

Beipackzettel.

»Bei Problemen fragen Sie ihren Bortz, Dorsch, Fisseni, Jäger, Grubitzsch oder Petermann.« - Sowohl die Fragen als auch die Antworten der Fragebögen sind sicherlich nicht voraussetzungslos zu verstehen. Auch wenn die Fragebögen in dieser Sammlung z.T. recht ausführlich beantwortet wurden, so bleibt doch immer genug ›drumherum-Wissen‹ übrig, das es im Zweifel nachzuschlagen gilt. Die kleinen Glossare / Lektüren

- ›Sadistik‹ &
- Diagnostik

sollen dies erleichtern.

Der kleine Sadistik-Glossar

Quellen:

Dorsch Psychologisches Wörterbuch. Häcker, H. & Stapf, K. H. (Hg.). Hans Huber: Bern, Göttingen, Toronto, Seattle. (1998)

Statistik für Sozialwissenschaftler. Bortz, J (Hg.). Springer Verlag: Berlin, Heidelberg.. (1999)

Arithmetisches Mittel. → *Mittelwert*

Chi-Quadrat-Methoden (χ^2 -Methoden). Signifikanztest zur Analyse von Häufigkeitsunterschieden. (Bortz)

Clusteranalyse. Heuristisches Verfahren zur systematischen Klassifizierung der Objekte einer gegebenen Objektmenge. (Bortz)

Deskriptivstatistik. Teil der Statistik, dessen Aufgabe in der Charakterisierung durch bestimmte Kennwerte (Mittelwert, Streuung, Korrelation u.a.) besteht. Darstellende oder beschreibende Statistik.

disjunkt. Zwei einander ausschließende (d.h. keine gemeinsamen Elementarereignisse beinhaltenden) Ereignisse sind disjunkt. Ihr Durchschnitt ($A \cap B$) ist die leere Menge. (Bortz)

diskret. Ein Merkmal ist diskret, wenn es nicht kontinuierliche, sondern nur bestimmte Werte annehmen kann. (Bortz)

empirisches Relativ. Aus empirischen Objekten bestehendes Relationensystem (im Gegensatz zu einem → *numerischen Relativ*). (Bortz)

Erwartungswert. „Mittelwert“ einer theoretischen (nicht empirischen) Verteilung einer Zufallsvariablen; bezeichnet durch den Buchstaben μ („mü“). (Bortz)

Exhaustion. Modifikation oder Erweiterung einer Theorie aufgrund von Untersuchungsergebnissen, die die ursprüngliche Form der Theorie falsifizieren. (Bortz)

Faktor. Im Rahmen der → *Varianzanalyse* ist ein Faktor eine unabhängige Variable, deren Bedeutung für eine abhängige Variable überprüft wird. (Bortz)

Faktorenanalyse. Datenreduzierendes Verfahren zur Bestimmung der dimensionalen Struktur korrelierter Merkmale. (Bortz)

Faktorladung. Korrelation zwischen einer Variablen und einem Faktor. (Bortz)

F-Test. Statistischer Signifikanztest, der zwei Stichprobenvarianzen miteinander vergleicht. (Bortz)

goodnes of fit test. Eindimensionaler → χ^2 -Test.

Grundgesamtheit (Population). Alle potentiell untersuchbaren Einheiten, die ein gemeinsames Merkmal aufweisen (Bsp.: Bewohner einer Stadt, Frauen, dreisilbige Substantive). (Bortz)

homomorph. Lässt sich ein → *empirisches* durch ein → *numerisches Relativ* so abbilden, dass eine bestimmte Relation im empirischen Relativ der Relation im numerischen Relativ entspricht, bezeichnet man diese Abbildung als homomorph. Bsp.: empirisches Relativ: Mathekenntnisse der Schüler einer Klasse; numerisches Relativ: Mathenoten (Bortz)

Inferenzstatistik (schließende Statistik). Statistik, die auf der Basis von Stichprobenergebnissen induktiv allgemeingültige Aussagen formuliert. Zur Inferenzstatistik zählen die Schätzung von → *Populationsparametern* (Schließen) und die Überprüfung von Hypothesen (Testen). (Bortz)
| Teil der Statistik, mit der Aufgabe, aus Stichprobenwerten auf die entsprechenden Populationswerte zu schließen (Bereichsschätzungen, Prüfung von Hypothesen). Schluß- und Prüfstatistik. (Dorsch)

