methevalreport
4(4) 2002

Ute Suhl & Christof Nachtigall

Warum kompliziert, wenn es auch einfach geht?

Teil 2: Ergebnisse einer Simulationsstudie zum Vergleich von Veränderungskennwerten

metheval
psychologische methodenlehre
und evaluationsforschung jena

Impressum

methevalreport
erscheint seit 1999
in unregelmäßigen Abständen
als „graue“ Schriftenreihe des Lehrstuhls für
Psychologische Methodenlehre und Evaluationsforschung
am Institut für Psychologie der Friedrich-Schiller-Universität Jena

Herausgeber:
Prof. Dr. Rolf Steyer
Sekr.: +49 (3641) 945 230
Durchwahl: +49 (3641) 945 231
Fax: +49 (3641) 945 232

rolf.steyer@uni-jena.de

Redaktion:
Dipl.Psych. Friedrich Funke
sff@uni-jena.de

Typographie:
cand.psych. Silke Zachariae
zachariae@web.de

Standort:
Thüringer Universitäts- und Landesbibliothek
Lesesaal Zweigstelle Psychologie

Internet
http://www.uni-jena.de/svw/metheval/report/

Bestellungen:
Methodenlehre und Evaluationsforschung
Institut für Psychologie
Steiger 3 Haus 1
D-07743 Jena
Deutschland

Warum kompliziert, wenn es auch einfach geht?
Teil 2: Ergebnisse einer Simulationsstudie zum Vergleich von Veränderungskennwerten

Ute Suhl und Christof Nachtigall

Institut für Psychologie

Friedrich-Schiller-Universität Jena

1. Einleitung

In der klinischen Praxis stellt sich häufig die Frage, ob sich ein krankheitsrelevantes Merkmal eines Klienten im Verlauf der Behandlung verändert. Zur Erfassung solcher intraindividuellen Veränderungen sind verschiedene Ansätze entwickelt worden. Diese lassen sich einteilen in sogenannte *direkte Verfahren* und *indirekte Verfahren*. Bei den direkten Verfahren wird der Klient gebeten, unmittelbar eine Einschätzung abzugeben, in welchem Ausmaß sich sein Zustand verändert hat. Bei den indirekten Verfahren versucht man dagegen, durch den Vergleich von Messwerten, die zu unterschiedlichen Zeitpunkten (z.B. vor und nach einer Intervention) erhoben wurden, zu erschließen, ob eine Veränderung stattgefunden hat. Dazu werden die Messwerte zu einem Veränderungsindex verrechnet, der dann mit Hilfe eines einfachen statistischen Tests auf Signifikanz geprüft werden kann.

Über die Art und Weise, wie (im einfachsten Fall: zwei) Messwerte eines Klienten am besten zu einem Index der Veränderung zusammengefasst werden können, herrscht allerdings keine Einigkeit. Es sind vielmehr verschiedene Ansätze entwickelt worden, und für den Praktiker stellt sich die Frage, welchen Index er denn nun im konkreten Anwendungsfall verwenden sollte. Basis einer solchen Entscheidung sind neben inhaltlichen Erwägungen, was ein solcher Index eigentlich genau abbildet, auch statistische Kriterien: Wie oft signalisiert ein solcher Index fälschlicherweise eine Veränderung, obwohl eigentlich keine Veränderung stattgefunden hat? Und mit welcher Zuverlässigkeit kann eine stattgefundene Veränderung auch tatsächlich aufgedeckt werden? Anders formuliert geht es also um den α-Fehler und die Power des statistischen Tests, mit dem ein Veränderungsmaß auf Signifikanz geprüft wird.

Im vorliegenden Beitrag werden wir mit Hilfe von Computersimulationen zwei Veränderungsmaße, die zur Beschreibung und Prüfung intraindividueller Veränderungen zwischen zwei Messzeitpunkten vorgeschlagen wurden, miteinander hinsichtlich ihrer statistischen Eigenschaften vergleichen. Dabei werden wir zum einen den sogenannten Reliable Change Index (Jacobson & Truax, 1991) betrachten, der auf der beobachtbaren Differenz zwischen einem z.B. nach einer Intervention erhobenen Posttestwert Y und einem vor der Behandlung erhobenen Prätestwert X aufbaut und eng verwandt ist mit der auf Lienert (1961) zurückgehenden kritischen Differenz. Zum anderen betrachten wir einen von Steyer, Hannöver, Telser und Kriebel (1997) vorgeschlagenen Veränderungsindex, der für sich in Anspruch nimmt, im Gegensatz zur manifesten Differenz Y-X dem Messfehlerproblem adäquat Rechnung zu tragen. Uns geht es im vorliegenden Beitrag nicht darum zu klären, was aus inhaltlicher Sicht betrachtet diese Indizes genau abbilden (vgl. hierzu Nachtigall & Suhl, in Vorbereitung); Zielsetzung ist vielmehr zu untersuchen, welche Konsequenzen sich aus der unterschiedlichen Art der Zusammenfassung von Post- und Prätestwert bei einer statistischen Absicherung hinsichtlich α-Fehler und Power des Tests ergeben: Wie oft signalisieren die

Tests eine Veränderung, obwohl keine Veränderung der Truescores vorliegt? Und wie zuverlässig decken die Tests tatsächlich vorliegende Truescore-Änderungen auf? Gibt es bzgl. dieser Kriterien Unterschiede zwischen den beiden Veränderungsindizes?

Um die Simulationsidee zu verdeutlichen, werden wir im nachfolgenden Abschnitt zunächst die beiden Veränderungsmaße mit ihren wesentlichen Eigenschaften kurz vorstellen und darauf aufbauend die genaue Fragestellung der Simulationsstudie erläutern. Anschließend wird das Vorgehen bei der Simulation beschrieben sowie die Ergebnisse dargestellt und hinsichtlich ihrer praktischen Relevanz diskutiert.

