MUNI FSS Katedra psychologie

Přednáška 2: Model klasické testové teorie

13. 10. 2019 | PSYn4790 | **Psychometrika: Měření v psychologii**

Katedra psychologie, Fakulta sociálních studií MU

Hynek Cígler

Obsah přednášky

Chyba měření a interval spolehlivosti.

- Opakování, inference o statisticky vs. klinicky významném rozdílu.

Model měření klasické testové teorie.

Pokročilé odhady reliability.

Cíle přednášky

Zvážení chyby měření v různých praktických situacích.

Co děláme když používáme CTT k měření?

Proč není alfa dobrý koeficient?

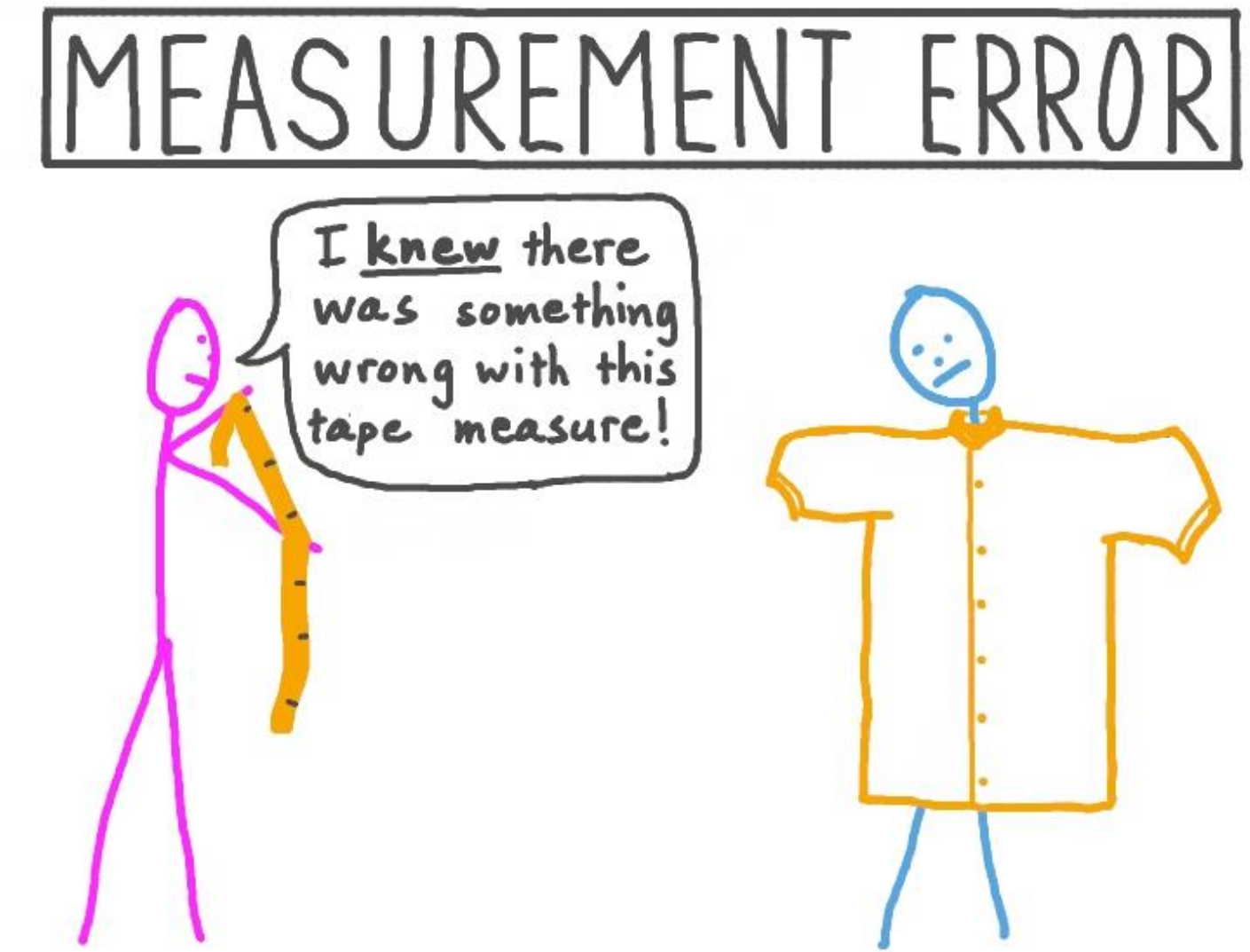
Jaké jsou alternativy ke koeficientu alfa?

Chyba měření a intervaly spolehlivosti

Opakování:
standardní chyba měření
standardní chyba predikce
standardní chyba rozdílu

Statisticky významný rozdíl

Klinicky významný rozdíl



https://www.nagwa.com/en/videos/138104137874/

Otázky spojené s chybou měření

Respondentovi naměřím výšku 178 cm.
Jaké otázky si mohu položit?

- Kolik měří právě teď?
- Kolik bude měřit příště?
- Kolik mu můžu naměřit příště, pokud se jeho výška nezmění?
- Kolik mu musím naměřit příště, abych mohl konstatovat, že se jeho výška změnila?

Kromě toho naměřím i jeho hmotnost 65 kg.
Jaké další otázky si mohu položit?

- Je „vyšší než těžší“?
- Je „vyšší než těžší“ oproti jiným respondentům?

Chyba měření

Standardní chyba měření: směrodatná odchylka pozorovaných hodnot okolo skutečné úrovně atributu

Příklad:

- https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm
- https://www.zoology.ubc.ca/~whitlock/Kingfisher/CLT.htm

Chyba měření a CI

Rozložení naměřených hodnot je normálně rozložené a definované svým M a SD.

Proto, když konstruujeme CI, musíme vědět:

- Okolo čeho? Jaký je průměr rozložení?
- Jak nepřesné? Jaká je směrodatná odchylka rozložení (SE?)

Tři klíčové vzorce (z nichž lze vše odvodit)

1. Základní teorém CTT:

$$X = \tau + e$$

- X – pozorované, τ – pravé skóre a e – chyba.

2. Reliabilita $r_{xx'}$ je podíl vysvětleného rozptylu:

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

- Symbol sigma (σ^2) označuje rozptyl.

3. Rozptyl součtu dvou náhodných proměnných A+B má rozptyl:

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

- $\sigma_{AB} = \mathrm{cov}(A, B)$ – kovariance, r_{AB} – jejich korelace (grafická ilustrace)
- Protože $r_{\tau e} = 0$, pak z 1 a 3 vyplývá $\sigma_x^2 = \sigma_\tau^2 + \sigma_e^2$.

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

- reliabilita – podíl vysvětleného rozptylu

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

- reliabilita – podíl vysvětleného rozptylu
- „nereliabilita“ – podíl nevysvětleného rozptylu

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

- reliabilita – podíl vysvětleného rozptylu
- „nereliabilita“ – podíl nevysvětleného rozptylu
- převod z rozptylu na směrodatnou odchylku
  - podíl směrodatné odchylky pravého skóru, která je „způsobena“ chybou

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

- reliabilita – podíl vysvětleného rozptylu
- „nereliabilita“ – podíl nevysvětleného rozptylu
- převod z rozptylu na směrodatnou odchylku
  - podíl směrodatné odchylky pravého skóru, která je „způsobena“ chybou
- převod z podílu (z-skóre) přímo na škálu směrodatné odchylky

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

- reliabilita – podíl vysvětleného rozptylu
- „nereliabilita“ – podíl nevysvětleného rozptylu
- převod z rozptylu na směrodatnou odchylku
  - podíl směrodatné odchylky pravého skóru, která je „způsobena“ chybou
- převod z podílu (z-skóre) přímo na škálu směrodatné odchylky
- směrodatná odchylka chyby měření

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

Ve vzorci je $r_{xx'}$ vysvětlený rozptyl; viz koeficient determinace (PSYb1170).

- Tedy rozptyl měření vysvětlený pravým skórem. Na rozdíl od koeficientu determinace tam není mocnina, protože reliabilita je už přímo „umocněná“.
- $r_{x\tau} = \sqrt{r_{xx'}}$ a tedy $r_{x\tau}^2 = r_{xx'}$

Středová hodnota

Chyba se nepohybuje kolem pozorovaného, ale kolem pravého skóre.

