Anotace jako *GLB*, *glb*, σ_+, ρ_{glb} apod.

V poslední době je Guttmanova λ_4 chápána jako synonymum pro GLB.

Položky jsou rozděleny tak, aby byla korelace polovin testu maximalizovaná.

- Může být analyticky náročné.
- Na malých vzorcích a krátkých testech vede k nadhodnocení z důvodu výběrové chyby („příliš dobré“ rozpůlení).
- Doporučení: N > 1000. Vyhnout se N < 200.

Cronbachovo alfa

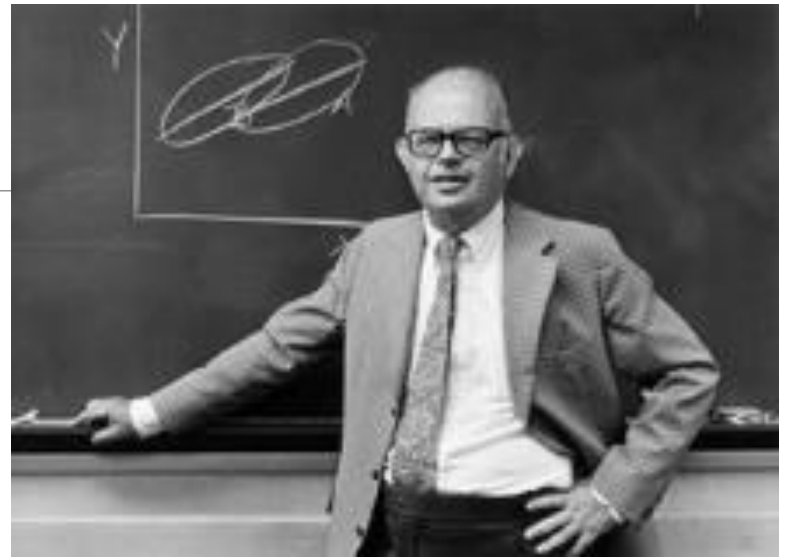
Co když jsou paralelními testy jednotlivé položky?

- Pokud měří všechny to samé, pak by spolu měly hodně korelovat – být vnitřně konzistentní.
- Položky měří totéž, pokud mají hodně sdíleného rozptylu.

Cronbachova (1951) alfa:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_x^2}\right)$$

- σ_i^2 – rozptyl položky *i*, $\sum_{i=1}^{k} \sigma_i^2$ je diagonála variančně-kovarianční matice (jedinečný/chybový rozptyl položek)
- σ_x^2 – rozptyl celého testu, tedy suma var-covar matice
- k – počet položek (ne celý jedinečný rozptyl položek je chybou, proto korekce $\frac{k}{k-1}$, aby reliabilita mohla být 1)
  - Bez této korekce jde o Guttmanovu λ_1.



	A	B	C
A	1	0,514	0,477
B	0,514	1	0,662
C	0,477	0,662	1

Část korelační matice Holzinger a Swineford (1937):

$$\alpha = \frac{3}{2}\left(1 - \frac{1+1+1}{1+1+1+2(0{,}514+0{,}477+0{,}662)}\right) = 0{,}786$$

Cronbachovo alfa: předpoklady

Tau-ekvivalentní položky

- Stejná lineární souvislost položky s pravým skóre...
- ... a tedy shodné faktorové náboje ve faktorové analýze (viz přednáška o FA).
- Při nedodržení podhodnocuje.

Unikátní rozptyl je celý chybovým rozptylem.

- A proto tzv. „spodní hranice reliability“.

Lokální nezávislost položek (**jednodimenzionalita**).

- Nedodržení může nadhodit i podhodnotit.

Alfa ale není ukazatelem jednodimenzionality!

- I vícedimenzionální testy mohou mít vysokou vnitřní konzistenci, viz např. Marko (2016).