2. Zur Messung intraindividueller Veränderung: *RCI* und V_{infer}

Der einfachste Weg, zwei Messwerte miteinander zu vergleichen, um zu beurteilen, ob eine Veränderung stattgefunden hat, besteht darin, ihre Differenz zu betrachten. Nehmen wir an, wir haben für einen Klienten zwei Messwerte erhoben, einen Prätestwert x, der seinen Zustand vor Beginn der Behandlung beschreibt, und einen Posttestwert y, der seinen Zustand nach einer Behandlung beschreibt. Die Differenz $y–x$ gibt dann Auskunft darüber, ob der Posttestwert y im Vergleich zu x größer oder kleiner geworden oder gleich geblieben ist. Leider sind psychologische Tests allerdings nicht 100%-tig zuverlässig: Die meisten psychologischen Messungen sind stets auch überlagert von vielen kleinen Zufallseinflüssen, die nichts mit dem eigentlich zu messenden Merkmal zu tun haben. Im Rahmen der Klassischen Testtheorie (vgl. z.B. Lienert, 1961; Steyer & Eid, 1993) wird dies durch den Ansatz formalisiert, dass sich ein Messwert X aus einem sogenannten „wahren Wert“ τ_X, der die eigentlich zu messende Eigenschaft widerspiegelt, und einem Messfehleranteil ε_X, der die bei der Messung ebenfalls wirksamen Zufallseinflüsse widerspiegelt, zusammensetzt. Erhält man beim Vergleich zweier Messwerte eine von Null verschiedene Differenz $Y–X$, dann heißt das nicht zwangsläufig, dass sich der wahre Wert der Person, d.h. das eigentlich interessierende Merkmal der Person, über die Zeit geändert hat. Solche Unterschiede zwischen Y und X können vielmehr auch allein durch Messfehlereinflüsse zustande kommen. In dieser Situation behilft man sich damit, dass man versucht, die nicht 100%-tige Zuverlässigkeit des psychologischen Tests in Rechnung zu stellen, und abschätzt, wie groß die Wahrscheinlichkeit ist, eine Differenz der Größenordnung $Y–X$ allein aufgrund von Messfehlereinflüssen zu erhalten. Ist diese Wahrscheinlichkeit kleiner als ein vorher festgelegtes Signifikanzniveau von beispielsweise α=0.05, dann schließt man daraus, dass sich offensichtlich doch die wahren Werte von Y und X unterscheiden, sich mithin das in Frage stehende Merkmal verändert hat. Technisch betrachtet geht man bei diesem Signifikanztest so vor, dass man die Differenz $Y–X$ zu ihrer eigenen Standardabweichung $Std(Y–X)$ ins Verhältnis setzt. Man erhält dann den Reliable Change Index (im folgenden abgekürzt mit *RCI*):

$$RCI = \frac{Y - X}{Std(Y - X)} = \frac{Y - X}{Std(X)\sqrt{2(1 - \mathrm{Rel}(X))}}. \tag{1}$$

In dieser Formel kennzeichnet Rel(X) die sogenannte Reliabilität (Zuverlässigkeit) des verwendeten Messinstrumentes, $Std(X)$ die Standardabweichung des Messwerts X in der Population. Sind Y und X in der Population normalverteilt mit gleichen wahren Werten und Varianzen und gilt ferner Rel(X)=Rel(Y), dann ist der *RCI*-Index ebenfalls normalverteilt. Die Nullhypothese gleicher wahrer Werte von Y und X (H_0: $\tau_X = \tau_Y$) kann dann über die Standardnormalverteilung geprüft werden.

Steyer et al. (1997) kritisieren die Verwendung eines solchen einfachen Differenzmaßes zur Erfassung intraindividueller Veränderung, da bei diesem Vorgehen dem Messfehler-

problem nicht adäquat Rechnung getragen werde. Ihr Verbesserungsvorschlag basiert auf der folgenden Idee: Erhebt man bei einer Person zwei Messwerte, dann wird der zweite Messwert im Durchschnitt näher am Erwartungswert der Populationsverteilung liegen als der erste Messwert, auch wenn die wahren Werte der beiden Messungen sich nicht unterscheiden. Dieser Effekt ist in der Literatur unter dem Stichwort „Regression zur Mitte“ bekannt. Der zu erwartende „Rutsch“ der zweiten Messung in Richtung auf den Populationsmittelwert fällt dabei umso größer aus,

- je stärker die erste Messung vom Populationsmittelwert abweicht, und
- je geringer die Reliabilität des verwendeten Messinstrumentes ist, d.h. je stärker die Messungen von Zufallseinflüssen überlagert sind.

Um den Effekt der „Regression zur Mitte“ zu kontrollieren, schlagen Steyer et al. vor, den Posttestwert Y nicht mit dem Prätestwert X selbst zu vergleichen, sondern mit einem „bereinigten“ Posttestwert Y', den man im Durchschnitt ausgehend vom Prätestwert X erwarten würde, wenn nur Zufallseinflüsse wirksam sind, aber keine Veränderung der wahren Werte stattgefunden hat. Y' lässt sich mit Hilfe eines linearen Regressionsmodells ausgehend von X wie folgt vorhersagen (zur Herleitung und der dabei benötigten Annahmen siehe Steyer et al., 1997, S. 294):

$$Y' := E_0(Y \mid X) = E(X) + \mathrm{Rel}(X)[X - E(X)]$$

Die Differenz (Y–Y') wird dann (wie beim *RCI* auch) für die Prüfung zu ihrer Standardabweichung $Std(Y - Y') = Std[Y - E_0(Y \mid X)] = Std(X)\sqrt{1 - \mathrm{Rel}^2(X)}$ in Beziehung gesetzt. Man erhält den kompliziert aussehenden Ausdruck:

$$V_{\mathrm{infer}} := \frac{Y - E_0(Y \mid X)}{Std[Y - E_0(Y \mid X)]} = \frac{[Y - E(X)] - \mathrm{Rel}(X)[X - E(X)]}{Std(X)\sqrt{1 - \mathrm{Rel}^2(X)}}$$

Gehen wir aber ohne Einschränkung der Allgemeinheit zur Betrachtung z-standardisierter Messwerte über, vereinfacht sich die Formel für diesen Veränderungsindex entscheidend und lässt sich einfacher mit dem *RCI* vergleichen. Man erhält:

(2) $$V_{\mathrm{infer}} := \frac{Y - \mathrm{Rel}(X) \cdot X}{\sqrt{1 - \mathrm{Rel}^2(X)}}$$

Dieser Veränderungsindex kann unter den gleichen Annahmen, die im Kontext des *RCI* genannt wurden, mit Hilfe der Standardnormalverteilung auf Signifikanz geprüft werden.

