MUNI FSS Katedra psychologie

Seminar in psychometrics (Charles University):

Reliability models in Classical Test Theory and latent trait theories

Hynek Cígler | cigler@fss.muni.cz | 23. 3. 2021

Katedra psychologie a Institut pro výzkum dětí, mládeže a rodiny
Fakulta sociálních studií, Masarykova univerzita

Obsah přednášky

Východiska a základní teorémy klasické testové teorie

Koncept paralelních testů

Zdroje chyby měření

Různá pojetí reliability

Přehled koeficientů, jejich předpoklady

Příklady výpočtů

Measuring children's height and weight, The University of Iowa, ca. 1920

https://www.flickr.com/photos/uiowa/8037651744

Model měření klasické testové teorie (CTT)

Charles Spearman (1904): Základy klasické testové teorie.

- Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Louis Guttman (1945): Systematičtější přehled odhadů reliability.

- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892

Lee J. Cronbach (1951): Konceptualizace vnitřní konzistence

- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, Rajaratnam & Gleser (1963): Zobecnění CTT pro více zdrojů chyby

- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Lord a Novick (1968): Ukončení hlavního vývoje CTT.

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Cronbach† a Shavelson (2004): **Jak jste mě mohli tak špatně pochopit?!**

- Cronbach, L. J., & Shavelson, R. J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. https://doi.org/10.1177/0013164404266386

Model měření CTT

Základní teorém CTT:

$$X = \tau + e$$

- Chyba je stejná pro všechna měření a všechny subjekty.
- Lineární vztah X a τ je shodný pro všechna měření a všechny subjekty.
- Chyba je normálně rozložená náhodná veličina.

Protože $\mathrm{cov}(\tau, e) = 0$:

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_e^2 + 2\mathrm{cov}(\tau, e)$$

Reliabilita je potom definovaná jako:

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

Model měření CTT

CTT imituje měření v přírodních vědách pomocí tzv. **paralelních testů**:

- A. Pravý skór je ve všech testech a pro každý měřený subjekt stejný: $\tau = \mathrm{E}(x) = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} x_i}{n}$.
- B. Rozptyl pravých skórů je v obou testech stejný (důsledek A).
- C. Chybový rozptyl je v obou testech a pro každý subjekt stejný.
- D. Shodný rozptyl pozorovaných skórů obou testů (důsledek A a C).

Reliabilita je definovaná jako vysvětlený rozptyl $\tau \to X$ (a analogicky $X \to \tau$).

- A tedy $\mathrm{cor}(\tau, X) = \sqrt{r_{xx'}}$.

Lze snadno dovodit:

- 1. Druhá mocnina reliability je proto vysvětlený rozptyl $X \to \tau \to X'$.
- 2. **Reliabilita je tedy korelace paralelních testů** a proto se zpravidla značí $r_{xx'}$.
- To platí, pokud (a.) X a X' jsou paralelní testy, **a (b.) chyby měření jsou nezávislé** ($\mathrm{cov}(e, e') = 0$).

Model měření CTT

Protože hodnota τ není známa, nelze odhadovat přímo $\sqrt{\mathrm{cor}(\tau, X)}$.

Tradiční odhady reliability jsou proto založeny na **korelaci paralelních testů**.

- Reliabilita paralelních forem, test-retest, shoda posuzovatelů...
- Paralelní testy jsou ovšem dostupné jen v minoritě případů.

Vnitřní konzistence – test X rozdělíme na n paralelních „subtestů“ x_i (položky, superpoložky, půlky):

$$X = \sum_{i=1}^{n} x_i$$

Pak platí

$$\sigma_X^2 = \mathbf{1}'\Sigma\mathbf{1}$$

- Rozptyl testu X je roven součtu prvků kovarianční matice Σ jednotlivých složek x_i.
  - Alternativně $X = \sum_{i=1}^{n} w_i x_i$ a $\sigma_X^2 = \mathbf{W}'\Sigma\mathbf{W}$, kde w_i jsou váhy jednotlivých testů a $\mathbf{W}$ je vektor těchto vah.

Potíže modelu měření CTT

Takto definované měření $X = \sum_{i=1}^{n} x_i$ má **značné množství potíží**.

