Cheatsheet: Statistische Methoden

Inhaltliche Bedeutsamkeit

Signifikante Ergebnisse (z. B. sign. von 0 verschiedener Mittelwert bei sehr großer Stichprobe) sind nicht notwendigerweise inhaltlich bedeutsam.

Zur Einschätzung der inhaltlichen Bedeutsamkeit standardisierte Effektstärken nützlich

Effektstärke der Abweichung des Mittelwerts $\bar{x}$ vom fixen Wert μ

$$d = \frac{\bar{x} - \mu}{\sigma_X}$$

Effektstärke der Differenz der Mittelwerte $\bar{x}_1$ und $\bar{x}_2$ aus unabhängigen Stichproben

$$d' = \frac{\bar{x}_2 - \bar{x}_1}{\sigma_{\text{inn}}}$$

wobei: σ_X = Standardabweichung von X in der Population; σ_{inn} = Standardabweichung innerhalb der beiden Teilpopulationen

Effektstärke - Daumenregel

Grundsätzlich: Ob eine Effektstärke inhaltlich bedeutsam ist oder nicht hängt maßgeblich vom Untersuchungsgegenstand ab.

Orientierungshilfe liefern Daumenregeln nach Cohen (1988)

- $|d| = 0,14 \rightarrow$ klein
- $|d| = 0,35 \rightarrow$ mittel
- $|d| = 0,57 \rightarrow$ groß
- $|d'| = 0,20 \rightarrow$ klein
- $|d'| = 0,50 \rightarrow$ mittel
- $|d'| = 0,80 \rightarrow$ groß

Reliable Change Index (RCI)

Zur Berechnung benötigt werden die Standardabweichungen s_1 und s_2 zu den Messzeitpunkten 1 und 2 und die Reliabilität r_{xx} des Messinstruments.

$$RCI_i = \frac{x_{2i} - x_{1i}}{SE_{\text{diff}}}$$

wobei

$$SE_1 = s_1\sqrt{1 - r_{xx}}, SE_2 = s_2\sqrt{1 - r_{xx}}$$
$$SE_{\text{diff}} = \sqrt{SE_1^2 + SE_2^2}$$

Unter Annahme von $SE_{diff} \sim N(0,1)$ kann z-Verteilung für inferenzstatistischen Test oder zur Bestimmung von Konfidenzintervallen genutzt werden: 1.04 (70% CI), 1.28 (80% CI), 1.64, (90% CI), 1.96 (95% CI).

Differenz auf Gruppenebene

Auch Veränderungen auf Gruppenebene können standardisiert werden. Hier die standardisierte Differenz d'' zwischen zwei Mittelwerten $\bar{x}_1$ und $\bar{x}_2$ sowie der Streuung der Differenzen:

$$d'' = \frac{\bar{x}_2 - \bar{x}_1}{\sigma_D} = \frac{\bar{x}_2 - \bar{x}_1}{\sigma_{(x_2 - x_1)}}$$

Für diese standardisierten Differenzen schlägt Cohen (1988) folgende Klassifikation vor:

$|d''| = 0.14 \rightarrow$ „klein"

$|d''| = 0.35 \rightarrow$ „mittel"

$|d''| = 0.57 \rightarrow$ " groß "

Inferenzstatistische Absicherung mit t-Test

Quantifizierung des Nutzens

Der monetäre Nutzen einer Maßnahme wird ermittelt, indem die Wirksamkeit mit einem Geldwert gewichtet wird.

$$\text{Nutzen} \;=\; \text{Wirksamkeit} \;\cdot\; \text{Wert}$$

Modell 1: Kosten-Nutzen-Analyse (KNA)

Basierend auf Nutzen N (=Wirksamkeit * Wert) und Kosten K lassen sich verschiedene Kennwerte bilden.