Die „Bereinigung[1]“ des Differenzmaßes um den Effekt der Regression zur Mitte bei der Veränderungskenngröße V_{infer} nach Steyer et al. beeinflusst sowohl den Zähler als auch den Nenner des Ausdrucks. Dies erkennt man, wenn man V_{infer} und den *RCI* miteinander vergleicht:

- So ist der Nenner von V_{infer} immer kleiner oder höchstens so groß wie der Nenner des *RCI*.

[1] Inwieweit durch die von Steyer et al. (1997) vorgeschlagene Korrektur der Effekt der Regression zur Mitte tatsächlich adäquat kontrolliert wird und ihre Argumentation überzeugend ist, soll an dieser Stelle nicht weiter kritisch diskutiert werden. Siehe dazu Nachtigall & Suhl (2002b).

- Die Stärke der Korrektur des Zählers von V_{infer} im Vergleich zum Zähler des *RCI* hängt, wie man erkennen kann, zum einen von der Reliabilität des Messinstrumentes ab, zum anderen aber auch davon, wie extrem die Prätestmessung ausfällt: Je geringer die Reliabilität des Messinstrumentes ist und je extremer X ausfällt, desto stärker wird die Korrektur wirksam. In welche Richtung sich die Korrektur auswirkt, lässt sich dagegen schwieriger beurteilen: Je nachdem, ob $|x|>|y|$ gilt oder $|x|<|y|$, fällt der Zähler von V_{infer} kleiner oder auch größer aus als der Zähler des *RCI*. Für bestimmte Messwertkonstellationen kann es aber auch zu Vorzeichenumkehrungen kommen.

Eine Diskussion der Unterschiede zwischen V_{infer} und *RCI* unter inhaltlichen Gesichtspunkten findet man bei Nachtigall & Suhl (in Vorbereitung). Neben inhaltlichen Kriterien, was eine Veränderungskenngröße eigentlich genau abbilden soll, spielen für die Auswahl eines bestimmten Testes aber auch dessen statistische Eigenschaften eine wesentliche Rolle. Hier steht zu erwarten, dass sich die aufgezeigten Unterschiede zwischen V_{infer} und *RCI* ebenfalls in deren statistischen Eigenschaften niederschlagen:

- Beide Prüfgrößen (1) und (2) sind so konstruiert, dass der Signifikanztest über die Standardnormalverteilung durchgeführt werden kann. Sofern die Verteilungsannahmen zutreffend sind, wird das festgelegte Signifikanzniveau daher global betrachtet[2] bei beiden Vorgehensweisen eingehalten. Da die Standardabweichung der Veränderungskenngröße V_{infer} allerdings kleiner oder maximal so groß ist wie die Standardabweichung des *RCI*, steht zu erwarten, dass global betrachtet der Signifikanztest bei Verwendung von V_{infer} eine größere Power besitzt als bei Verwendung des *RCI*. Dieser Powervorteil dürfte dabei umso stärker ausgeprägt sein, je geringer die Reliabilität des verwendeten Messinstrumentes ist.

- Darüber hinaus dürften sich aber bei einer differenzierten Betrachtungsweise weitere Unterschiede zeigen. Wie oben dargestellt wurde, hängt die Stärke der Korrektur um den Effekt der Regression zur Mitte bei V_{infer} neben der Reliabilität des Messinstrumentes auch von der absoluten Größe des Prätestwertes X ab: Je extremer X ist, desto stärker wirkt sich die Korrektur im Vergleich zum *RCI* aus. Bei gegebener Differenz $Y–X$ kann dies durchaus zu unterschiedlichem Entscheidungsverhalten führen, je nachdem ob V_{infer} oder *RCI* verwendet wird. Dies spiegelt sich dann möglicherweise ebenfalls in den statistischen Eigenschaften des Tests wider.

Die Zielsetzung der durchgeführten Computersimulationen besteht darin, eben diese statistischen Eigenschaften der beiden Prüfgrößen genauer zu untersuchen: Ergeben sich tatsächlich, wie oben vermutet, global betrachtet Vorteile für V_{infer} hinsichtlich der Power? Und welche Eigenschaften ergeben sich, wenn man zu einer bedingten Betrachtungsweise übergeht, d.h. wenn man α-Fehler-Wahrscheinlichkeit und Power des Tests in Abhängigkeit vom Prätestwert X bzw. dessen Truescore-Wert τ_x betrachtet? Wird hier jeweils das α-Fehler-Niveau eingehalten, und ergeben sich hier für eine der beiden Prüfgrößen Vorteile hinsichtlich der Power, die als Argumente für die Auswahl der entsprechenden Prüfgröße herangezogen werden können?

3. Das Design der Simulationsstudie

Die Computersimulationen zur Untersuchung der statistischen Eigenschaften der beiden Veränderungskenngrößen *RCI* und V_{infer} wurden unter Verwendung der Programmiersprache

[2] „Globale Betrachtungsweise“ meint hier, dass alle Messwertkombinationsmöglichkeiten gemeinsam betrachtet werden, also insbesondere nicht danach unterschieden wird, wie extrem der Prätestwert X ausfällt.