- 1. Vztahy subtestů x_i, x_j zpravidla nejsou lineární, a nelze je sčítat.
- 2. Vztah celého testu X a subtestů x_i tedy rovněž nejsou lineární.
- 3. Proměnné X a x_i zpravidla nejsou normálně rozložené, intervalové a spojité.
- 4. Psychické atributy (a tedy i τ) nejsou extensivní, ale naopak intensivní povahy.
  - Součet dílčích měření x_i atributu τ tak nedává věcně smysl (aka součet teplot?).

V 1. polovině 20. stol. značné množství pokusů obtíže vyřešit.

- Např. Guttmanova, Thurstonova, Likertova škála a další.
- Nejjednodušší součet ovšem funguje jako uspokojivý „estimátor“.
- Částečně vyřešeno modely faktorové analýzy, IRT a zejm. teorií spojitého měření.

Míra paralelnosti

Paralelní testy v psychologii prakticky neexistují. Proto mluvíme o „míře paralelnosti“.

Model faktorové analýzy:

$$x_{ip} = \lambda_i \eta_p + \nu_i + \varepsilon_{ip}$$

- kde λ_i –faktorový náboj, ν_i – intercept, $\varepsilon_{ip} \sim \mathrm{N}(0, \vartheta_i)$ – chyba měření a η_p – hodnota faktoru osoby p.

Kongenerické testy: položky jsou vybrané ze stejné domény, měří stejný rys, jsou lokálně nezávislé (a nebo jsou tyto závislosti adekvátně modelovány).

Tau-ekvivalentní testy: + shodné faktorové náboje ($\lambda_i = \lambda$ pro všechny položky)

Paralelní testy: + shodné reziduální rozptyly ($\vartheta_i = \vartheta$ pro všechny položky)

Striktně paralelní: + shodné obtížnosti ($\nu_i = \nu$ pro všechny položky).

Konceptualizace chyby měření a reliability v klasické testové teorii

Zdroje chyby měření (absolutní vs. relativní zobecnění):

- *Pochází z teorie zobecnitelnosti:*
- Nekonzistence položek (neshoda v pořadí): nízké faktorové náboje.
- Rozdíly v obtížnosti (neshoda v průměru): kriteriální měření, indukce na obsahové univerzum.

Pojetí reliability:

- *Pochází z faktorové analýzy:*
- Korelace paralelních testů (dimension-free reliability).
- Vysvětlený rozptyl (model-based reliability).

Absolutní vs. relativní zobecnění

Teorie zobecnitelnosti; dva typy D-studie: absolutní vs. relativní.

V jednofasetovém designu interakcí položek (I) a osob (P) $P \times I$ pozorujeme tři zdroje variability:

- σ_p^2 – rozptyl pravých skórů; σ_i^2 – rozptyl obtížností položek; a $\sigma_{e,p:i}^2$ – rozptyl interakcí (reziduální rozptyl).

Relativní zobecnění: neusuzuji na univerzum všech možných položek.

- Jediným zdrojem „chyby“ je tak rozdíl seřazení osob napříč položkami, tedy $\sigma_{e,p:i}^2$.
- Koeficient zobecnitelnosti G: $r_{xx'} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{e,p:i}^2}{n_i}}$

Absolutní zobecnění: usuzuji na univerzum.

- Hraje roli, zda jsem vybral lehké či snadné položky. Rozptyl položek σ_i^2 je zdrojem chyby.
- Koeficient dependability Φ: $r_{xx'} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{e,p:i}^2}{n_i} + \frac{\sigma_i^2}{n_i}}$

Absolutní vs. relativní zobecnění

Teorie zobecnitelnosti; dva typy D-studie: absolutní vs. relativní.

V jednofasetovém designu interakci položek (I) a osob (P) $P \times I$ pozorujeme tři zdroje variability:

- σ_p^2 – rozptyl pravých skórů; σ_i^2 – rozptyl obtížností položek; a $\sigma_{e,p:i}^2$ – rozptyl interakcí (reziduální rozptyl).