Cronbachovo alfa: varianty

Standardizovaná Cronbachova alfa:

- Korelační, nikoliv kovarianční matice.
- Vnitřní konzistence *standardizovaných* položek.
- Robustnější při výrazně rozdílné obtížnosti položek (slabší předpoklad tau-ekvivalence).

Kuderův-Richardsonův (1931) vzorec 20 a 21

$$KR_{20} = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} P_i(1-P_i)}{\sigma_x^2}\right]$$

- V případě binárních položek, kdy $P_i(1-P_i)$ je rozptyl dichotomické položky.
- KR_{20} = α, KR_{21} pro položky stejné obtížnosti.
- Spíše historické kvůli snadnosti výpočtu.

Psychometrický paradox

Hypotetický dotazník extroverze:

- Rád se vídám s lidmi.
- Rád jsem v kontaktu s lidmi.
- Vyhledávám společnost lidí.
- Jsem rád mezi lidmi.
- Dělá mi dobře společnost lidí.
- ...

Psychometrický paradox

Reliabilita testu je funkcí korelací mezi položkami a jejich počtem.

Čím více spolu položky korelují, tím „ostřeji“ se zaměřují na specifický rys.

„*Alfa tuning*“ škál: výběr nejvíce korelujících položek a zvýšení reliablity.

- Měříme stále přesněji stále méně (menší výsek konstruktu) – ztráta (výběrové) validity.
- Někdy i jako cílená aktivita; de facto může jít o podvod (synonymní páry položek...).

Nikoli vždy!

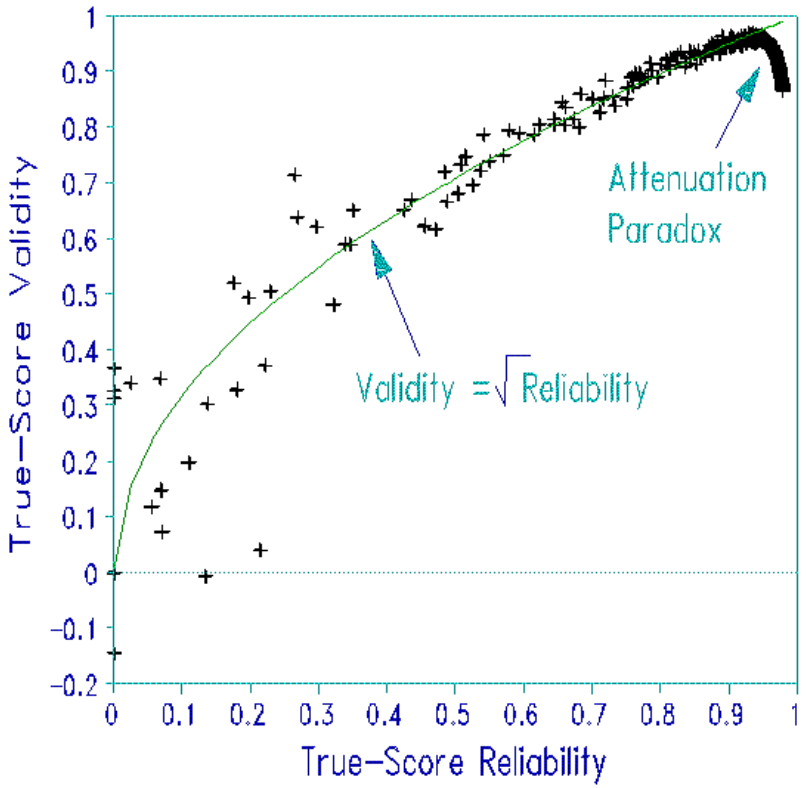


Figure 1. The Attenuation Paradox

https://www.rasch.org/rmt/rmt94a.htm

Kdy použít split-half?

Vnitřní konzistence (alfa, omega...) bývá výhodnější než split-half.

- Přesnější a robustnější odhad.

Výjimka: časované/rychlostní testy nebo testy s pravidlem ukončení.