Mathematica 3.0 durchgeführt. Mit der dort implementierten Zufallszahlen-Prozedur wurden zunächst normalverteilte Werte für die Truescorevariable τ_X der Prätestmessungen X mit dem Erwartungswert $E(\tau_X) = 0$ und der Varianz $Var(\tau_X) = 0.8$ erzeugt. Für die Messfehlervariablen ε_X und ε_Y wurden ebenfalls Normalverteilungen vorgegeben mit $E(\varepsilon_X) = E(\varepsilon_Y) = 0$ und $Var(\varepsilon_X) = Var(\varepsilon_Y) = 0.2$. Die Prätestvariable $X := \tau_X + \varepsilon_X$ ist dann normalverteilt mit $E(X)=0$ und $Var(X)=1$. Für die Reliabilität des Messinstrumentes ist damit $Rel(X)=0.8$ vorgegeben.
Da die Power eines Signifikanztests von der Größe des zu testenden Effektes abhängt, wurde bei den Simulationen auch die Differenz der Truescores von Post- und Prätestmessung $\Delta\tau := \tau_Y - \tau_X$ systematisch variiert. Die verwendeten Truescoredifferenzen sind in der nachfolgenden Tabelle 1 zusammengestellt. Bei gegebener Truescoredifferenz $\Delta\tau_j$ lässt sich nun die Posttestvariable Y über den Ansatz $Y := \tau_X + \Delta\tau_j + \varepsilon_Y$ erzeugen. Y ist normalverteilt mit $E(Y)= \Delta\tau_j$ und $Var(Y)=1$.

Tabelle 1: Bei den Simulationen verwendete Truescoredifferenzen

$\Delta\tau_j$	Truescoredifferenz (in Einheiten der Standardabweichung von τ_X)	Truescoredifferenz $\Delta\tau:=\tau_Y-\tau_X$
$\Delta\tau_1$	–2.0	–1.78885
$\Delta\tau_2$	–1.5	–1.34164
$\Delta\tau_3$	–1.0	–0.894427
$\Delta\tau_4$	–0.5	–0.447214
$\Delta\tau_5$	–0.25	–0.223607
$\Delta\tau_6$	0.0	0.0
$\Delta\tau_7$	0.25	0.223607
$\Delta\tau_8$	0.5	0.447214
$\Delta\tau_9$	1.0	0.894427
$\Delta\tau_{10}$	1.5	1.34164
$\Delta\tau_{11}$	2.0	1.78885

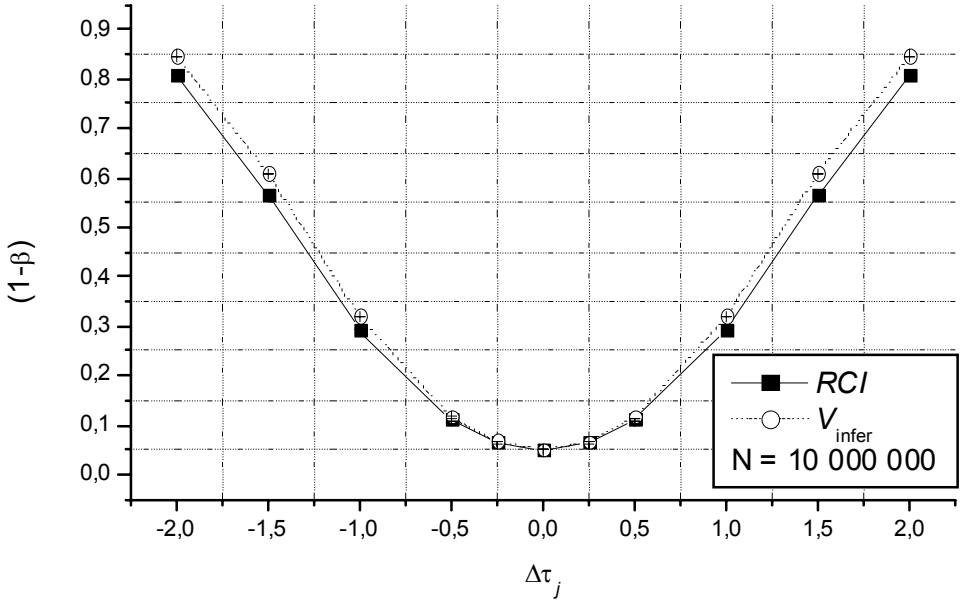
Für jede betrachtete Truescoredifferenz $\Delta\tau_j$ wurden auf dem oben beschriebenen Weg N = 10.000.000 Messwertpaare (x, y) erzeugt und für jedes Messwertpaar die Veränderungskenngrößen RCI und V_{infer} berechnet. Über die Standardnormalverteilung wurde dann jeweils die Nullhypothese „keine wahre Veränderung“ auf dem 5%-α-Fehler-Niveau geprüft: Gilt $|RCI| > 1.959964$ resp. $|V_{infer}| > 1.959964$, dann wird die Nullhypothese verworfen. Für die Truescoredifferenz $\Delta\tau_6=0$ erhält man über die relative Häufigkeit, mit der die Nullhypothese verworfen wird, eine Schätzung für die tatsächliche α-Fehler-Wahrscheinlichkeit. Für alle anderen Truescoredifferenzen $\Delta\tau_j \neq 0$ erhält man auf diesem Weg eine Schätzung der Power des Tests.

4. Die Ergebnisse der Simulationsstudien

Abbildung 1 zeigt die auf dem Simulationsweg gewonnenen Schätzungen für die Power und die α–Fehler-Wahrscheinlichkeit in Abhängigkeit von der vorgegebenen Truescoredifferenz. Auf der Abszisse sind die 11 verschiedenen Truescoredifferenzen aus Tabelle 1 abgetragen (jeweils ausgedrückt in Standardabweichungen der Truescoreverteilung von X; dies entspricht den Werten in der mittleren Spalte von Tabelle 1). Auf der Ordinate sind die Schätzungen der Power bzw. der α-Fehler-Wahrscheinlichkeit dargestellt. Jede Schätzung basiert auf N=10.000.000 Messwertpaaren. Ebenfalls in die Graphik eingetragen sind für jede Schätzung die zugehörigen 95%-Konfidenzintervalle: Da N sehr groß ist, sind die Konfidenzintervalle entsprechend klein; die Schätzungen sind nahezu punktgenau bzw. exakt. Die durchgezogene Linie zeigt die Schätzungen bei Verwendung des RCI, die gepunktete Linie die Schätzungen bei Verwendung von V_{infer}.