Jaká je nejpravděpodobnější hodnota pravého skóre při určitém pozorovaném skóre x?

O trochu blíže k průměru (protože pravé skóry mají menší rozptyl než pozorované skóry).

Regresní model CTT:

$$\mathrm{E}(T|x) = r_{xx'}x + (1 - r_{xx'})M_x$$

- $\mathrm{E}(T|x)$: očekávané (expected), nejpravděpodobnější pravé skóre.
- $r_{xx'}$: reliabilita; „směrnice“.
- M_x : průměrné skóre; $(1 - r_{xx'})M_x$ je „průsečík“.
- Čím větší reliabilita, tím větší vliv pozorovaného skóre a menší vliv průměru (a naopak).

Směrodatná odchylka pravého skóre: $\sigma_\tau = \sqrt{r_{xx'}}\sigma_x$

Chyba měření (v CTT)

Takto spočítanou chybu měření mohu použít pro konstrukci intervalu spolehlivosti.

$$CI_i = \mathrm{E}(X) \pm z_i \sigma_e$$

- $\mathrm{E}(X)$ = očekávaná hodnota, okolo které interval konstruuji.
- σ_e = chyba měření
- z_i = kvantil normálního rozdělení

Kvantily normálního rozdělení:

- 95% CI: $z_{95\%} \cong 1{,}96$
- 90% CI: $z_{90\%} \cong 1{,}64$
- 80% CI: $z_{80\%} \cong 1{,}28$
- 68% CI: $z_{68\%} \cong 1{,}00$

Shrnutí: Důležité prvky práce s SE

Co je očekávanou hodnotou, okolo které interval konstruuji?

- Pozorované skóre?
- Odhad pravého skóre?
- Nula (pro rozdíl dvou skórů)?

Jak spočítám chybu pro daný účel/diagnostickou otázku?

Jaký odhad reliability nejlépe použiju pro daný účel?

Scénář 1: Standardní chyba měření

Pokud jsme naměřili pozorované skóre X, jaké jiné alternativní X jsme mohli rovněž naměřit?

Slouží pro popis chyby měření a intervalu spolehlivosti jednoho jediného měření.

Velikost chyby:

$$\sigma_e = \sigma_x \sqrt{1 - r_{xx'}}$$

Středová hodnota: odhad pravého skóre

$$\mathrm{E}(T|x) = r_{xx'}x + (1 - r_{xx'})M_x$$

Scénář 2: Chyba odhadu pravého skóre

Pokud jsme naměřili pozorované skóre X, jaká je chyba odhadu pravého skóre τ?

Vzorec je stejný, jen namísto SD pozorovaného skóre použijeme odhad SD pravého skóre:

Velikost chyby:

$$\sigma_{e(\tau)} = \sigma_\tau \sqrt{1 - r_{xx'}} = \sigma_x \sqrt{r_{xx'}} \sqrt{1 - r_{xx'}}$$

Středová hodnota:

$$\mathrm{E}(T|x) = r_{xx'}x + (1 - r_{xx'})M_x$$

Někteří autoři tento postup doporučují, ale potíž s interpretací.

- Zajímá nás chyba na škále použité při konstrukci norem. Zpravidla tedy nepoužitelné.
- Nicméně např. WISC-5[UK] – pro standardizaci na IQ použil právě σ_τ
  - Standardizace $IQ = 15\frac{(X - M_x)}{\sigma_x \sqrt{r_{xx'}}} + 100$ namísto běžného $IQ = 15\frac{(X - M_x)}{\sigma_x} + 100$

Scénář 3: Standardní chyba predikce

Naměřil jsem X. V jakém rozsahu bude ležet příští měření, pokud se úroveň atributu nezmění?

- *„Zlepšil se klient v terapii?“ „Je účinný výukový program?“*

Velikost chyby:

$$\sigma_{pred} = \sigma_x \sqrt{1 - r_{xx'}^2}$$

- $r_{xx'}^2$ - druhá mocnina (test-retest) reliability.
- jde o úpravu $\sigma_{pred} = \sqrt{\sigma_e^2 + \sigma_{e(\tau)}^2}$, tedy rozdíl chyby odhadu pravého skóru a chyby měření

Středová hodnota = očekávaný skór při retestu: odhad pravého skóre:

$$\mathrm{E}(T|x) = r_{xx'}x + (1 - r_{xx'})M_x$$

Scénář 4: Statisticky významný rozdíl

Standardní chyba rozdílu. Rozdíl dvou nezávislých testů jedné osoby; případně rozdíl dvou osob.

Jaká je očekávaná odlišnost v měření dvěma testy?

- „Dosáhla vyššího skóru Anežka nebo Bedřich?“ „Je Cyril vyšší nebo těžší?“
- Musí být ve stejných jednotkách.

Velikost chyby:

$$\sigma_{e(A-B)} = \sqrt{\sigma_{e(A)}^2 + \sigma_{e(B)}^2} = \sigma_{ab}\sqrt{2 - r_{aa'} - r_{bb'}}$$

- Pokud jde o měření jediným testem (dvěma testy se stejnou reliabilitou), lze zjednodušit:

$$\sigma_{e(A-B)} = \sqrt{2}SE = \sigma_x\sqrt{2}\sqrt{1 - r_{xx'}}$$

Středová hodnota:

- Jde o rozdíl a očekávaný rozdíl je zpravidla žádný rozdíl, **proto zpravidla 0**.
- To není úplně pravda; pokud $r_{aa'} \neq r_{bb'}$, pak je střední hodnotou $E(\tau'_A - \tau'_B) = \sqrt{r_{AA'}}(A - M) - \sqrt{r_{BB'}}(B - M)$, ale výsledek bude velmi podobný. Zanedbejte.

Scénář 5: Klinicky významný rozdíl

Liší se dva skóry téhož respondenta více či méně než u „běžných“ respondentů?

- To, že se skóry liší, neznamená, že se liší více, než bychom čekali u náhodně vybraného člověka.
- Klinické hypotézy: „*Rozkolísaný profil schopností...*“, „*Je rozdíl ‚klinicky‘ významný?*“ atd.

Příklad:

- **Statisticky významný rozdíl:** *„Člověk má vyšší váhu než výšku (ve standardních jednotkách, např. IQ skórech)“*.
- **Klinicky významný rozdíl:** *„Člověk má vyšší váhu, než by odpovídalo jeho výšce, je tedy obézní.“*

Scénář 5: Klinicky významný rozdíl

Více postupů. Nejjednodušší používá pouze korelaci a je zcela shodný s postupem pro chybu predikce.

Odhad chyby:

$$\sigma_{A-B} = \sigma_{AB}\sqrt{1 - r_{AB}^2}$$

- r_{AB} je korelace testů A a B, σ_{AB} je směrodatná odchylka obou testů (musí být shodná)

Středová hodnota:

$$\mathrm{E}(B|A) = r_{AB}A + (1 - r_{AB})M_{AB}$$

Scénář 6: Více měření

Lze testovat, zda má klient celkově „rozkolísaný profil“.

- Např.: *„Liší se subtesty ve WAIS-III od celkového IQ více, než bychom čekali?“*
- Analogie F-testu u lineární regrese s více prediktory.

Poskytují jen některé diagnostické metody, není pravidlem.

Technicky vzato není ideální interpretovat „profil“, pokud test celkového rozdílu není signifikantní na zvolené *p*-hladině.

Ruční výpočet je příliš náročný.

Studijní zdroje

Viz interaktivní osnova.

Pracovní verze Diagnostické kalkulačky (bez záruky).

Klasická testová teorie

Classical Test Theory (CTT)

Klasická testová teorie (CTT)

Tři pilíře CTT (Traub, 1997):

- Chyby I. typu, chyba měření jako náhodná veličina, korelace.

Koeficient proti oslabení korelace (Spearman, 1904).

- Vztah reliability, chyby měření a koncept paralelních testů.
- Attenuation formula, $r_{pq}^{*} = \frac{r_{pq}}{\sqrt{r_{pp'} r_{qq'}}}$.

