Inhaltsverzeichnis

Zusammenfassung

Draxler & Zessin (2015) haben eine Klasse pseudo-exakter oder konditionaler Tests zur Power-Berechnung von Annahmen des Rasch Modells vorgeschlagen. Zum Simulieren der für die Power-Berechnung notwendigen Daten bedarf es Sampling-Algorithmen. Verhelst (2008) hat mit dem *Rasch Sampler* einen relativ schnellen Algorithmus entworfen, der die wahre Verteilung mithilfe von *Markov Chain Monte Carlo* Prozeduren approximiert. Miller & Harrison (2013) haben mit dem *Exact Sampler* einen Algorithmus entwickelt, der die exakte Verteilung abzählen und daraus ziehen kann. Die Genauigkeit der beiden Sampler wird verglichen, indem potentielle Einflüsse der Stichprobengröße, DIF-Parameter und Itemschwierigkeit auf die Genauigkeit der Power-Berechnung untersucht werden. Darüber hinaus werden die Burn-In Phase und der Step-Parameter als Einflussfaktoren auf den *Rasch Sampler* überprüft. Die Genauigkeit der Sampler unterscheidet sich nicht wesentlich. Bei steigender Stichprobengröße steigt die Power an. Auch bei größeren Modellabweichung im Positiven wie im Negativen kann eine höhere Power beobachtet werden. Bei moderater Itemschwierigkeit ist die Power bei positivem und negativem DIF-Parameter nahezu gleich groß und bei Modellabweichung eines leichten Items ist die Power bei positiver Abweichung größer als bei negativer. Mit einem schwierigen Item ist mit dem Unterschied, dass die Streuung deutlich höher ausfällt, ein gegensätzlicher Trend zu beobachten. Weder die Burn-In Phase noch der Step-Parameter hat einen Einfluss auf die Genauigkeit des *Rasch Samplers*. Aufgrund von effizienterer Berechnung sollte in jedem Fall der *Rasch Sampler* verwendet werden. Die Ergebnisse bezüglich des Verhaltens der Power unter Variation verschiedener Parameter entsprechen den Beobachtungen von Draxler & Zessin (2015).

Schlüsselwörter: Rasch Modell, Power, Pseudo-exakte Tests, Konditionale Tests, Rasch Sampler, Exact Sampler

1 Einleitung

Die psychologische Testtheorie ist in der Psychologie eine der Grundlagen für neue Erkenntnisse. Anders als zum Beispiel bei der Körpergröße, sind viele interessierende Dimensionen in der Psychologie nicht direkt messbar. Derartige Merkmale nennt man latent. Jedes Mal, wenn ein latentes Merkmal sichtbar gemacht werden soll, muss man einen psychologischen Test anwenden. Dabei können die Anwendungsbereiche von Intelligenzquotienten über Persönlichkeitsmerkmale bis hin zur Erfassung ganzheitlicher Arbeitsbedingungen alle nur erdenklichen Bereiche der Psychologie abdecken. Aber auch in anderen Fachbereichen finden testtheoretische Ansätze Anwendung. Es gibt grundsätzlich zwei Ansätze: die klassische und die probabilistische Testtheorie. Für diese Arbeit ist ausschließlich Letztere von Bedeutung. Die probabilistische Testtheorie geht davon aus, dass die Ausprägung des gemessenen Merkmals sowohl von der Personenfähigkeit als auch von der Itemschwierigkeit abhängt. Dabei gibt es je nach Anwendungsbezug im Rahmen der *Item response theory* (IRT) Modelle, mithilfe man die Wahrscheinlichkeit ausrechnen kann, einen bestimmten Wert für ein bestimmtes Item zu erzielen. Bei einer erneuten Testung der Personen würde die Wahrscheinlichkeit für eine Antwort x gleich bleiben, aber das Ergebnis könnte variieren, da der Zusammenhang nicht deterministisch festgelegt ist. Die probabilistische Testtheorie hat in den letzten Jahren vermehrt an Bedeutung gewonnen, was weniger an der kürzlichen Erschließung des theoretischen Fundaments, sondern viel mehr an dem rasanten Fortschritt der Technik liegt, die es benötigt, um Kalkulationen unter der IRT durchzuführen. Ein bekanntes Anwendungsbeispiel sind die PISA-Studien (Prenzel, Walter & Frey, 2007). Da die Lernpläne zwischen den verschiedenen Ländern nicht identisch sind, können nur IRT Modelle einen adäquaten Vergleich schaffen.