In der Abbildung 1 kann man zum einen sehen, dass sowohl bei Verwendung des *RCI* als auch bei Verwendung von V_{infer} das vorgegebene Signifikanzniveau von $\alpha = 0.05$ eingehalten wird: Für beide Veränderungsindizes ergibt sich für die Truescoredifferenz $\Delta\tau_6=0.0$ eine Schätzung der α-Fehler-Wahrscheinlichkeit von nahezu genau 0.05. Zum anderen kann man erkennen, dass sich tatsächlich wie erwartet Power-Vorteile für die Veränderungskenngröße V_{infer} im Vergleich zum *RCI* zeigen, da deren Ergebniskurve (gepunktete Linie) immer oberhalb der Kurve des *RCI* (durchgezogene Linie) liegt. Allerdings sind diese Powerunterschiede für kleine Truescoredifferenzen $|\Delta\tau_j| \leq 0.5$ vernachlässigbar gering; in der Graphik sind sie kaum erkennbar. Erst für große Differenzen $|\Delta\tau_j| \geq 1$ sind deutliche, wenn auch weiterhin kleine (<0.05) Unterschiede zu sehen.

Abbildung 1: Schätzungen der Power bzw. der α-Fehler-Wahrscheinlichkeit in Abhängigkeit von der vorgegebenen Truescoredifferenz $\Delta\tau_j$ für *RCI* (durchgezogene Linie) und V_{infer} (gepunktete Linie). Für $\Delta\tau_j=0$ erhält man eine Schätzung der α-Fehler-Wahrscheinlichkeit, für $\Delta\tau_j\neq 0$ Schätzungen der jeweiligen Power. Erläuterung siehe Text.

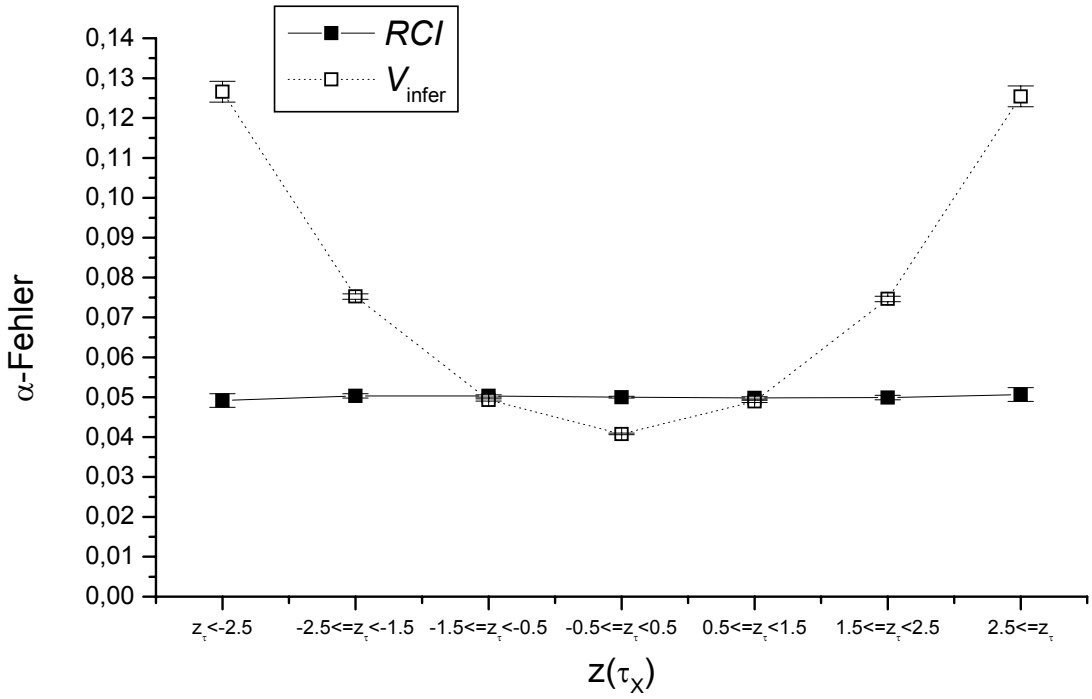


Doch welche Schätzungen für die α-Fehler-Wahrscheinlichkeit und die Power ergeben sich, wenn man nun zu einer „bedingten Betrachtungsweise“ übergeht, d.h. die Simulationsdaten in Abhängigkeit vom Truescorewert τ_x der Prätestmessung X auswertet? Wird auch dann bei beiden Prüfgrößen (1) und (2) das vorgegebene α-Fehler-Niveau eingehalten, und ergeben sich Unterschiede bzgl. der Power der Tests? Zur Beantwortung dieser Fragen wurden die Simulationsdaten noch einmal auf andere Art ausgewertet: Alle pro Truescoredifferenz $\Delta\tau_j$ per Simulation erzeugten Messwertpaare (x, y) wurden zunächst in 7 Gruppen eingeteilt. Die Gruppenzugehörigkeit richtet sich dabei nach der Stärke der Abweichung des Truescorewertes τ_x der Prätestmessung X vom Erwartungswert der Truescoreverteilung, gemessen in Standardabweichungen der Truescoreverteilung, also nach dem z-Wert z_τ von τ_x. Die verwendeten Intervallgrenzen lassen sich aus den Ergebnisgraphiken (Abbildung 2 und 3) ablesen. Anschließend wurde pro Gruppe ermittelt, wie oft die durchgeführten Signifikanztests eine Änderung der Truescores signalisierten: Für die vorgegebene Truescoredifferenz $\Delta\tau_j=0$ erhält man über die relative Häufigkeit pro Gruppe eine Schätzung der α-Fehler-Wahrscheinlichkeit in Abhängigkeit davon, wie extrem τ_x von $E(\tau_x)=0$ abweicht; für jedes $\Delta\tau_j\neq 0$ gewinnt man eine Schätzung für die Power des Test in Abhängigkeit von τ_x.