Vývoj CTT byl prakticky ukončen do 60. let: Lord a Novick (1968).

Klasická testová teorie (CTT)

Důležitým impulzem byla Fergusonova komise (1932– 1940).

- Striktní požadavek aditivity (a zřetězení).
- Psychologové zřetězení nedokázali → **CTT není vědeckou teorií měření.**
- Reakcí byla Stevensova „nevědecká“ „operační teorie měření“, která rozšířila definici měření: *„...measurement, in the broadest sense, is defined as the **assignment of numerals to objects and events according to rules.**“* (Stevens, 1946, s. 677). Klíčový pojem je „**matching**“.
  - Ve skutečnosti zjednodušení konsenzu z přírodních věd: „Measurement is a method of *assigning numbers to magnitudes*“ (např. Helmholtz, 1887).

Vývoj CTT byl prakticky ukončen do 60. let: Lord a Novick (1968).

von Helmholtz, H. (1887/1971). An Epistemological Analysis of Counting and Measurement. In R. Kahl (ed.), *Selected Writings of Hermann von Helmholtz*. Wesleyan University Press.
Pro interpretaci doporučuji: Michell, J. (1993). The origins of the representational theory of measurement: Helmholtz, Hölder, and Russell. *Studies in History and Philosophy of Science Part A, 24*(2), 185–206. doi:10.1016/0039-3681(93)90045-I

Měření v přírodních vědách

Existuje nějaký atribut, který opakovaně měříme tím stejným nástrojem.

Každé jedno měření má nějakou chybou, kterou neznáme.

- Jednotlivá měření se pohybují okolo skutečné hodnoty v důsledku náhodné chyby měření.

Výsledkem opakovaných měření je proto rozložení, které použijeme pro odhad skutečné hodnoty:

- **Průměr rozložení:** odhad míry atributu, $\mathrm{E}(x) = \frac{\sum_{i=1}^{N} x_i}{N}$.
  - N – počet měření; x_i – i-tá naměřená hodnota; $\mathrm{E}(x)$ – expected value (průměr, nejpravděpodobnější hodnota příštího měření.
- **Standardní chyba průměru:** odhad standardní chyby měření, $SE = \frac{s_d}{\sqrt{N}}$
  - SE – standardní chyba měření (Standard Error), s_d – výběrová směrodatná odchylka jednotlivých měření.
- Lze využít pro konstrukci CI atd. (za pomoci Studentova t-rozložení).

Předpoklady

Odhad průměru (standardní chyba měření) je přibližně normálně rozložený.

- Centrální limitní teorém: potřebujeme alespoň 30 měření.
- Příklady zde a zde ☺

To v psychologii není možné. Nemůžu člověka měřit 30krát tím stejným testem (vyjma jednoduchých psychofyzikálních úloh).

Kudy z toho ven? Shodná chyba měření pro všechny respondenty.

- Nikoliv *„standardní chybu průměru“* pro každého respondenta zvlášť.

Jednotlivá měření jako paralelní testy.

Paralelní testy

„Dobré“ měření je takové, kdy různí lidé v různých časech dojdou různými nástroji ke stejným naměřeným hodnotám, pokud se míra samotného objektu nezměnila.

Paralelní testy/měření jsou takové, pro které platí:

- A. Pravý skór je v paralelních 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).
- C. Chybový rozptyl je v paralelních testech a pro každý subjekt stejný.
  - Důsledkem je navíc shodný rozptyl pozorovaných skórů obou testů.

Paralelní testy

Korelace paralelních testů je reliabilita: $r_{xx'} = \mathrm{cor}(x, x')$

- To je právě Spearmanův objev.
- Test-retest, paralelní formy, shoda posuzovatelů, split-half...

Původně CTT považovala za paralelní testy pouze jejich výsledek (celkové skóre).

- Způsob konstrukce tohoto skóre je irelevantní.
- Operacionalismus: pravé skóre (a tedy měřený atribut) je definovaný měřením.

CTT tedy chápe reliabilitu jako „stabilitu“ odhadu pravého skóre napříč podmínkami (paralelním testováním).

S postupem času otázka: Jak se celkové skóre vytváří?

- **Položky jako paralelní testy.**

CTT: Paralelní testy

Úrovně paralelnosti položek (založené na faktorové analýze):

$$X_{ip} = i_i + a_i\tau_p + e_{ip}$$

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 (založené na faktorové analýze):

$$X_{ip} = i_i + a_i\tau_p + e_{ip}$$

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 (založené na faktorové analýze):

$$X_{ip} = i_i + a\tau_p + e_{ip}$$

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 (založené na faktorové analýze):

$$X_{ip} = i_i + a\tau_p + e_{ip}$$

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. $a_i = a, \mathrm{var}(e_{ip}) = \mathrm{var}(e)$

- Shodné reziduální rozptyly. V případě binárních položek je shodné s předchozím, $\mathrm{var}(x) = {}^{\mathrm{P}(x)}/_{1-\mathrm{P}(x)}$

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):

$$X_{ip} = i + a\tau_p + e_{ip}$$

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.

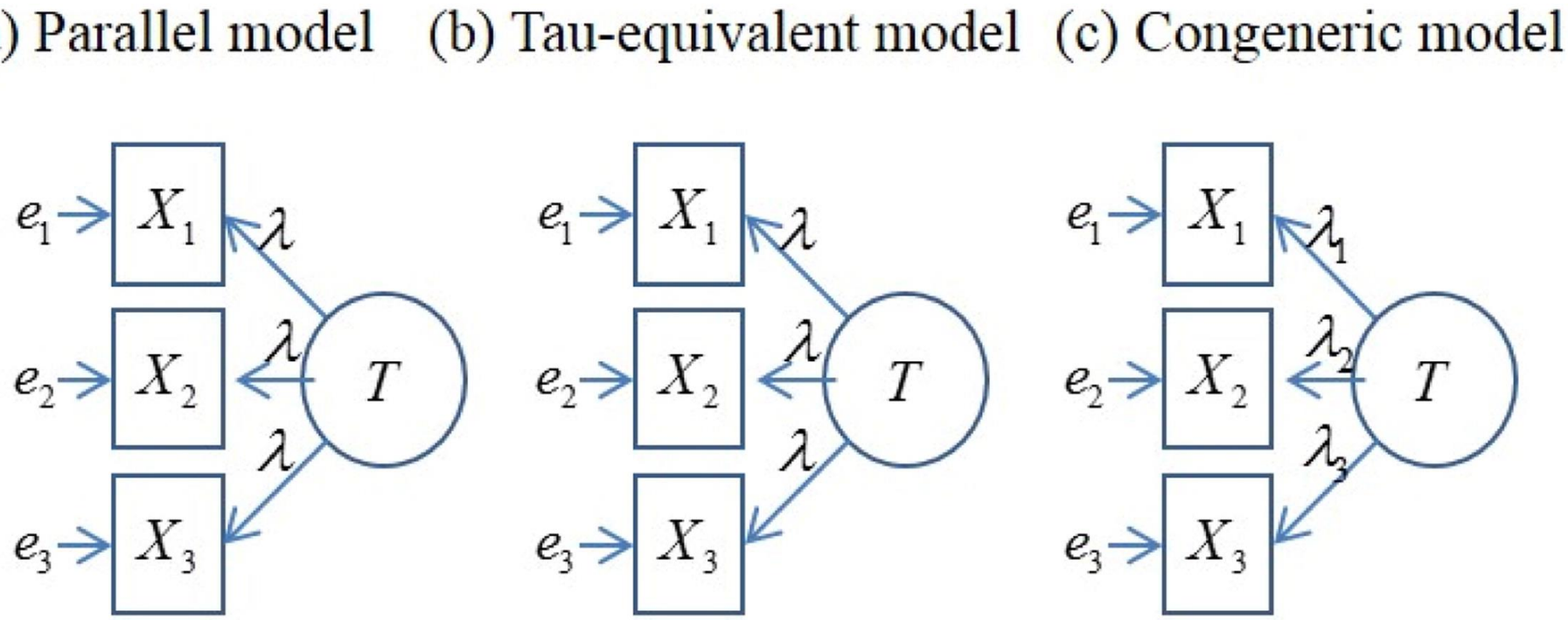
- Shodné reziduální rozptyly.