Klassifikation. Mit Klassifikationsverfahren kann man überprüfen, zu welche von k Gruppen ein Individuum aufgrund eines individuellen Merkmalsprofils am besten passt (Diskriminanzanalyse). (Bortz)

Konfidenzintervall. Derjenige Bereich eines Merkmals, in dem sich 95% bzw. 99% aller möglichen → *Populationsparameter* befinden, die den empirische ermittelten Stichprobenkennwert erzeugt haben können. M.a.W., der in der Stichprobe ermittelte Mittelwert gehört mit 95%iger bzw. 99%iger Wahrscheinlichkeit zu einer Population, deren Parameter μ sich im berechneten Intervall befindet.

(Bortz)

Konsistenz. Kriterium der Parameterschätzung: ein Schätzwert ist konsistent, wenn er sich mit wachsendem Stichprobenumfang (n) dem zu schätzenden Parameter nähert. (Bortz)

Kontrollvariable (Moderatorvariable). Merkmal, das bei einem (Quasi-)Experiment weder abhängige noch unabhängige Variable ist, sondern nur miterhoben wird, um im Nachhinein prüfen zu können, ob es einen Einfluss auf das Untersuchungsergebnis hatte. (Bortz)

Kovarianz. Mittelwert aller Produkte von korrespondierenden Abweichungen zweier gemeinsam erhobener Variablen; m.a.W., die Kovarianz ist ein Maß für den Grad des Miteinander-Variierens zweier Messwertreihen x und y . Eine positive Kovarianz besteht, wenn viele Versuchspersonen bei einem hohen x -Wert auch einen hohen y -Wert haben, eine negative Kovarianz besteht, wenn viele Versuchspersonen bei einem hohen x -Wert einen niedrigen y -Wert haben. (Bortz)

Kreuzvalidierung. Verfahren, bei dem zwei Regressionsgleichungen aufgrund von zwei Teilstichproben bestimmt werden, deren Vorhersagekraft in bezug auf die Kriteriumswerte der anderen Stichprobe geprüft wird. (Bortz)

Kriteriumsvariable. Variable, die mittels einer \rightarrow *Prädiktorvariablen* und einer \rightarrow *Regressionsgleichung* vorhergesagt werden kann. (Bortz)

Maximum-likelihood-Methode. Methode, nach der \rightarrow *Populationsparameter* so geschätzt werden, dass die „Wahrscheinlichkeit“ (Likelihood) des Auftretens der beobachteten Daten maximiert wird. (Bortz)

Median. Derjenige Wert einer Verteilung, der die Gesamtzahl der Fälle in zwei Hälften teilt, so dass 50% aller Werte unter dem Median, 50% aller Fälle über ihm liegen. (Bortz)

Mittelwert (arithmetisches Mittel). Derjenige Wert, der sich ergibt, wenn die Summe aller Werte einer Verteilung durch die Gesamtzahl der Werte (n) geteilt wird. (Bortz)

Modalwert. Derjenige Wert einer Verteilung, der am häufigsten vorkommt. In einer graphischen Darstellung der Verteilung deren Maximum. Eine Verteilung kann mehrere Modalwerte (und somit Maxima) besitzen. (Bortz)

Moderatorvariable. \rightarrow *Kontrollvariable* (Bortz)

Multivariate Methoden. Mit multivariaten Methoden werden Hypothesen geprüft, die auf das Zusammenwirken vieler abhängiger und unabhängiger Variablen beziehen. (Bortz)

Normalverteilung. Wichtigste Verteilung der Statistik; festgelegt durch die Parameter μ (Erwartungswert) und σ (Streuung); glockenförmig, symmetrisch, zwischen den beiden Wendepunkten ($\mu \pm 1\sigma$) liegen ca. 68% der gesamten Verteilungsfläche (Standardnormalverteilung). (Bortz)

numerisches Relativ. Aus Zahlen bestehendes Relationensystem (z.B. die Menge der reellen Zahlen); mit einem numerischen Relativ lässt sich ein \rightarrow *empirisches Relativ* \rightarrow *homomorph* abbilden. (Bortz)

Operationalisierung. Umsetzung einer eher abstrakten Variable bzw. eines theoretischen Konstruktes in ein konkret messbares Merkmal; Bsp.: Operationalisierung der Variable „mathematische Begabung“ durch die Variable „Mathematiknote“¹. Wichtig ist, dass die operationalisierte Variable die abstrakte Variable tatsächlich widerspiegelt. (Bortz)