Abbildung 2 zeigt die Ergebnisse der Simulationsstudie bzgl. der α-Fehler-Wahrscheinlichkeit in Abhängigkeit von τ_x. Die durchgezogene Linie gibt die Schätzungen der α-Fehler-Wahrscheinlichkeit bei Verwendung des *RCI* wieder, die gestrichelte Linie die Schätzungen bei Verwendung von V_{infer}. Für jede Schätzung sind wiederum die 95%-Konfidenzintervalle mit dargestellt, um die Genauigkeit der Schätzungen zu veranschaulichen. Wie man erkennen kann, liegt die α-Fehler-Wahrscheinlichkeit bei Verwendung des *RCI* nahe bei dem vorgegebenen Signifikanzniveau $\alpha=0.05$, unabhängig davon, welchen Wert τ_x annimmt. Im Unterschied dazu zeigt sich bei den Ergebnissen für V_{infer} eine deutliche Abhängigkeit von τ_x: Für große Werte von τ_x, d.h. $|z_\tau|>1.5$, ist der tatsächliche α-Fehler deutlich größer als das vorgegebene nominelle α-Fehler-Niveau von $\alpha=0.05$; für extreme Werte von τ_x mit $|z_\tau|>2.5$

ist der tatsächliche α-Fehler mehr als doppelt so groß wie eigentlich vorgegeben. Demgegenüber ist für Werte von τ_x nahe dem Erwartungswert ($|z_\tau|<0.5$) der tatsächliche α-Fehler deutlich kleiner als das vorgegebene α-Fehler-Niveau von $\alpha=0.05$; das Testverhalten ist hier also bei Verwendung von V_{infer} konservativer als eigentlich gewünscht. Gemittelt über den gesamten Wertebereich von τ_x resultiert ein globaler α-Fehler von 0.05: Das konservativere Testverhalten für kleine τ_x kompensiert gewissermaßen das progressivere Testverhalten für große τ_x, da bei Normalverteilung kleine τ_x wesentlich häufiger auftreten als große τ_x.

Abbildung 2: α-Fehler-Wahrscheinlichkeit in Abhängigkeit von τ_x für die Prüfgrößen *RCI* (durchgezogene Linie) und V_{infer} (gepunktete Linie). Die waagerechten kleinen Striche kennzeichnen die 95%-Konfidenzintervalle für die relativen Häufigkeiten.

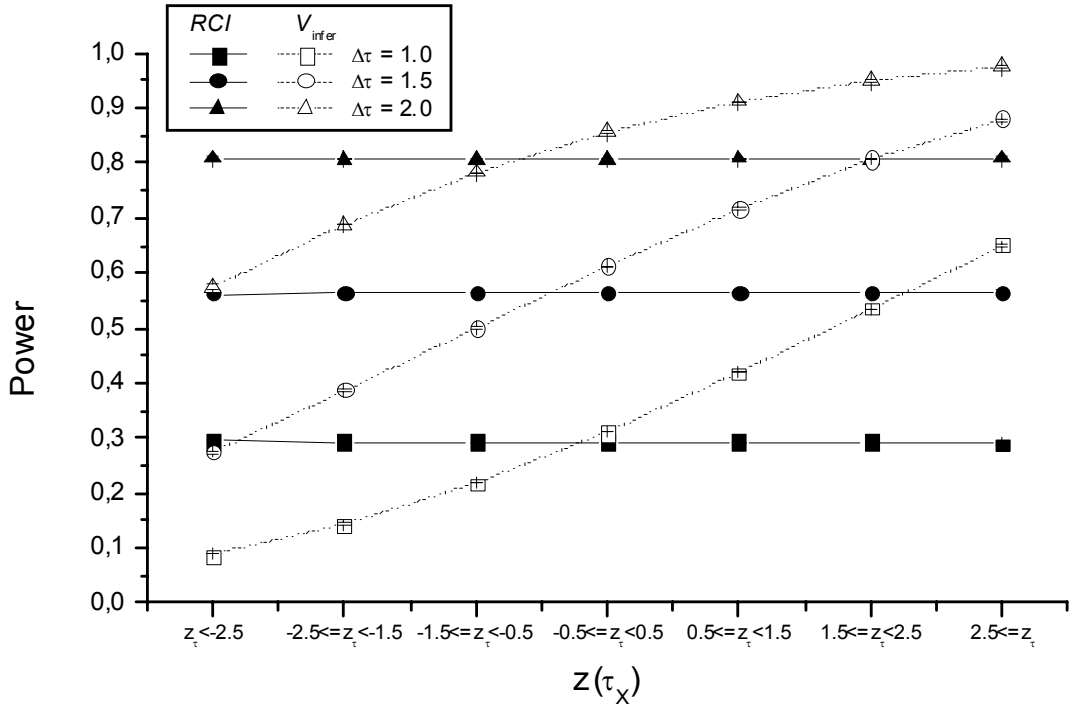


Betrachten wir nun die Simulationsergebnisse bezüglich der Power der Signifikanztests: Abbildung 3 zeigt die auf dem Simulationsweg gewonnenen Schätzungen der Power bei Verwendung von *RCI* und V_{infer} in Abhängigkeit von τ_x. Aus Platzgründen werden nur die Ergebnisse für die drei Truescoredifferenzen $\Delta\tau_9 = 1.0$ (Quadrat), $\Delta\tau_{10} = 1.5$ (Kreis) und $\Delta\tau_{11}=2.0$ (Dreieck), jeweils gemessen in Einheiten der Standardabweichung der Verteilung von τ_x, dargestellt; die Ergebnisse für die anderen Truescoredifferenzen $\Delta\tau_j \neq 0$ sind aber vergleichbar. Wiederum sind, um die Genauigkeit der Schätzungen zu veranschaulichen, die jeweiligen 95%-Konfidenzintervalle in die Graphik eingetragen.

Wie in der Abbildung 3 zu erkennen ist, hängt die Power des Signifikanztests bei Verwendung der Prüfgröße *RCI* (durchgezogene Linie mit ausgefüllten Symbolen), wie zu erwarten ist, von der Größe der vorgegebenen Truescoredifferenz $\Delta\tau_j$ ab: Je größer $\Delta\tau_j$, desto größer ist auch die Power des Tests. Allerdings zeigt sich keine Abhängigkeit der Power vom Truescore τ_x der Prätestmessung: Alle durchgezogenen Linien verlaufen horizontal parallel zur Abszisse. Im Unterschied dazu ist in den Ergebnissen für die Prüfgröße V_{infer} (gepunktete Linien mit nicht ausgefüllten Symbolen) eine deutliche Abhängigkeit der Power von τ_x zu

erkennen: Zwar gilt auch hier „je größer $\Delta\tau_j$, desto größer die Power des Tests“, die einzelnen Linien verlaufen jedoch nicht horizontal, sondern weisen eine starke Steigung auf.