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

- 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

CTT: Paralelní testy



$$Var(e_1) = Var(e_2) = Var(e_3)$$

Cho, E. (2016). Making Reliability Reliable: A Systematic Approach to Reliability Coefficients. *Organizational Research Methods, 19*(4), 651–682. https://doi.org/10.1177/1094428116656239

Reliabilita

„The term reliability has been used in two ways in the measurement literature. First, the term has been used to refer to the ***reliability coefficients of classical test theory****, defined as the* ***correlation between scores*** *on two equivalent forms of the test, presuming that taking one form has no effect on performance on the second form.*

Second, the term has been used in a more general sense, to refer to the ***consistency of scores across replications*** *of a testing procedure, regardless of how this consistency is estimated or reported (e.g., in terms of standard errors, reliability coefficients per se, generalizability coefficients, error/tolerance ratios, item response theory (IRT) information functions, or various indices of classification consistency).“*

(AERA, 2014, s. 33)

Stačí to?

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Dvě pojetí reliability

Stabilita měření.

- Bez ohledu na to, jaký je „význam“ měření.

Vysvětlený rozptyl.

- Vysvětlený rozptyl **čím**?
- Co považujeme za pravé skóre?

Dvě pojetí reliability

1. Dimension-free reliability (důraz na korelaci paralelních testů)

- Odhad vztahu (korelace) dvou paralelních měření týmž testem bez ohledu na to, co test měří.
- split-half, alfa, celková omega, *glb*

2. Model-based reliability (důraz na vysvětlený rozptyl)

- Odhad vztahu (vysvětleného rozptylu) měřeného atributu a pozorovaného skóru.
- Rodina koeficientů omega (McDonaldova hierarchická omega).

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
- Cho, E. (2016). Making Reliability Reliable: A Systematic Approach to Reliability Coefficients. *Organizational Research Methods*, *19*(4), 651–682. doi:10.1177/1094428116656239

Systematický přístup k reliabilitě

Table 3. Names of Reliability Coefficients Currently Used in the Literature.

	Unidimensional		Multidimensional
	Split-Half	General	General
Parallel	Spearman–Brown formula	Standardized alpha	(Not yet published)
Tau-equivalent	Flanagan–Rulon formula Flanagan formula Rulon formula Guttman's λ_4	Cronbach's alpha Coefficient alpha Guttman's λ_3 Hoyt method KR-20	Stratified alpha
Congeneric	Raju (1970) coefficient Angoff–Feldt coefficient Angoff coefficient	Composite reliability Construct reliability Congeneric reliability Omega Unidimensional omega Raju (1977) coefficient Classical congeneric reliability coefficient	Omega Omega total McDonald's omega Multidimensional omega

Cho (2016)

Systematický přístup k reliabilitě

Table 4. Names and Notations of Reliability Coefficients Suggested in This Study.

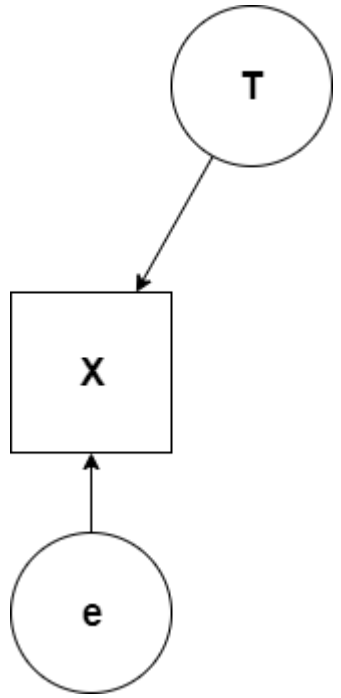
	Unidimensional		Multidimensional
	Split-Half	General	General
Parallel	Split-half parallel reliability (ρ_{SP})	Parallel reliability (ρ_P)	Multidimensional parallel reliability (ρ_{MP})
Tau-equivalent	Split-half tau-equivalent reliability (ρ_{ST})	Tau-equivalent reliability (ρ_T)	Multidimensional tau-equivalent reliability (ρ_{MT})
Congeneric	Split-half congeneric reliability (ρ_{SC})	Congeneric reliability (ρ_C)	Bifactor model Bifactor reliability (ρ_{BF}) Second-order factor model Second-order factor reliability (ρ_{SOF}) Correlated factors model Correlated factors reliability (ρ_{CF})

Cho (2016)

Potíž 1: Spodní hranice reliability

Lower-bound of reliability.

Zpravidla předpokládáme, že unikátní rozptyl položek je chyba.



Potíž 1: Spodní hranice reliability

Lower-bound of reliability.

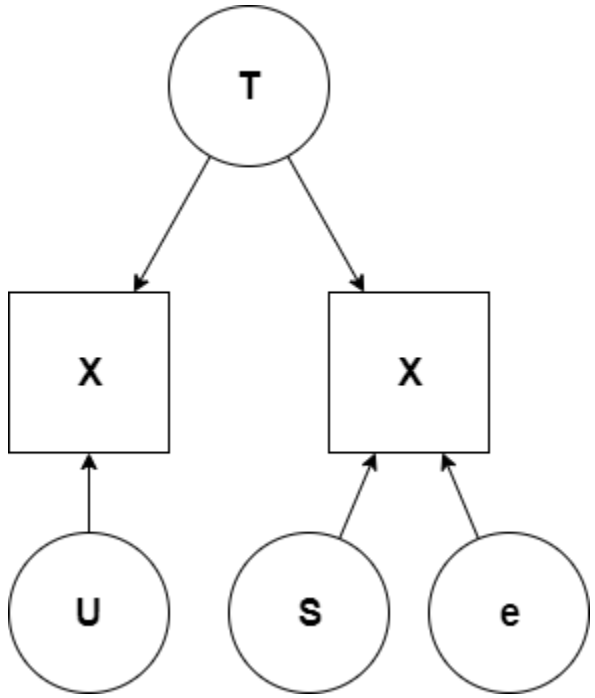
Zpravidla předpokládáme, že unikátní rozptyl položek je chyba.

Unikátní rozptyl ale lze rozdělit na:

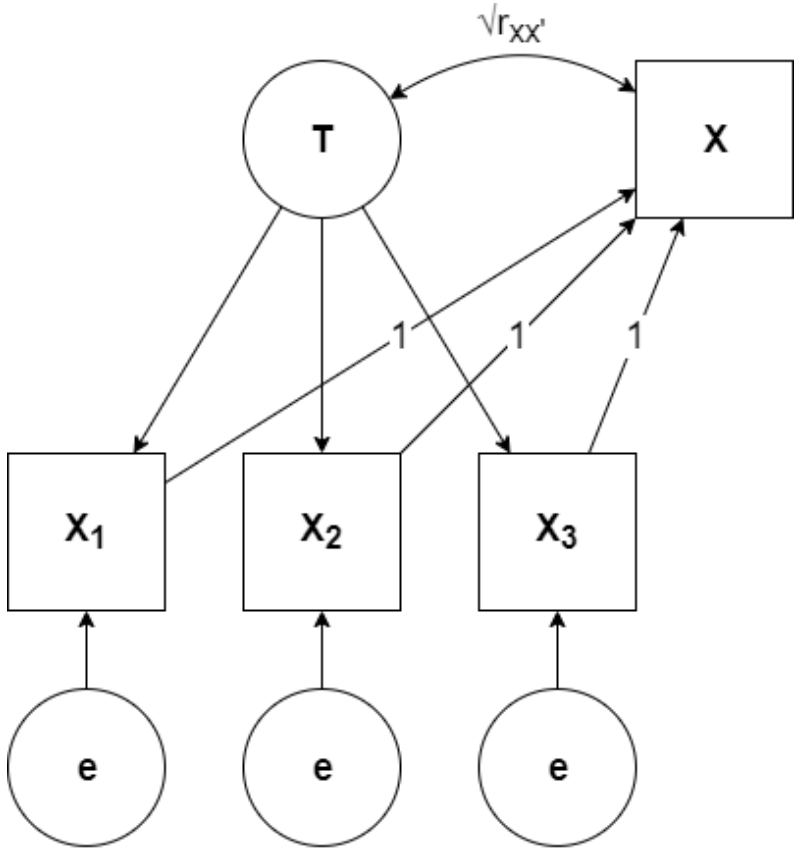
- specifický (systematický pro daného člověka)
- chybový (náhodný)