- Počet správně vyřešených položek za 1 minutu (např. Test pozornosti d2).
- „Ukončete administraci po 5 chybných odpovědích“ (např. Wechslerovy testy).
- Datová matice obsahuje řadu chybějících dat na koncích řádků.

Určité výhody i u velkých datasetů (N>1000, ideálně N>5000).

- GLB, menší statistické předpoklady (např. ve srovnání s binárními pol.).

Specifické příklady vnitřní konzistence

Reliabilita celkového skóre v multidimenzionálních testech.

- Např. reliabilita celkového skóre v inteligenčním testu (WISC, WAIS).

Reliabilita váženého skóre.

- Celkové skóre je váženým součtem dílčích položek/subtestů.

Reliabilita rozdílového skóre.

- Např. reliabilita rozdílu rychlosti a správnosti v testu pozornosti d2.

Reliabilita v testech založených na jiné teorii měření.

- Typicky IRT nebo jiné modely s latentními proměnnými, případně teorie zobecnitelnosti.

V těchto případech je pro odhad vnitřní konzistence použít jiné postupy.

- Postupů pro odhad reliability je mnoho – představili jsme jen nejzákladnější postupy.

Model-based reliabilita

CTT je antirealistická – reliabilita jako „vysvětlený rozptyl“ nedává moc smysl.

Při dodržení předpokladů je ale korelace paralelních testů rovna R^2.

Lze proto využít realistický model měření pro odhad reliability.

- Podrobně viz Bentler P. M. (2009). Alpha, Dimension-Free, and Model-Based Internal Consistency Reliability. *Psychometrika*, *74*(1), 137–143. doi:10.1007/s11336-008-9100-1

Dva odhady reliability:

- **Dimension-free reliabilita** – prostě jen odhad korelace paralelních testů bez ohledu na vnitřní strukturu testu.
- **Model-based reliabilita** – rozptyl hrubého skóre vysvětlený latentním rysem.

Model-based reliabilita: **omega**

Rodina koeficientů; Betlerova, Raykovova, ... a zejm. **McDonaldova omega**.

Obecný vzorec (Bollen, 1980; Raykov, 2001):

$$\omega = \frac{\left(\sum_{i=1}^{n} \lambda_i\right)^2 \sigma_\psi^2}{\left(\sum_{i=1}^{n} \lambda_i\right)^2 \sigma_\psi^2 + \sum_{i=1}^{n} \sigma_{e;i}^2 + 2\sum_{i<j} \sigma_{ij}^2} = \frac{\left(\sum_{i=1}^{n} \lambda_i\right)^2 \sigma_\psi^2}{\sigma_x^2}$$

- λ_i = faktorový náboj položky i
- σ_ψ^2 = rozptyl faktoru, σ_x^2 = celkový pozorovaný rozptyl
- $\sigma_{e;i}^2$ = reziduální rozptyl položky i
- σ_{ij}^2 = kovariance položek i, j

Bez předpokladu tau-ekvivalence (rozdílné faktorové náboje jsou zohledněny).

Model-based reliabilita: **omega**

Rodina koeficientů; Betlerova, Raykovova, ... a zejm. **McDonaldova omega**.

Obecný vzorec (Bollen, 1980; Raykov, 2001):

$$\omega = \frac{\left(\sum_{i=1}^{n} \lambda_i\right)^2 \sigma_\psi^2}{\left(\sum_{i=1}^{n} \lambda_i\right)^2 \sigma_\psi^2 + \sum_{i=1}^{n} \sigma_{e;i}^2 + 2\sum_{i<j} \sigma_{ij}^2} = \frac{\left(\sum_{i=1}^{n} \lambda_i\right)^2 \sigma_\psi^2}{\sigma_x^2}$$

- λ_i = faktorový náboj položky i
- σ_ψ^2 = rozptyl faktoru, σ_x^2 = celkový pozorovaný rozptyl
- $\sigma_{e;i}^2$ = reziduální rozptyl položky i (náhodný chybový rozptyl)
- σ_{ij}^2 = kovariance položek i, j (systematický chybový rozptyl)

- vysvětlený rozptyl
- chybový rozptyl
- celkový rozptyl

Bez předpokladu tau-ekvivalence (rozdílné faktorové náboje jsou zohledněny).