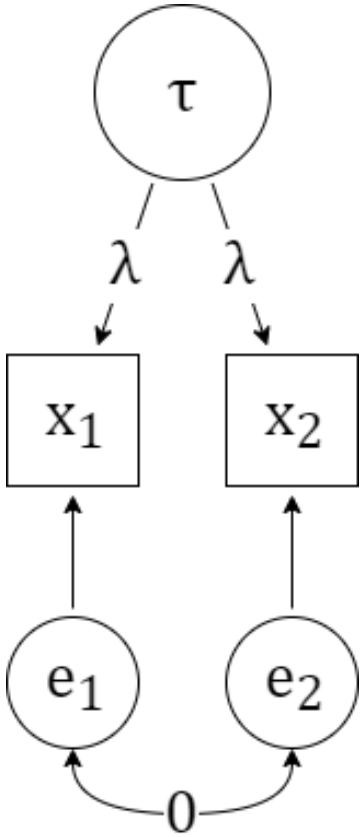
Relativní zobecnění: neusuzuji na univerzum všech možných položek.

- Jediným zdrojem „chyby“ je tak rozdíl seřazení osob napříč položkami, tedy $\sigma_{e,p:i}^2$.
- Koeficient zobecnitelnosti G: $r_{xx'} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{e,p:i}^2}{n_i}}$

Absolutní zobecnění: usuzuji na univerzum.

- Hraje roli, zda jsem vybral lehké či snadné položky. Rozptyl položek σ_i^2 je zdrojem chyby.
- Koeficient dependability Φ: $r_{xx'} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{e,p:i}^2}{n_i} + \frac{\sigma_i^2}{n_i}}$

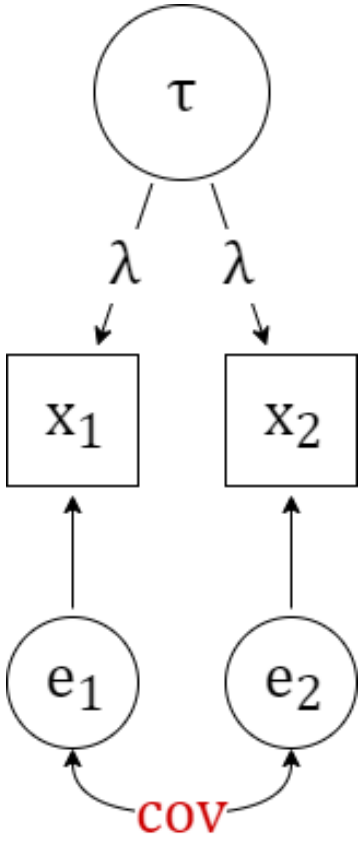
Vysvětlený rozptyl vs. korelace paralelních testů



Model 1:
Nekorelovaná rezidua, jednoduchá struktura.

$$r_{xx'} = \mathrm{cor}(x_1, x_2) = \lambda^2$$

Korelace paralelních testů je rovna rozptylu vysvětlenému pravým skóre

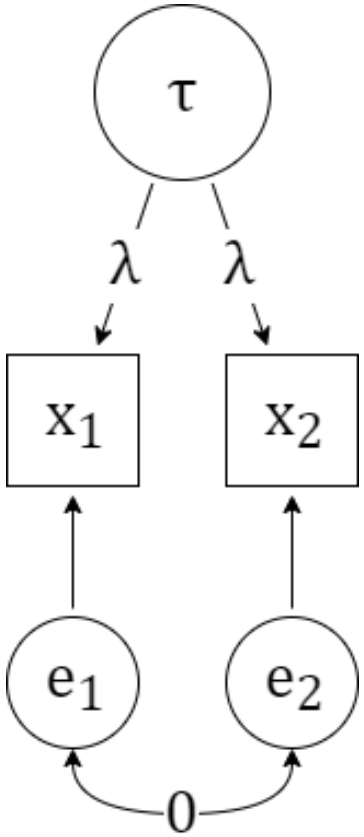


Model 2:
Korelovaná rezidua

$$r_{xx'} = \mathrm{cor}(x_1, x_2) = \lambda^2 + \mathrm{cor}(e_1, e_2)$$

Korelace paralelních testů je navýšena (nebo snížena) reziduální kovariancí oproti vysvětlenému rozptylu.