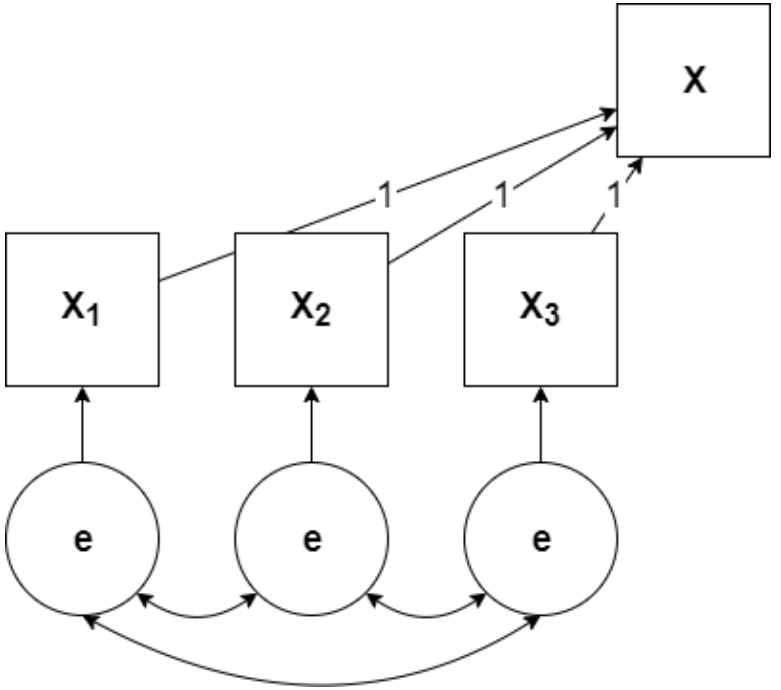
Tyto složky ale nelze oddělit při jediné administraci testu.

- V longitudinálních SEM modelech korelovaná rezidua v čase.

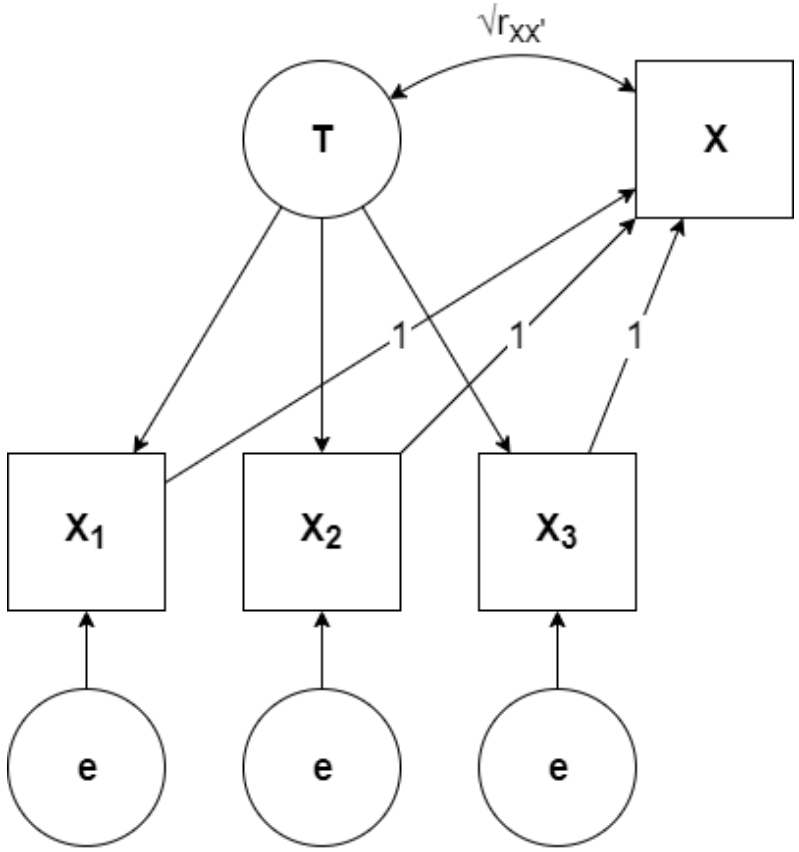


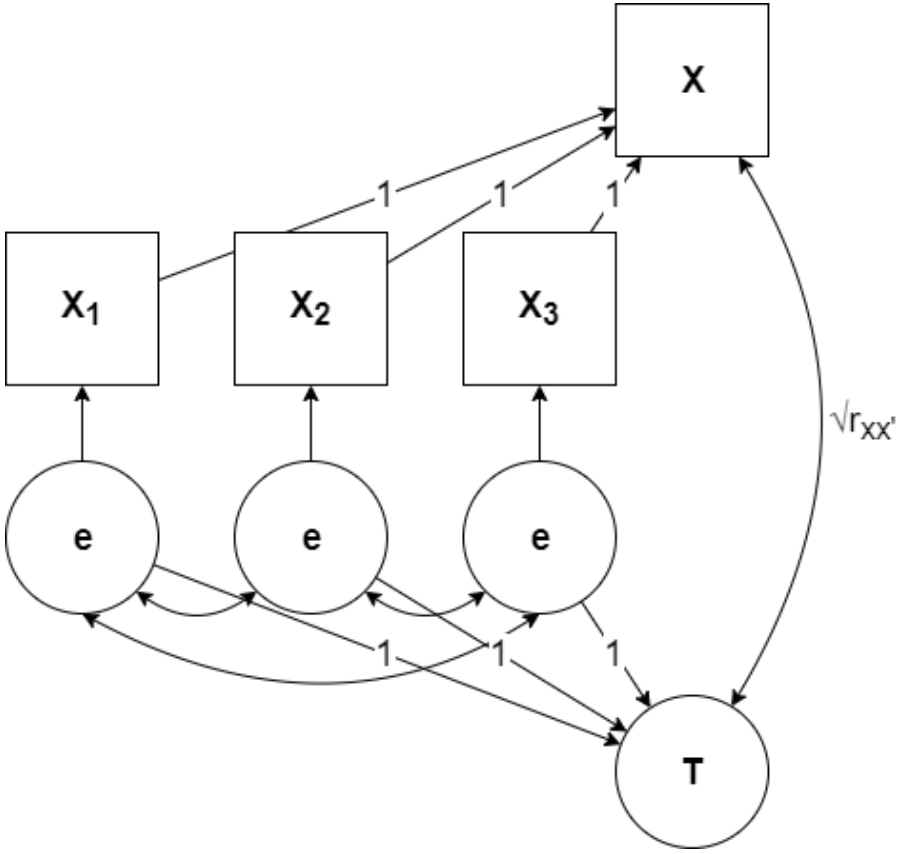
Potíž 2: Formativní vs. reflektivní model





Potíž 2: Formativní vs. reflektivní model





Split-half

Reliabilita jako stabilita.

Problémy se split-half:

- Nelze ověřit předpoklady paralelnosti.
- Test je zkrácený na polovinu.
- Existuje velké množství rozdělení testu na dvě poloviny.
  - Různá rozdělení → různé odhady.

Split-half

SPEARMAN-BROWNŮV PŘÍSTUP

Spearmanův-Brownův věštecký vzorec:

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

- N – změna délky testu, v případě split-half N=2:

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

Předpoklad: **paralelní poloviny**.

- Při nedodržení příliš „optimistický“, může nadhodnocova nebo podhodnocovat.

GUTTMANOVA λ_4

Guttman (1945) publikoval λ_{1-6}:

$$\lambda_4 = \frac{4\sigma^2_{pq}}{\sigma^2_x}$$

- σ^2_{pq} – kovariance polovin testu
- $\sigma^2_x = \sigma^2_p + \sigma^2_q + 2\sigma^2_{pq}$ – rozptyl celého testu.