Eine zentrale Größe vor und nach der Durchführung eines Tests ist die Anzahl der richtigen und falschen Ergebnisse. Wie oft erkennt der Test zum Beispiel eine pathologische Diagnose, obwohl der Patient gesund ist und wie oft wird der Patient

gesund eingeschätzt, obwohl dieser eine Erkrankung hat? Während ersterer Frage in der Vergangenheit bereits viel Aufmerksamkeit zuteil wurde, hat man die Wichtigkeit letzterer länger unterschätzt. Dabei kann dieser Fehler je nach Kontext nicht nur genauso relevant sein wie der erste, sondern noch tiefgreifender. Angenommen, ein Test soll Schizophrenie diagnostizieren. Nach der Diagnose erhält der Patient ein starkes Neuroleptikum mit schwerwiegenden Nebenwirkungen. Der erste Fehler beschreibt die Wahrscheinlichkeit, dass ein Patient, der keine Schizophrenie hat, das Neuroleptikum verschrieben bekommt. Das ist sicherlich nicht erwünscht, aber der zweite Fehler beschreibt in diesem Fall die Wahrscheinlichkeit, dass eine schizophrene Person als gesund diagnostiziert und unbehandelt nach Hause geschickt wird. Diese letztere Wahrscheinlichkeit kann kontrolliert werden. Man sollte die Power – die Gegenwahrscheinlichkeit besagten Fehlers – eines statistischen Tests, wie Draxler (2010) vorgeschlagen hat, genau wie den Fehler erster Art a priori festlegen und aufgrund dessen die für eine gewünschte Genauigkeit benötigte Stichprobengröße ausrechnen.

Im Anwendungskontext von Modellen unter der IRT kann die Power mithilfe von Simulationen berechnet werden. Für diese Simulationen gibt es zwei verschiedene Ansätze. Einerseits ein relativ schnelles, approximatives Verfahren und andererseits ein langsameres Verfahren, bei dem die exakte Verteilung bekannt ist. Da der exakte Algorithmus extrem rechenintensiv ist, muss die Frage gestellt werden, in welchen Szenarien es genügt, die zugrunde liegende Verteilung zu approximieren und wann oder ob überhaupt die Notwendigkeit der exakten Berechnung besteht. Auch ist es durch die exakte Funktion zum ersten Mal möglich, die Genauigkeit des approximativen Verfahrens im Allgemeinen zu überprüfen, da die genaue Verteilung bekannt ist. Diesen Fragen sowie dem Abklären potentieller Einflüsse verschiedener Parameter auf die Power-Berechnung evaluiert diese Arbeit.

1.1 Theoretischer Hintergrund

Im Folgenden werden die für diese Arbeit notwendigen theoretischen Fundamente vom binären Rasch Modell über die Berechnung der Power pseudo-exakter oder konditionaler Tests von Annahmen des Modells nach Draxler und Zessin (2015) bis hin zum Vergleich der Funktionsweise verschiedener Sampling-Algorithmen erklärt.

2 Methodik

Im Folgenden wird das genaue Studiendesign zur Beantwortung der sechs Fragestellungen und die für die Datenanalyse verwendete Hardware und Software vorgestellt.

2.1 Studiendesign

Für sämtliche Berechnungen wird α mit 5 Prozent gewählt. Die Anzahl der gezogenen Matrizen wird bei Fragestellung A auf 3 000 und bei Fragestellung B bis E auf 8 000 festgesetzt, da diese als genügend groß angenommen werden kann, um die gesamte Verteilung abzubilden ...

3 Ergebnisse

[...] Auffällig ist die bedeutend höhere Standardabweichung bei 150 Personen und 4 Items. Diese ist mit *SD* = 0.063 beim *Rasch Sampler* und *SD* = 0.064 beim *Exact Sampler* mit der Ausnahme von 60 Personen, da dort einige Ausreißer zu beobachten sind, bedeutend höher als bei anderen Personenkonstellationen, die in einem Bereich von 0.009 und 0.026 liegen. Des Weiteren ist die Power bei beiden Samplern bei der 30 x 4 Matrix mit *M* = .25 und *SD* = 0.014 respektive *M* = .25 und *SD* = 0.013 beim *Exact Sampler* deutlich niedriger als die der anderen Matrizen. An diesen Beispielen kann man exemplarisch auch erkennen, wie gering die Unter-

schiede zwischen den Samplern sind. Die Matrix mit zehn Personen weist mit *M* = .59 die größte mittlere Power auf. Auffällig ist außerdem die fehlende Kontinuität beider Sampler. Bei steigender Personenzahl ist kein linearer Zusammenhang in Form von steigender Power beobachtbar. In Abbildung 1 wird der in Fragestellung B untersuchte Zusammenhang zwischen der Power und der Stichprobengröße illustriert. [...]