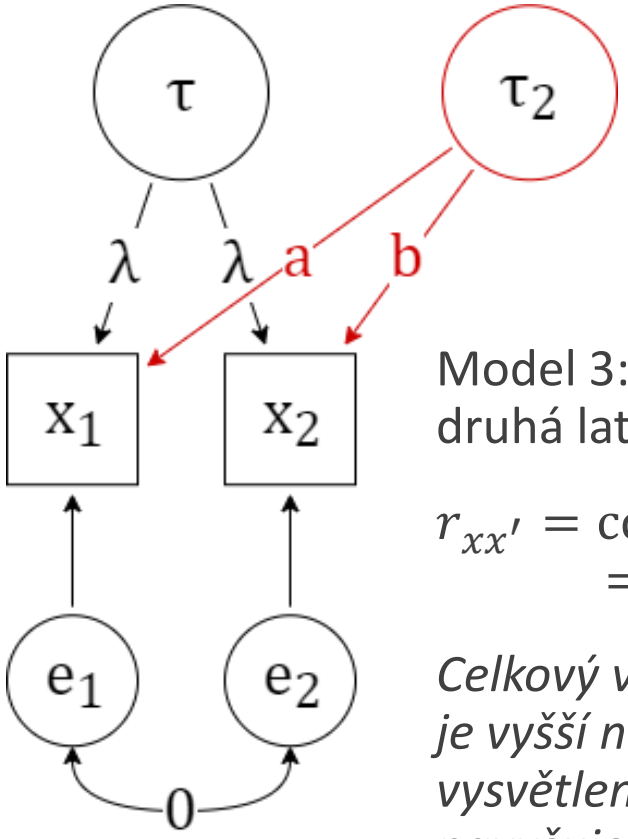
Vysvětlený rozptyl vs. korelace paralelních testů



Model 1:
Nekorelovaná rezidua.

$r_{xx'} = \mathrm{cor}(x_1, x_2) = \lambda^2$

Korelace paralelních testů je rovna rozptylu vysvětlenému pravým skóre



Model 3:
druhá latentní proměnná

$$r_{xx'} = \mathrm{cor}(x_1, x_2) = \lambda^2 + ab$$

Celkový vysvětlený rozptyl je vyšší než rozptyl vysvětlený jen τ, což navyšuje korelaci testů.

Vysvětlený rozptyl vs. korelace paralelních testů

Dimension free reliabilita.

- Odhad korelace paralelních testů.
- Bez ohledu na model měření.
- Tradiční odhady reliability.
- Operacionalismus, antirealismus.

Model-based reliabilita.

- Odhad vysvětleného rozptylu... čím?
- Nutí nás definovat měřený atribut.
- Svázaná s faktorovou analýzou.
- Realismus.

Společné předpoklady:

- Unikátní rozptyl položek je chybový rozptyl (neobsahuje systematickou složku).
- Pokud neplatí (a obsahuje systematický specifický rozptyl), pak je vnitřní konzistence spodním odhadem reliability.
- **Lower-bound of reliability.**
- Tohle je důvodem, proč je někdy test-retest reliabilita vyšší než vnitřní konzistence, přestože předpokládáme fluktuaci rysu v čase.

Greatest-lower bound of reliability

- Viz např. Sijtsma (2008), Revelle a Zinbarg (2009), Bentler (2009) či Cho (2016).

Cronbachova alfa

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

Odhad rozptylu vysvětleného obecným faktorem...
... nadhodnocený reziduálními kovariancemi položek.

Nebo odhad korelace paralelních testů...
... podhodnocený předpokladem jednodimenzionality (jeden obecný faktor).

Další předpoklady: **tau-ekvivalence položek.**

- Všechny položky mají stejnou váhu.
- Při porušení podhodnocuje.

Varianty koeficientu alfa

Standardizovaná Cronbachova alfa: řeší předpoklad tau-ekvivalence $\alpha = \frac{k^2 \overline{\sigma}_{ij}}{\sigma_x^2}$

- Odhad s pomocí korelační, nikoli kovarianční matice.
- Jde o reliabilitu součtu standardizovaných položek.
- Ovšem v případě výrazného porušení tau-ekvivalence méně zkreslený estimátor reliability.
- Využití v případě výrazně rozdílných rozptylů (obtížnosti) položek.

Stratifikovaná Cronbachova alfa: řešení předpokladu jednodimenzionality.

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

- Předpoklad tau-ekvivalence nikoli jen položek, ale i testů.
- Nezávislost disturbancí (lokální nezávislost subtestů).
- Může nadhodnotit.

Spearman-Brown (split-half)

$$r_{SB} = \frac{2r_{xx'}}{1 + r_{xx'}}$$

Odhad korelace paralelních testů z korelace polovin testu.

Předpoklad: **paralelní poloviny**.