$\lambda_4 = \alpha$ (ve dvoupoložkovém testu)

- **tau-ekvivalentní poloviny** (jinak podhodnocuje)
- Proto je λ_4 dnes chápána jako maximalizovaná split-half pomocí nejlepšího možného rozdělení.

„Příliš dobré rozdělení“→ na malých vzorcích nadhodnocuje.

Pokud je kovariance větší než kterýkoli z rozptylů: hrubé podhodnocení.

Založeno na jediné korelaci → nepřesný odhad reliability.

Split-half: Nestejné poloviny

Spearmanův-Brownův i Guttmanův přísup předpokládá stejně dlouhé poloviny testu.

Odvozeno z SB-vzorce (při stejné délce by poloviny byly paralelní):

- Horstova (1951)[1]: $r_H = \frac{r_{12}\sqrt{r_{12}^2+4\pi_1\pi_2(1-r_{12}^2)}-r_{12}^2}{2\pi_1\pi_2(1-r_{12}^2)}$, kde π_1 a π_2 jsou délky polovin testu.

Odvozeno z Guttmanovy λ_4 (při stejné délce by poloviny byly tau-ekvivalentní):

- Raju (1977): $\beta = \frac{\sigma_{12}}{\pi_1\pi_2\sigma_x^2}$
- Délku polovin lze odhadnout na základě jejich rozptylu jako $\pi_1 = \frac{\sigma_1^2+\sigma_{12}}{\sigma_x^2}$, $\pi_2 = \frac{\sigma_2^2+\sigma_{12}}{\sigma_x^2}$, což lze dosadit:
- Angoffův-Feldtův koeficient (1953, 1975): $r_{AF} = \frac{4\sigma_{12}}{\sigma_x^2-\frac{\left(\sigma_1^2-\sigma_2^2\right)^2}{\sigma_x^2}}$

[1] Horst (1951) má chybu ve vzorci 2, pro korektbní vzorec viz např. Warrense (2016).

Cronbachovo alfa (Guttmanova λ_3)

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

- σ_i^2 – rozptyl položky *i*, $\sum_{i=1}^{k} \sigma_i^2$ je diagonála var-kovar matice (unikátní rozptyl položek = chyba)
- σ_x^2 – rozptyl celého testu, tedy suma var-kovar matice (sdílený rozptyl položek)
- k – počet položek (ne celý unikátní rozptyl je chybou, proto korekce $\frac{k}{k-1}$, aby reliabilita mohla být 1)
- V případě binárních položek je výsledek shodný s výpočetně jednodušším KR-20.

Předpoklady:

- Tau-ekvivalentní položky (při nedodržení je korekce $\frac{k}{k-1}$ nedostatečná → podhodnocení reliability).
- Jednodimenzionalita (nadhodnocení i podhodnocení dle typu).
- Alfa není ukazatelem jednodimenzionality (viz např. Marko, 2016).

Výhody: Přesný odhad (ve srovnání se split-half), tradice.

Varianty koeficientu alfa

Standardizované alfa.

- Pro výpočet použita korelační matice → reliabilita součtu standardizovaných položek.
- Použitelné v případě položek s rozdílnou odpověďovou škálou, tedy i pozorovaným rozptylem a výrazným narušením předpokladu tau-ekvivalence.

Ordinální alfa (Zumbo, Gadermann, Zeisser, 2007)

- Alfa spočítané nad maticí polychorických korelací.
- Zcela jiný význam, není použitelné pro běžnou praxi.
- Není srovnatelné s jinými odhady reliability (viz např. Chalmers, 2017).

Stratifikované Cronbachovo alfa

Nejjednodušší odhad reliability součtu subtestů – Cronbach (1965):

$$\alpha_{strat} = 1 - \frac{\sum_{i=1}^{k}\left[\omega_i^2 \sigma_i^2 (1 - r_{ii'})\right]}{\sigma_Z^2}$$

- ω_i „váha“ testu *i*
- σ_i^2 rozptyl testu *i*
- $r_{ii'}$ reliabilita testu *i*
- Pro výpočet stačí kovarianční matice a alfy subtestů.

Předpokladem je nejen tau-ekvivalence položek v testech, ale i tau-ekvivalence testů.

- A nekorelované chyby měření testů.

Např.: *„Jaká bude test-retest korelace celkového IQ skóre, pokud jsou obě měření paralelní?“*

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 přímo započítány).

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

- vysvětlený rozptyl
- chybový rozptyl
- celkový rozptyl

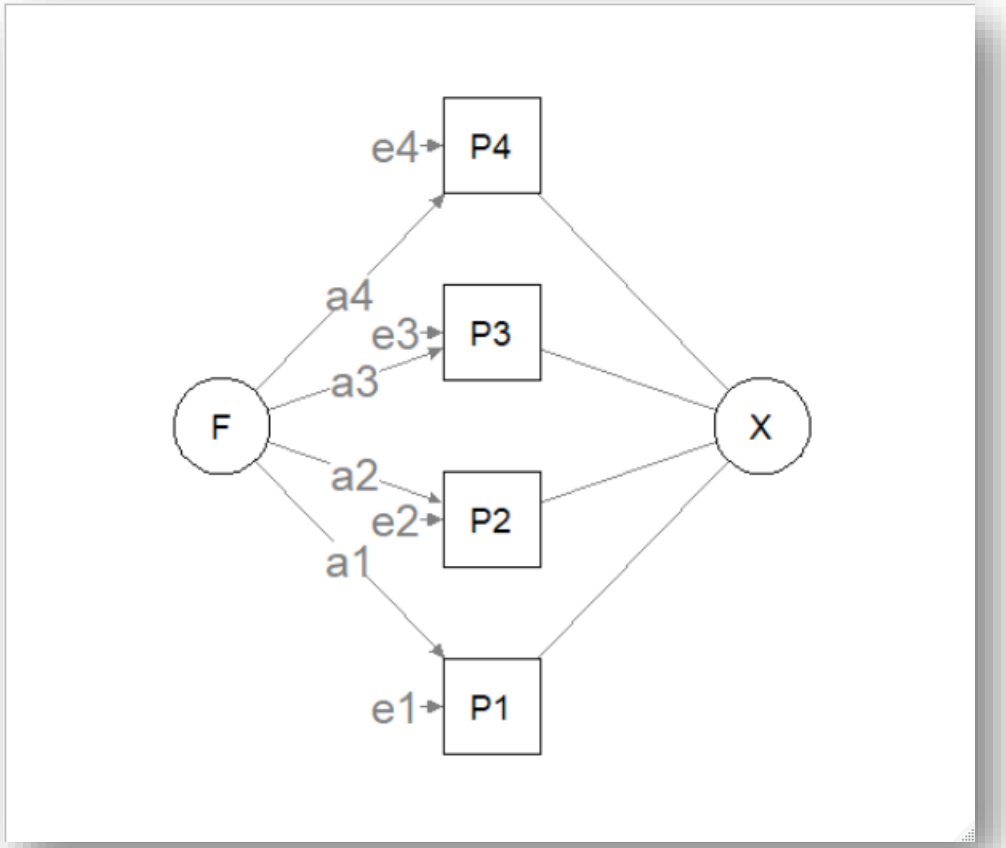
Bez předpokladu tau-ekvivalence (rozdílné faktorové náboje jsou zohledněny).