Abbildung 3: Power in Abhängigkeit von τ_x für die Prüfgrößen *RCI* (durchgezogene Linie mit ausgefüllten Symbolen) und V_{infer} (gepunktete Linie mit nicht ausgefüllten Symbolen). Die unterschiedlichen Symbole veranschaulichen unterschiedliche vorgegebene Truescoredifferenzen $\Delta\tau_j$. Die waagerechten kleinen Striche kennzeichnen die 95%-Konfidenzintervalle für die relativen Häufigkeiten.



5. Diskussion der Ergebnisse

Die durchgeführten Computersimulationen zeigen,

- dass sowohl für den *RCI* als auch für V_{infer} global betrachtet das vorgegebene Signifikanzniveau eingehalten wird; für V_{infer} ergeben sich aufgrund der geringeren Varianz der Prüfgröße geringe Powervorteile im Vergleich zum *RCI* (vgl. Abbildung 1),
- dass sich bei einer differenzierteren Betrachtungsweise in Abhängigkeit vom Truescore τ_x der Prätestmessung für die Prüfgröße V_{infer} unerwartete statistische Eigenschaften offenbaren: Während bei Verwendung des *RCI* das Signifikanzniveau unabhängig von τ_x eingehalten wird und auch die Power des Test bei gegebener Truescoredifferenz $\Delta\tau_j$ nicht von τ_x abhängt, zeigt sich in den Ergebnissen für V_{infer} ein anderes Bild: Sowohl der α-Fehler als auch die Power variieren deutlich in Abhängigkeit von τ_x (siehe Abbildungen 2 und 3).

Hier stellt sich nun einerseits die Frage, worauf diese Unterschiede zwischen den Prüfgrößen zurückzuführen sind, und andererseits, welche praktischen Konsequenzen sich aus den Simulationsergebnisse ableiten lassen.

Zur Beantwortung der ersten Frage betrachten wir zunächst Erwartungswert und Varianz der beiden Prüfgrößen für einen vorgegebenen Wert τ_a der Truescorevariablen τ_x. Gilt die Nullhypothese H_0: $\tau_x = \tau_y$, dann lässt sich zeigen, dass die Prüfgröße *RCI* weiterhin standardnormalverteilt ist, insbesondere gilt $E(RCI \mid \tau_X = \tau_a) = 0$ und $Std(RCI \mid \tau_X = \tau_a) = 1$ (siehe

Anhang). Die Verwendung der Standardnormalverteilung als Prüfgrößenverteilung ist also auch bei der bedingten Betrachtungsweise ($\tau_X=\tau_a$) korrekt. Ein anderes Bild zeigt sich für die Prüfgröße V_{infer}: Hier ergibt sich

$$E(V_{\text{infer}} \mid \tau_X = \tau_a) = \frac{\tau_a(1-\text{Rel}(X))}{\sqrt{1-\text{Rel}^2(X)}} \text{ und}$$

$$Var(V_{\text{infer}} \mid \tau_X = \tau_a) = \frac{(1-\text{Rel}(X))\cdot(1+\text{Rel}^2(X))}{1-\text{Rel}^2(X)}$$

(Herleitung siehe Anhang).

V_{infer} ist damit also nicht mehr standardnormalverteilt, und die Verwendung der Standardnormalverteilung als Prüfgrößenverteilung führt zu falschen Abschätzungen der α-Fehler-Wahrscheinlichkeit. Da der Erwartungswert der Prüfgröße vom Truescore der Prätestmessung abhängt, variiert der tatsächliche α-Fehler in Abhängigkeit von τ_x. In der Konsequenz variiert auch die Power des Tests in Abhängigkeit von τ_x.
In der Praxis lässt sich der Entscheidungsbias nicht beheben bzw. ausrechnen, da der wahre Wert der Prätestmessung nicht bekannt ist. Der Anwender wird also bei Verwendung von V_{infer} letztlich nicht angeben können, wie groß sein tatsächliches Fehlentscheidungsrisiko im Einzelfall ist. Dies ist eine Situation, die der Intention der statistischen Prüfung von intraindividuellen Veränderungen zuwiderläuft. Aus diesem Grund ist für die Praxis dringend von der Verwendung von V_{infer} abzuraten, zumal mit dem *RCI* eine Veränderungsprüfgröße mit wesentlich besseren statistischen Eigenschaften zur Verfügung steht.

6. Literatur

Jacobson, N. S. & Truax, P. (1991). Clinical significance. A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology, 59, 1, 12-19.

Lienert, G. (1961). Testaufbau und Testanalyse. (3. Auflage). Weinheim: Beltz.

Nachtigall, C. & Suhl, U. (2002 a). Warum kompliziert, wenn es einfach geht. Teil 1: Zur Analyse intraindividueller Veränderung. *metheval report 4 (3).*

Nachtigall, C. & Suhl, U. (2002 b). Der Regressionseffekt. Mythos und Wirklichkeit. *metheval report 4 (2).*

Steyer, R. & Eid, M. (1993). Messen und Testen. Berlin, Heidelberg, New York: Springer.

Steyer, R., Hannöver, W., Telser, Ch. & Kriebel, R. (1997). Zur Evaluation intraindividueller Veränderung. Zeitschrift für Klinische Psychologie, 26, 291-299.

7. Anhang: Ableitung der Verteilungseigenschaften der Veränderungskennwerte bei bedingter Betrachtungsweise

Gezeigt werden soll in diesem Abschnitt, dass der *RCI* auch bei einer bedingten Betrachtungsweise seine Verteilungseigenschaften behält, während sich bei V_{infer} Erwartungswert und Varianz ändern. Zur Vereinfachung der Ableitung werden zunächst nur die Zähler der beiden Prüfgrößen betrachtet. Vorausgesetzt werden die üblichen Eigenschaften gleicher Fehlervarianzen, der Normalverteilung der Messfehler sowie der Unkorreliertheit der Messfehler. Ferner gehen wir ohne Einschränkung der Allgemeinheit zur Vereinfachung von z-standardisierten Messwerten X und Y aus. Die Ableitung erfolgt unter der Annahme der Gültigkeit der Nullhypothese H_0: $\tau_X=\tau_Y$. Hergeleitet werden Erwartungswert und Varianz der Prüfgrößen für $\tau_X=\tau_a$.