Nettonutzen: $NN = N - K$

Nutzenquotient: $NQ = \frac{N}{K}$

Im wirtschaftlichen Bereich auch Return on Investment (ROI)

$NQ > 1 \rightarrow$ Maßnahme ist effizient; über Maßnahmen hinweg vergleichbar

Profitrate: $PR = \frac{NN}{N} = \frac{N-K}{N}$

Nettonutzen in Relation zum Gesamtnutzen

Klassische Testtheorie

KTT geht davon aus, dass das interessierende Merkmal kontinuierlich ist, und dass sich die mit einem Test ermittelte Merkmalsausprägung X_i von Individuum i aus dem wahren Wert T_i des Individuums und einem zufälligen Messfehler E_i zusammensetzt.

$$X_i = T_i + E_i$$

KTT fokussiert vor allem auf Bestimmung des Anteils des Messfehlers

Reliabilität definiert als Anteil der „wahren Varianz" an der beobachteten Varianz:

$$\text{Rel} = \frac{\sigma_T^2}{\sigma_X^2}$$

Verknüpfungsfunktion

In Modellen mit latenten Variablen wird die Beziehung zwischen beobachteten Indikatoren und latenten Variablen mit einer mathematischen Verknüpfungsfunktion (link function) definiert. - Bei Modellen

mit kontinuierlichen Indikatoren bspw. eine lineare Funktion (z. B. bei konfirmatorischer Faktorenanalyse).

$$x = \alpha + \lambda\theta + \varepsilon$$

α = Konstante (Intercept) λ = Gewicht (Faktorladung) ε = Messfehler

IC-Funktion

IC-Funktion im Raschmodell enthält 2 Parameter.

Wahrscheinlichkeit, dass Person j mit Fähigkeit θ Aufgabe i mit Schwierigkeit b richtig beantwortet, ist gegeben durch:

$$P\left(x_{ij} = 1 \mid \theta_j, b_i\right) = \frac{\exp\left(\theta_j - b_i\right)}{1 + \exp\left(\theta_j - b_i\right)}$$

Entscheidend für Lösungswahrscheinlichkeit ist Differenz zwischen individueller Merkmalsausprägung und Itemschwierigkeit

Logistische IC-Funktion

IC-Funktion des Raschmodells ist eine logistische Funktion

Wird auch in der logistischen Regression verwendet

Differenz zwischen Merkmalsausprägung θ_j und Itemschwierigkeit b_i entspricht den logarithmierten Odds („Wettquotient") einer richtigen zu einer falschen Antwort.

$$\log\left(\frac{P\left(x_{ij} = 1\right)}{P\left(x_{ij} = 0\right)}\right) = \theta_j - b_i$$

Mit der logistischen Funktion wird die Antwortwahrscheinlichkeit $P\left(x_{ij} = 1\right)$ von 0 bis 1 auf den Wertebereich von $-\infty$ bis $+\infty$ projiziert:

Der nicht logarithmierte Wettquotient hat Wertebereich von 0 bis $+\infty$.

$$\log\left(\frac{P\left(x_{ij} = 1\right)}{P\left(x_{ij} = 0\right)}\right)$$

IC-Funktion Raschmodell

$$P\left(x_{ij} = 1 \mid \theta_j, b_i\right) = \frac{\exp\left(\theta_j - b_i\right)}{1 + \exp\left(\theta_j - b_i\right)}$$

Beziehung zwischen Merkmal und Antwortwahrscheinlichkeit streng monoton

Itemschwierigkeit b_i ist Punkt auf dem Merkmalskontinuum, an dem Lösungswahrscheinlichkeit 50% beträgt (=Wendepunkt)

Unterschied 1PL / Raschmodell

1PL kann auch notiert werden als:

$$P\left(x_{ij} = 1 \mid \theta_j, b_i, a_i\right) = \frac{\exp\left(a\left(\theta_j - b_i\right)\right)}{1 + \exp\left(a\left(\theta_j - b_i\right)\right)}$$

1PL: $a = \text{const}$ für alle $i \in I$

Rasch-Modell: $a = 1$ für alle $i \in I$

1PL und Rasch-Modell sind mathematisch äquivalent.