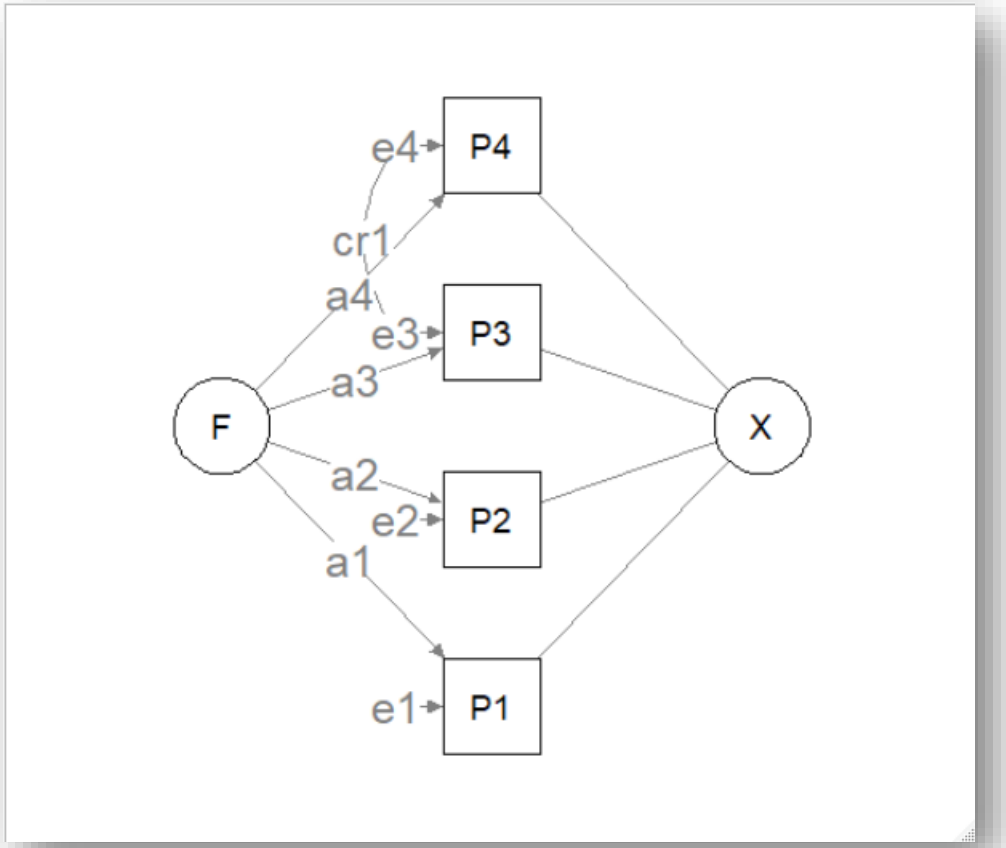
Model-based reliabilita: **omega**

Použití koeficientu omega nás nutí zamyslet se, co je pravým skóre.

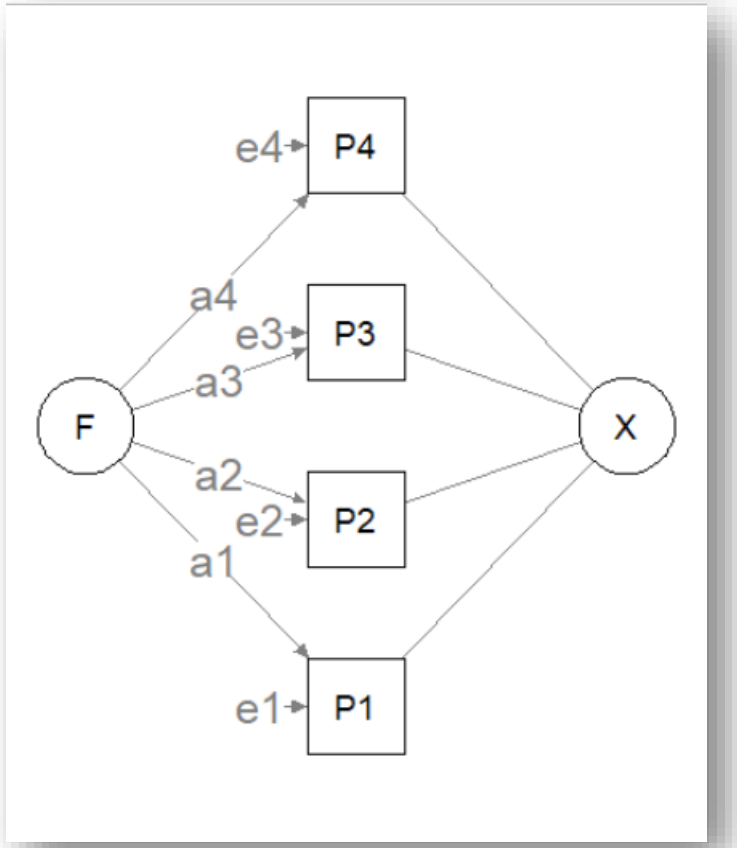
Co je to, co chceme měřit?

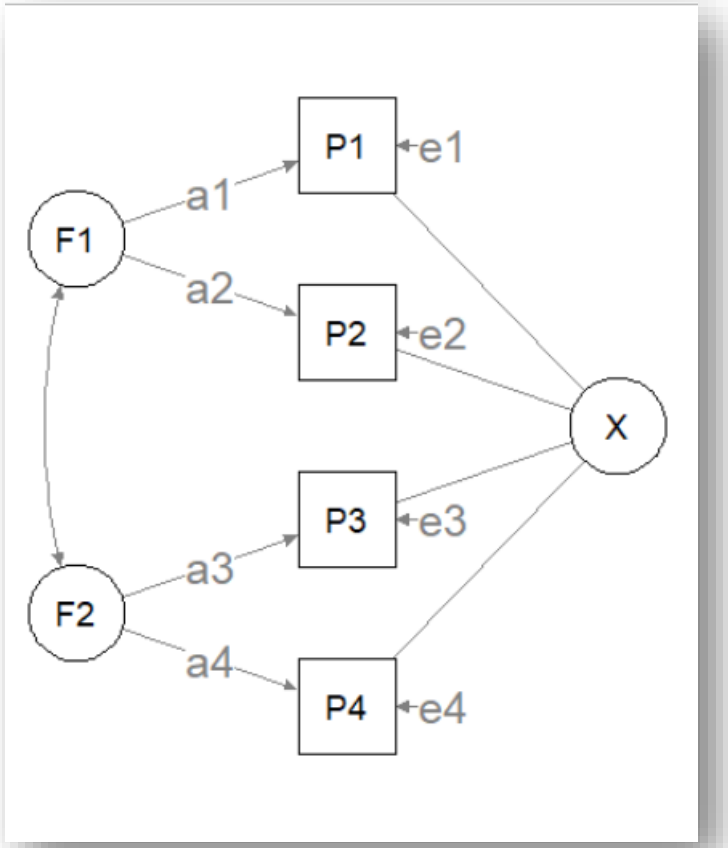
Omega: Multidimenzionalita

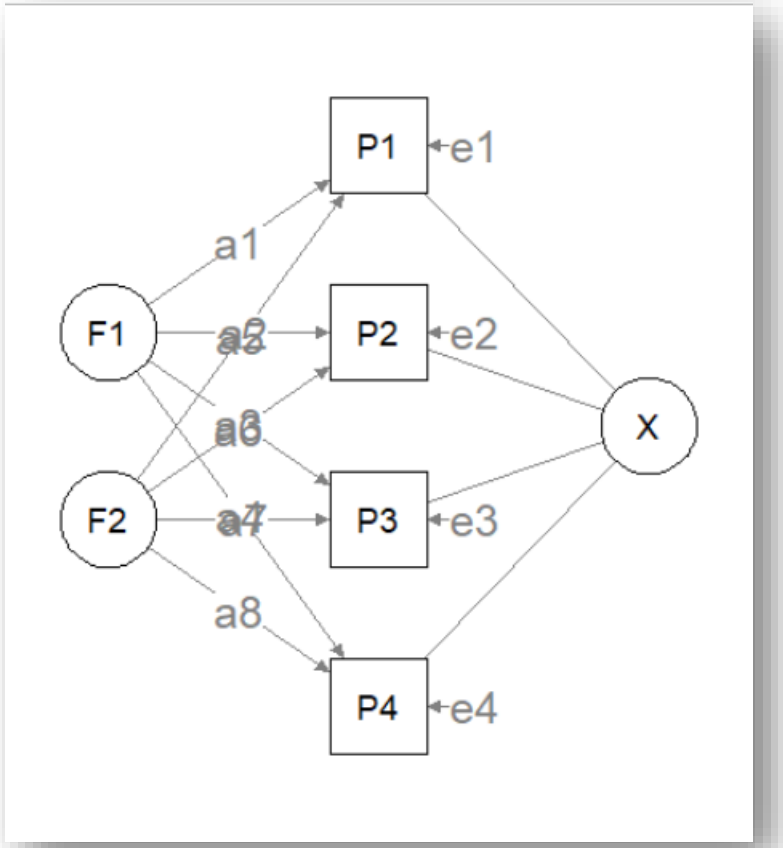




Omega: Multidimenzionalita







Omega: Multidimenzionalita

Hierarchická omega (omega hierarchical):