- Při porušení může podhodnotit, **ale i nadhodnotit** skutečnou reliabilitu.

Důsledek: není vhodné hledat optimální rozdělení na poloviny.

- Mohlo by vést k nadhodnocení reliability.

Model free: Guttmanova λ_4 (split-half)

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

Guttman (1945): $\lambda_4 = \lambda_3 = \alpha$ v případě, že se test skládá ze dvou položek.

Stejně jako alfa je nutně spodní hranicí reliability.

- Porušení předpokladu tau-ekvivalence podhodnocuje.

To poskytuje prostor pro hledání optimálního rozdělení na poloviny.

Model free: další varianty split-half

Revellova (1979) beta

Postup:

- Nalezení nejhoršího možného rozdělení.

$$\beta = \frac{k^2 \bar{\sigma}_{ij'}}{\sigma_x^2}$$

- kde $\bar{\sigma}_{ij'}$ je průměrná korelace položek napříč těmito polovinami.

Nejmenší možný společný rozptyl vysvětlitelný jediným obecným faktorem.

- Nadhodnocený reziduálními kovariancemi.
- Zinbarg a kol. (2005).

Greatest-lower bound of reliability

- λ_4 je dnes často chápaná jako synonymum pro GLB estimátor.
- např. Woodhouse & Jackson (1977)

Rozdělení na poloviny tak, aby byla λ_4 maximalizovaná.

- Z definice musí jít o spodní hranici reliability za předpokladu nezávislých reziduí.

Problém: nadhodnocení kovariance v důsledku výběrové chyby u malých vzorků

Sijtsma (2009), Revelle a Zinbarg (2009)

Model-based: Rodina koeficientů omega

Odstranění předpokladu tau-ekvivalence díky faktorové analýze.

- Položky testu přestávají být paralelními testy pravého skóre (operacionalismus), ale jde o indikátory způsobené latentní proměnnou (realismus).

Odhaduje rozptyl vysvětlený *latentním rysem* a dělí jej *celkovým rozptylem*.

Více pojetí:

- Liší se ve jmenovateli, resp. v tom, co považují za celkový rozptyl.
- Bollen (1980): parciální omega, kontroluje ostatní faktory. Omezené využití.
- Bentler (1972): celkový rozptyl je implikován modelem.
- McDonald (1999): pracuje s pozorovaným celkovým rozptylem.

Model-based: McDonaldova omega

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2 \mathrm{var}(\psi)}{\mathbf{1}'\Sigma\mathbf{1}}$$

- λ_i – faktorové náboje položek i
- $\mathrm{var}(\psi)$ – rozptyl faktoru
- $\mathbf{1}'\Sigma\mathbf{1}$ – součet prvků pozorované kovarianční matice.
- Vektor faktorových nábojů a kovarianční matice je omezena na položky, které jsou sečteny do celkového skóre.

Faktorové náboje λ_i jsou přímo kontrolovány na odhadované reziduální kovariance položek.

- Předpoklad: model dobře popisuje data.

Model-based: Případy koeficientu omega

Hierarchická omega (ω_h):

- Rozptyl měření vysvětlený jediným (obecným) faktorem.

Celková omega (ω_{tot}):

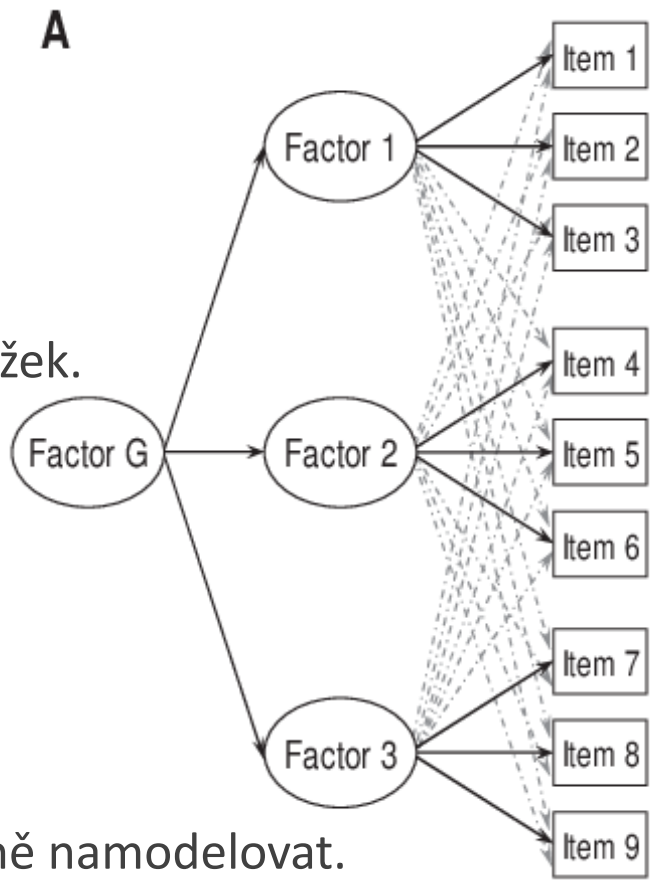
- Rozptyl měření vysvětlený všemi modelovanými faktory.
- Odhad korelace paralelních testů *očištěný* o reziduální kovariance položek.