ML-Schätzung von $\hat{\theta}$

Exemplarisch für $\theta = -3$:

Schritt 1:

$$p(x_1 = 1 \mid \theta = -3, b_1 = -1.90) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} = \frac{\exp(-3 + 1.9)}{1 + \exp(-3 + 1.9)} = 0.2497$$
$$p(x_2 = 1 \mid \theta = -3, b_1 = -0.60) = 0.0832$$
$$p(x_3 = 0 \mid \theta = -3, b_1 = -0.25) = 1 - p(x_3 = 1 \mid \theta = -3, b_1 = -0.25) = 0.9399$$
$$p(x_4 = 0 \mid \theta = -3, b_1 = -0.25) = 1 - p(x_4 = 1 \mid \theta = -3, b_1 = -0.25) = 0.9644$$
$$p(x_5 = 0 \mid \theta = -3, b_1 = -0.25) = 1 - p(x_5 = 1 \mid \theta = -3, b_1 = -0.25) = 0.9692$$

Schritt 2:

$$p(x_1 = 1) * p(x_2 = 1) * p(x_3 = 0) * p(x_4 = 0) * p(x_5 = 0) =$$
$$0.2497 * 0.0832 * 0.9399 * 0.9644 * 0.9692 = 0.0182$$

Likelihoodfunktion

Betrachtet man den Vektor $\boldsymbol{x_j}$ der Antworten von $\boldsymbol{j} = 1, 2, \ldots$, J Personen sowie den Vektor $\boldsymbol{b}_i$ der Itemschwierigkeiten von $i = 1, 2, \ldots$, I Items gemeinsam, so lässt sich dies notieren als:

$$L(\boldsymbol{x}_j \mid \theta, \boldsymbol{b}) = \prod_{i=1}^{I} p_i^{x_{ij}} (1 - p_i)^{(1 - x_{ij})}$$

wobei: x_{ij} = beobachtete Antwort von Person j auf Item i p_i = Wahrscheinlichkeit von $x_{ij} = 1; p(x_{ij} = 1 \mid \theta_j, b_i)$

Log-Likelihood

Umfasst ein Test viele Items, wird die Likelihood sehr klein.

Dies erschwert die Verarbeitung durch den Computer.

Daher wird bei ML-Schätzungen der natürliche Logarithmus (ln) der Likelihood verwendet, die Log-Likelihood:

$$\ln L(\boldsymbol{x}_j \mid \theta, \boldsymbol{b}) = \sum_{i=1}^{I} x_{ij} \ln(p_i) + (1 - x_{ij}) \ln(1 - p_i)$$

Log-Likelihood ist eine streng monotone Transformation der Likelihood aber in der Handhabung deutlich einfacher.

Eine Optimierung der Log-Likelihood optimiert zugleich die Likelihood.

Standardfehler von θ

Wie für die meisten statistischen Parameter lässt sich auch für die geschätzte Merkmalsausprägung $\hat{\theta}_j$ ein Standardfehler berechnen.

In diesen Standardfehler SE(θ) ($\sigma_e(\theta)$ in de Ayala) gehen im Raschmodell/1PL die individuellen Antwortwahrscheinlichkeiten $p_i = p(x_{ij} = 1 \mid \theta_j, b_i)$ für alle beantworteten Items ein:

$$SE(\theta) = \sqrt{\frac{1}{\sum_{i=1}^{I} p_i (1 - p_i)}}$$

An dieser Formel sind unmittelbar zwei Eigenschaften des Standardfehlers zu erkennen:

1. $SE(\theta)$ fällt für verschiedene Personen und Items unterschiedlich aus, da die individuellen Antwortwahrscheinlichkeiten eingehen.

2. $SE(\theta)$ wird aufgrund der Summe über / Items im Nenner kleiner, je mehr Items beantwortet wurden.

Testinformation

Die Gesamtinformation eines Tests ergibt sich aus der Summe aller Iteminformationen.