Model-based reliabilita: **omega**

Omega má předpoklad pouze kongenerických položek a lokální nezávislosti.

- Předpokladem alfy jsou tau-ekvivalentní položky. Omega proto bývá o něco vyšší.
- Zejména v případě silného porušení tau-ekvivalence.
- Zejména v případě malého počtu položek (méně než 5).

Stále „spodní hranice reliability“.

- Celý unikátní rozptyl je považován za chybový.

Při výpočtu lze jednoduše vzít v potaz další aspekty modelu.

- Reziduální kovariance, hierarchická struktura dat, vícedimenzionalita...
- Defaultní odhad v JASP ale s těmito aspekty nepracuje!

Reliabilita: interpretace

Reliabilita je ukazatelem kvality testu.

- Řada doporučení ohledně minimální hranice přípustné reliability. Typicky Klineovo pravidlo: $r_{xx'} > 0{,}7$.
- Záleží ale na účelu testu: nižší nároky pro výzkumné metody, vyšší nároky pro metody určené do praxe, nejvyšší nároky na high-stakes testy (SCIO, inteligenční test...).
- V případě výzkumu záleží i na způsobu využití (SEM vs. pozorované skóry).

Doporučené hodnoty reliability:

- „Nejlepší“ metody (celkový skór IST-2000-R) nebo testy základních kognitivních funkcí (Bourdonova zkouška): $r_{xx'} > 0{,}95$.
- Dobré testy: $r_{xx'} > 0{,}90$. Ve výzkumu výjimečně i $r_{xx'} > 0{,}70$.
- Osobně považuji testy s $r_{xx'} < 0{,}80$ za problematické. **Vždy ale záleží na účelu měření!**

Reliabilita rozdílu

Jak reliabilní je používání rozdílu mezi dvěma testy?

- Například VIQ a PIQ ve WAIS-III?

$$r_{x-y} = \frac{\sigma_x^2 r_{xx'} + \sigma_y^2 r_{yy'} - 2 r_{xy} \sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2 - 2 r_{xy} \sigma_x \sigma_y},$$

- kde σ_x^2 a σ_y^2 jsou rozptyly obou testů, $r_{xx'}$ a $r_{yy'}$ jejich reliability a r_{xy} je jejich korelace.
- jmenovatel je roven rozptylu výsledných rozdílů.

Pokud $\sigma_x^2 = \sigma_y^2 = \sigma_{xy}^2$ (v případě standardizovaných testů), pak:

- $r_{x-y} = \sigma_{xy}^2 \frac{r_{xx'} + r_{yy'} - 2 r_{xy}}{2 - 2 r_{xy}}$

Reliabilita rozdílu

$r_{xx'}$	$r_{yy'}$	r_{xy}	r_{x-y}	SD_{x-y}	SE_{x-y}	$CI_{95\%}$
0,7	0,8	0	**0,75**	21,2	10,6	20,8
0,7	0,8	0,2	**0,69**	19,0	10,6	20,8
0,7	0,8	0,4	**0,58**	16,4	10,6	20,8
0,7	0,8	0,6	**0,38**	13,4	10,6	20,8
0,7	0,7	0,6	**0,25**	13,4	11,6	22,8
0,9	0,9	0,8	**0,50**	9,5	6,7	13,1
0,9	0,9	0,45	**0,82**	15,7	6,7	13,1
0,6	0,6	0,5	**0,20**	15,0	13,4	26,3
0,7	0,7	0,65	**0,14**	12,5	11,6	22,8

Standardní chybu rozdílu lze spočítat s pomocí SD a SE vlevo, nebo prostřednictvím vzorce.

- Viz seminář.

Toto je důvod, proč je problematická interpretace rozdílu vysoce korelovaných subtestů.

- Téměř u nikoho se neliší...