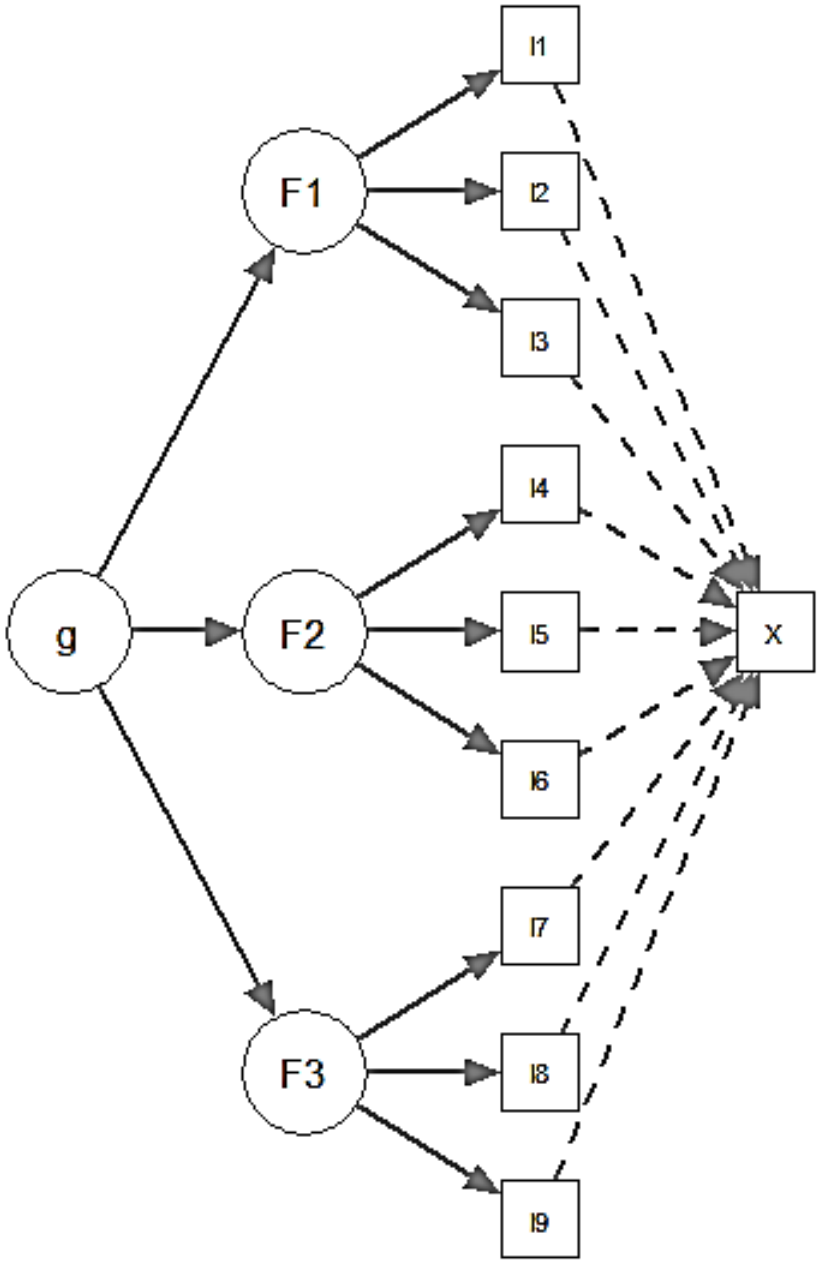
- Rozptyl součtu položek vysvětlený daným faktorem.
- V případě faktoru druhého řádu (g) jsou specifické rozptyly faktorů prvního řádu považovány za chybu.
- **Model based reliabilita:** velmi záleží na definici modelu.

Celková omega (omega total):

- Rozptyl součtu položek vysvětlený všemi faktory prvního řádu.
- Odhad test-retest reliability součtu položek, pokud se míra žádného z atributů nezmění.

Explorační omega (Revelle):

- Celková omega spočítaná na základě EFA.
- omega funkce v psych balíčku v R.



Přehled dalších FA koeficientů

Revellova β (1978): Nejnižší podíl rozptylu, který lze vysvětlit jediným společným faktorem.

- $\beta = \frac{k^2 \bar{\sigma}_{ij}}{\sigma_x^2}$, kde $\bar{\sigma}_{ij}$ je průměrná kovariance napříč dvěma nejhůře rozdělenými polovinami testu.

Bentlerův koeficient *glb* (Greatest Lower-Bound of reliability, 1980):

- Dimension-free vnitřní konzistence.
- Princip: odhad ω_{tot} pro tolik faktorů, kolik jich nevede k negativnímu reziduálnímu rozptylu žádné z položek.
- $\rho_{glb} = 1 - \max \frac{1' \Psi 1}{1' \Sigma 1}$, s pozitivně semi-definitní maticí $(\Sigma - \Psi)$ (kde Σ je pozorovaná matice, Ψ reziduální matice a 1 je jednotková matice.

SW implementace

Pozor: omega v JASPu a JAMOVI je dobrým ukazatelem jen tehdy, pokud jednodimenzionální model sedí na data.

Balíček `psych` v R (funkce `splitHalf`, `omega`, `glb.fa`).

- Pozor: funkce `omega` defaultně využívá korelační, nikoliv kovarianční matici.

Funkce `reliability` v `semTools` balíčku odhadne reliabilitu `lavaan` modelu.

Určitost faktorových skórů

Factor score determinacy.

Koeficienty omega pracují se součtem položek (všechny položky mají váhu 1).

Občas pracujeme s odhady faktorových skórů.

- Vážený průměr všech položek; váha je spočítaná na základě f. nábojů a reziduálních rozptylů.
- $C = \Sigma_y \Lambda_y^T (\Lambda_y \Sigma_y \Lambda_y^T + \Theta_y)^{-1}$ maticový vzorec výpočtu, není podstatné.

Výhody: Vyšší reliabilita (váhy položek jsou optimálně zvolené).

Nevýhody: Sample dependency (zvláště u malých vzorků nepřesný odhad parametrů FA modelu).

Factor score determinacy (FSD) = podíl rozptylu odhadu faktorového skóre vysvětlený faktorem.

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

Standardní chybu (SE) rozdílu lze spočítat s pomocí SD a SE vpravo, nebo prostřednictvím vzorce.

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

- $r_{xx'}$, $r_{yy'}$ – reliability testů x a y
- r_{xy} – korelace testů x a y
- **r_{x-y} – reliabilita rozdílu**
- SD_{x-y} – SD rozdílu
- SE_{x-y} – standardní chyba rozdílu
- $CI_{95\%}$ – šířka 95% intervalu spolehlivosti

$r_{xx'}$	$r_{yy'}$	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

Kompozitní reliabilita

Srovnání reliability rozdílu a kompozitní reliability (stratifikovaná Cronbachova alfa).

Je evidentní, že korelace testů má opačný vliv na výslednou reliabilitu.

$r_{xx'}$	$r_{yy'}$	r_{xy}	r_{x-y}	r_{x+y}
0,7	0,8	0	**0,75**	**0,75**
0,7	0,8	0,2	**0,69**	**0,79**
0,7	0,8	0,4	**0,58**	**0,82**
0,7	0,8	0,6	**0,38**	**0,84**
0,7	0,7	0,6	**0,25**	**0,81**
0,9	0,9	0,8	**0,50**	**0,94**
0,9	0,9	0,45	**0,82**	**0,93**
0,6	0,6	0,5	**0,20**	**0,73**
0,7	0,7	0,65	**0,14**	**0,82**

Doporučení na závěr

Reliabilita čeho?

Pravého skóre?

Stabilita skóre napříč (jakými?) podmínkami?

Reliabilita není jedna.