Ergebnis ist die Testinformationsfunktion (Engl.: Test Information Curve [TIC]; de Ayala: total information)

Testinformationsfunktion gibt Messgenauigkeit der Merkmalsschätzung in Abhängigkeit von θ an.

Zusammenhang zwischen Testinformation I für einen speziellen θ-Wert und Standardfehler:

$$SE(\hat{\theta} \mid I) = 1/\sqrt{I(\hat{\theta} \mid I)}$$

2PL

Vorherrschendes Modell in den USA

Bei groß angelegten Vergleichsstudien wie IGLU, TIMSS und PISA (seit 2015) genutzt

Eingeführt von Birnbaum (1968)

Ein Personenparameter θ und zwei Itemparameter a und b (α and δ in de Ayala, 2022)

$$P\left(x_{ij} = 1 \mid \theta_j, a_i, b_i\right) = \frac{\exp\left(a_i\left(\theta_j - b_i\right)\right)}{1 + \exp\left(a_i\left(\theta_j - b_i\right)\right)}$$

Iteminformation im 2PL

Die Itemdiskrimination geht im 2PL mit in die Iteminformation ein:

$$I_i(\theta) = \alpha_i^2 p_i\left(1 - p_i\right)$$

Items mit höherer Diskrimination sind dementsprechend informativer bei der Erfassung des interessierenden Merkmals.

3PL

Seltener genutzt als 1PL und 2PL

Einsatz bspw. beim National Assessment of Educational Progress (NAEP) in den USA

Ein Personenparameter θ, zwei Itemparameter a, b und zusätzlich der Pseudo-Rateparameter $c(\alpha, \delta, \chi$ in de Ayala, 2022)

$$P\left(x_{ij} = 1 \mid \theta_j, a_i, b_i, c_i\right) = c_i + \left(1 - c_i\right)\frac{\exp\left(a_i\left(\theta_j - b_i\right)\right)}{1 + \exp\left(a_i\left(\theta_j - b_i\right)\right)}$$

Iteminformation im 3PL

Beim 3PL gehen Itemdiskrimination und Pseudo-Rateparameter mit in die Iteminformation ein:

$$I_i(\theta) = a_i^2 \left(\frac{p_i(\theta) - c_i}{1 - c_i} \right)^2 \frac{q_i(\theta)}{p_i(\theta)}$$

Formale Definition von uniformem DIF

Im Rahmen von IRT-Modells für dichotome Daten lässt sich DIF durch gruppenspezifische Itemparameter darstellen, im Rahmen das dichotomen Raschmodells z. B. wie folgt:

$$P\left(x_{ij} = 1\right) = \frac{\exp\left(\theta_j - b_{gi}\right)}{1 + \exp\left(\theta_j - b_{gi}\right)}$$

b_{gi} ist hier eine gruppenspezifische Itemschwierigkeit.

Ein Item i weist DIF auf, wenn sich die Itemschwierigkeiten der Referenzgruppe ($g = R$) oder der Fokalgruppe ($g = F$) unterscheiden ($b_{\mathrm{Ri}} \neq b_{Fi}$).

Formale Definition von non-uniformem DIF

Im Rahmen von IRT-Modellen für dichotome Daten lässt sich DIF durch gruppenspezifische Itemparameter darstellen, im Rahmen des 2PL z. B. wie folgt:

$$P\left(x_{ij} = 1\right) = \frac{\exp\left(\alpha_{gi}\left(\theta_j - b_{gi}\right)\right)}{1 + \exp\left(\alpha_{gi}\left(\theta_j - b_{gi}\right)\right)}$$

b_{gi} ist hier wieder die gruppenspezifische Itemschwierigkeit, α_{gi} eine gruppenspezifische Itemdiskrimination. Ein Item i weist non-uniformen DIF auf, wenn sich die Itemdiskriminationen der Referenzgruppe ($g = R$) oder der Fokalgruppe ($g = F$) unterscheiden ($\alpha_{Ri} \neq \alpha_{Fi}$).