Herleitung für die Prüfgröße *RCI*:

Erwartungswert:

$$\begin{aligned} E(Y-X \mid \tau_X=\tau_a) &= E(\tau_Y+\varepsilon_Y-\tau_X-\varepsilon_X \mid \tau_X=\tau_a) \\ &= E(\varepsilon_Y-\varepsilon_X \mid \tau_X=\tau_a) \\ &= E(\varepsilon_Y \mid \tau_X=\tau_a)-E(\varepsilon_X \mid \tau_X=\tau_a) \\ &= 0 \end{aligned}$$

Daraus folgt sofort

$$E\left(\frac{Y-X}{\sqrt{2(1-\text{Rel}(X))}} \mid \tau_X=\tau_a\right)=\frac{1}{\sqrt{2(1-\text{Rel}(X))}}E(Y-X \mid \tau_X=\tau_a)=0$$

Varianz:

$$\begin{aligned} Var(Y-X \mid \tau_X=\tau_a) &= Var(\tau_Y+\varepsilon_Y-\tau_X-\varepsilon_X \mid \tau_X=\tau_a) \\ &= Var(\varepsilon_Y-\varepsilon_X \mid \tau_X=\tau_a) \\ &= Var(\varepsilon_Y \mid \tau_X=\tau_a)+Var(\varepsilon_X \mid \tau_X=\tau_a) \\ &= 2(1-\text{Rel}(X)) \end{aligned}$$

Daraus lässt sich nun ableiten, dass die folgende Beziehung gilt:

$$\begin{aligned} Var\left(\frac{Y-X}{\sqrt{2(1-\text{Rel}(X))}} \mid \tau_X=\tau_a\right) &= \frac{1}{2(1-\text{Rel}(X))}Var(Y-X \mid \tau_X=\tau_a) \\ &= \frac{1}{2(1-\text{Rel}(X))}\cdot 2(1-\text{Rel}(X)) \\ &= 1 \end{aligned}$$

Damit ist gezeigt, dass bei der bedingten Betrachtungsweise $\tau_X=\tau_a$ für den *RCI* weiterhin $E(RCI|\tau_X=\tau_a)=0$ und $Var(RCI|\tau_X=\tau_a)=1$ gilt.

Herleitung für die Prüfgröße V_{infer}:

Erwartungswert:

$$\begin{aligned} E(Y-\text{Rel}(X)\cdot X \mid \tau_X=\tau_a) &= E(\tau_Y+\varepsilon_Y-\text{Rel}(X)\cdot(\tau_X-\varepsilon_X) \mid \tau_X=\tau_a) \\ &= E(\tau_Y-\text{Rel}(X)\cdot\tau_X \mid \tau_X=\tau_a)+E(\varepsilon_Y-\text{Rel}(X)\cdot\varepsilon_X \mid \tau_X=\tau_a) \\ &= \tau_a(1-\text{Rel}(X)) \end{aligned}$$

Daraus folgt dann weiter

$$E\left(\frac{Y-\mathrm{Rel}(X)\cdot X}{\sqrt{1-\mathrm{Rel}^2(X)}}\,|\,\tau_X=\tau_a\right)=\frac{1}{\sqrt{1-\mathrm{Rel}^2(X)}}E\left(Y-\mathrm{Rel}(X)\cdot X\,|\,\tau_X=\tau_a\right)$$
$$=\frac{\tau_a(1-\mathrm{Rel}(X))}{\sqrt{1-\mathrm{Rel}^2(X)}}$$

Varianz:

$$Var(Y-\mathrm{Rel}(X)\cdot X\,|\,\tau_X=\tau_a)=Var(\tau_Y+\varepsilon_Y-\mathrm{Rel}(X)\cdot(\tau_X+\varepsilon_X)\,|\,\tau_X=\tau_a)$$
$$=Var(\varepsilon_Y\,|\,\tau_X=\tau_a)+\mathrm{Rel}^2(X)Var(\varepsilon_X\,|\,\tau_X=\tau_a)$$
$$=(1-\mathrm{Rel}(X))+\mathrm{Rel}^2(X)(1-\mathrm{Rel}(X))$$
$$=(1-\mathrm{Rel}(X))\cdot(1+\mathrm{Rel}^2(X))$$

Daraus folgt nun für die bedingte Varianz der Prüfgröße V_{infer}

$$Var\left(\frac{Y-\mathrm{Rel}(X)\cdot X}{\sqrt{1-\mathrm{Rel}^2(X)}}\,|\,\tau_X=\tau_a\right)=\frac{1}{1-\mathrm{Rel}^2(X)}Var(Y-\mathrm{Rel}(X)\cdot X\,|\,\tau_X=\tau_a)$$
$$=\frac{(1-\mathrm{Rel}(X))\cdot(1+\mathrm{Rel}^2(X))}{1-\mathrm{Rel}^2(X)}$$

Die Herleitung zeigt, dass V_{infer} bei der bedingten Betrachtungsweise nicht mehr standardnormalverteilt ist. Insbesondere der bedingte Erwartungswert $E(V_{\mathrm{infer}}|\tau_X=\tau_a)$ ist in der Regel von Null verschieden, da er von τ_a abhängt. Nur für den Ausnahmefall $\tau_a=0$ gilt auch $E(V_{\mathrm{infer}}|\tau_X=\tau_a)=0$.

AUTORENHINWEIS

Korrespondenz bzgl. dieses Artikels ist zu richten an:

Dr. Christof Nachtigall
Friedrich Schiller Universität Jena
Institut für Psychologie
Am Steiger 3, Haus 1
D-07743 Jena
Tel.: 03641 945234
Fax: 03641 945232
Email: christof.nachtigall@uni-jena.de