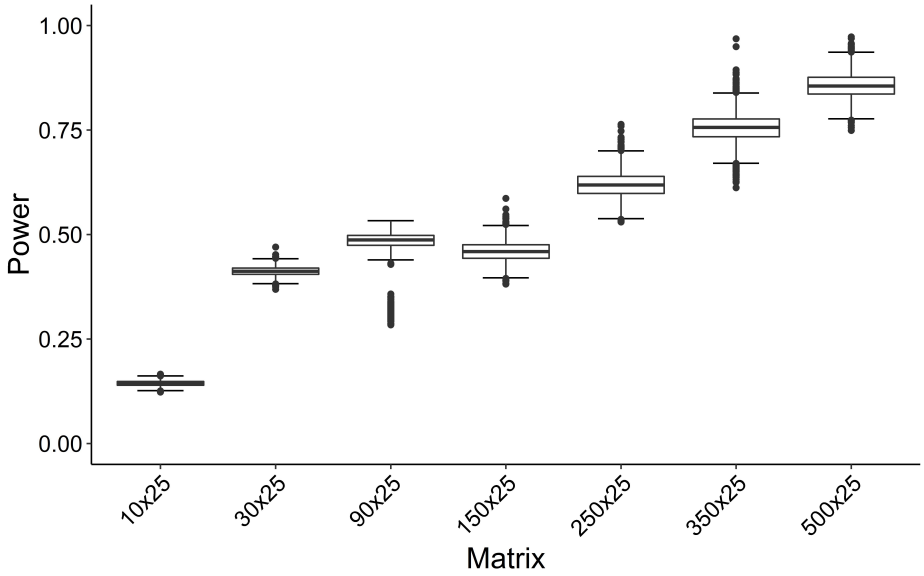

Abbildung 1. Box Plots der Wahrscheinlichkeitsverteilung der Power verschiedener Stichprobengrößen bei gleichbleibender Itemanzahl

Tabelle 1

Deskriptive Statistik zu Abbildung 1 mit Minimum (Min), 2.5% Quantil (Q.025), 25% Quantile (Q.25), Median, Mittelwert (Mean), 75% Quantil (Q.75), 97.5% Quantil (Q.975), Maximum (Max) und Standardabweichung (SD).

Matrix	Min	Q.025	Q.25	Median	Mean	Q.75	Q.975	Max	SD
10x25	0.12	0.13	0.14	0.14	0.14	0.15	0.16	0.17	0.007
30x25	0.37	0.39	0.40	0.41	0.41	0.42	0.43	0.47	0.011
90x25	0.28	0.31	0.47	0.49	0.47	0.50	0.51	0.53	0.051
150x25	0.38	0.42	0.44	0.46	0.46	0.48	0.51	0.59	0.024
250x25	0.53	0.56	0.60	0.62	0.62	0.64	0.68	0.76	0.031
350x25	0.61	0.68	0.73	0.76	0.76	0.78	0.82	0.97	0.035
500x25	0.75	0.80	0.84	0.86	0.86	0.88	0.92	0.97	0.031

4 Diskussion

5 Literaturverzeichnis

Draxler, C. (2010). Sample size determination for rasch model tests. *Psychometrika*, *75* (4), 708–724.

Draxler, C. & Zessin, J. (2015). The power function of conditional tests of the rasch model. *AStA Advances in Statistical Analysis*, *99* (3), 367–378.

Miller, J. W. & Harrison, M. T. (2013). Exact sampling and counting for fixed-margin matrices. *The Annals of Statistics*, *41* (3), 1569–1592.

Prenzel, M., Walter, O. & Frey, A. (2007). Pisa misst kompetenzen. *Psychologische rundschau*, *58* (2), 128–136.

Verhelst, N. D. (2008). An efficient mcmc algorithm to sample binary matrices with fixed marginals. *Psychometrika*, *73* (4), 705–728.

A Funktionen in R

A.1 Powerfunktion für Fragestellung A

```
pwr_A <- function(rows_sums, cols_sums,
                  n_repeats = 1000, n_matrices = 3000,
                  alpha = .05, dev = .6,
                  item_pos = 2, burnIn = 300,
                  step = 16, folder = ""){
  ##################################################################
  # INPUTS:
  # rows_sums: Zeilenrandsummen
  # cols_sums: Spaltenrandsummen
  # n_repeats: Anzahl an Power-Werte pro Szenario
  # n_matrices: Anzahl an Matrizen pro Power-Wert
  # alpha: Fehler 1. Art
  # dev: DIF-Parameter
  # item_pos: Items mit Modellabweichung
  # burnIn: Burn-In Phase fuer Rasch Sampler
  # step: Step-Parameter fuer Rasch Sampler
  # folder: Speicherort
  ##################################################################
  model <- sample(rows_sums, cols_sums, 1)
  half_length <- length(cols_sums) / 2
  groups <- c(rep(1, half_length), rep(0, half_length))
  dif <- rep(0, length(cols_sums))
  dif[item_pos] <- dev

  path <- paste0(folder, "/",
                 as.character(length(rows_sums)), "x",
```

```
                as.character(length(cols_sums)), ".csv")

  mcmc <- exact <- vector("numeric", n_repeats)

  count(rows_sums, cols_sums)

  mcmc <- replicate(n_repeats,
                    pwr_mcmc(mat = model,
                             group = groups,
                             dif = dif,
                             repetitions = n_matrices,
                             alpha = alpha,
                             burn = burnIn,
                             steps = step))
  exact <- replicate(n_repeats,
                     pwr_exact(rows = rows_sums,
                               cols = cols_sums,
                               group = groups,
                               dif = dif,
                               repetitions = n_matrices,
                               alpha = alpha))

  rio::export(data.frame(power = c(mcmc, exact),
                         method = rep(c("mcmc", "exact"),
                                      each = n_repeats)), path)

}
```