Několik druhů **second-order omega**:

- Hierarchická faktorová analýza s faktory prvního a vyššího řádu.
- Rozptyl (1st order? položek? položek po kontrole 1st order?) vysvětlený faktorem vyššího řádu.

Výhody koeficientů omega:

- Člověk má kontrolu nad tím, čeho reliabilitu odhaduje, a může si ji volně namodelovat.



Dimension-free omega

Hypoteticky lze odhadovat koeficient omega s pomocí EFA o určitém počtu faktorů.

- Větší množství faktorů ve skutečnosti nenadhodnocuje odhad reliability.
- Revellův přístup (`psych::omega`).
- Záleží ale na výzkumníkovi, kolik faktorů zvolí.

Bentlerův koeficient ρ_{glb} (1980, 2009):

- MTFA (minimum trace factor analysis) dekompozice kovarianční matice.
- Zjednodušeně: Extrakce takového množství ortogonálních faktorů, který nevede k Heywoodovu případu (komunalitě větší než 1).
- Tento postup má právě jedno řešení a jde o nejvyšší možnou spodní hranici reliability za předpokladu nezávislosti reziduálních kovariancí položek.

Specifické případy

Vážený součet položek: vážený koeficient omega

$$\omega_w = \frac{\left(\sum_{i=1}^{k} w_i \lambda_i\right)^2 \mathrm{var}(\psi)}{\mathbf{W}'\Sigma\mathbf{W}}$$

- V případě, že je namísto součtu položek používán odhad faktorového skóre, lze požít váženou omegu, nebo rovnou FSD (**factor score determinacy**) jako odhad reliability.

V případě, že je modelem měření **ordinální faktorová analýza**, pak je nutné použít Green-Yangův (2009) vzorec 21, jinak jde o tzv. „ordinální alfu“.

Ordinální koeficient alfa, omega... (Zumbo, Gadderman,, 2007, 2012).

- Koeficienty alfa či omega odhadnuté nad maticí polychorických korelací.
- Nejde o odhad reliability v pravém slova smyslu (viz Chalmers, 2018), ale užitečnou deskripci.

Děkuji za pozornost!

MUNI FSS Katedra psychologie

Hynek Cígler | cigler@fss.muni.cz
Katedra psychologie & Institut pro výzkum dětí, mládeže a rodiny
Fakulta sociálních studií, Masarykova univerzita MU
https://is.muni.cz/osoba/hynek.cigler | https://osf.io/t6ufg/

- Bentler, P. M. (2009). Alpha, Dimension-Free, and Model-Based Internal Consistency Reliability. *Psychometrika*, *74*(1), 137–143.
- Chalmers, R. P. (2018). On Misconceptions and the Limited Usefulness of Ordinal Alpha. *Educational and Psychological Measurement*, *78*(6).
- Cho, E. (2016). Making Reliability Reliable: A Systematic Approach to Reliability Coefficients. *Organizational Research Methods*, *19*(4), 651–682.
- Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating Ordinal Reliability for Likert-Type and Ordinal Item Response Data: A Conceptual, Empirical, and Practical Guide. *Practical Assessment, Research, and Evaluation*, *17*(3), 1–13.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates, Inc.
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154.
- Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120
- Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika*, *42*(4), 579–591.
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's, α Revelle's β and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.

- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's, α Revelle's β and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.

- Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*(1), 21–29.