Parameter. Kennwerte einer theoretischen Verteilung oder Grundgesamtheit (im Gegensatz zu Stichprobenkennwerten) wie z.B. Erwartungswert, Streuung etc. Bezeichnung durch griechische Großbuchstaben. (Bortz)

Prädiktorvariable. Variable, mittels derer unter Verwendung der \rightarrow *Regressionsgleichung* eine Vorhersage über eine andere Variable (Kriteriumsvariable) gemacht werden kann. (Bortz)

Prozentränge. In Prozentwerte umgerechnete kumulierte Häufigkeiten. (Bortz)

Range. \rightarrow *Variationsbreite*. (Bortz)

Regression, multiple. Vorhersage eines Kriteriums mittels eines linearen Gleichungsmodells aufgrund mehrerer \rightarrow *Prädiktorvariablen*. (Bortz)

¹ Übernommenes Beispiel, mit einigen Fehlern: Mathe-Note soll zunächst nur eine Leistung konstatieren, die keine notwendige Verbindung zu etwas wie „Begabung“ haben muss; weiterhin kann je nach Notengebung die Leistung eher absolut (kriteriumsorientierte Leistungsmessung) oder relativ (Notengebung nach Normalverteilung) bemessen sein.

Regressionsgleichung. (Meist lineare) Gleichung, die die Beziehung zwischen zwei Merkmalen x und y charakterisiert. Mit Hilfe der Regressionsgleichung kann ein Vorhersagewert für y (\rightarrow *Kriteriumsvariable*) geschätzt werden, wenn x (\rightarrow *Prädiktorvariable*) bekannt ist. Die Regressionsgleichung wird so ermittelt, dass sie die Summe der quadrierten Vorhersagefehler minimiert. (Bortz)

Signifikanzniveau (α -Fehler-Niveau). Die Irrtumswahrscheinlichkeit, die ein Untersuchungsergebnis maximal aufweisen darf, damit die Alternativhypothese als bestätigt gelten kann. Im allgemeinen spricht man von einem signifikanten Ergebnis, wenn die Irrtumswahrscheinlichkeit höchstens 5%, von einem sehr signifikanten Ergebnis, wenn sie höchstens 1% beträgt. (Bortz)

Standardabweichung (Streuung). Wurzel aus der \rightarrow *Varianz*; bezeichnet durch s für Stichproben und σ für theoretische Verteilungen (Bortz). | Statistischer Kennwert der Streuung oder Dispersion einer Verteilung. Die Standardabweichung ist die positive Quadratwurzel aus dem Durchschnitt der quadratischen Abweichungen der Maßzahlen vom Mittelwert (Dorsch).

Standardfehler / Standardmessfehler. Streuung einer Stichprobenkennwerteverteilung. Sie informiert darüber, wie unterschiedlich Stichprobenkennwerte (z.B. Mittelwerte) von Stichproben aus einer Population bei einem gegebenen Stichprobenumfang sein können. (Bortz) | Maß für den \rightarrow *Stichprobenfehler* (Dorsch)

Standardnormalverteilung. Normalverteilung mit \rightarrow *Erwartungswert* (μ) 0 und \rightarrow *Standardabweichung* (σ) 1. Jede Normalverteilung kann durch \rightarrow *z-Transformation* in die Standardnormalverteilung überführt werden, was den Vergleich verschiedener Normalverteilungen ermöglicht. (Bortz)

Standardschätzfehler. Kennzeichnet die Streuung der y -Werte um die Regressionsgerade und ist damit ein Gütemaßstab für die Genauigkeit der Regressionsvorhersagen. Je kleiner der Standardschätzfehler, desto genauer ist die Vorhersage. (Bortz) | Fehler der gemacht wird, wenn man von einem Testwert auf einen Kriteriumswert schließt. Er kommt zustande, weil der Kriteriumswert auch mit Fehlervarianz behaftet ist. (Dorsch)

Statistik. Quantitative Erfassung und Analyse von Merkmalen, deren zahlenmäßige Ausdrücke bei Messwiederholungen (am gleichen Merkmalsträger oder gleichartigen Merkmalsträgern) unter sonst gleichen Bedingungen durch nichtsystematische Zufallseinflüsse variieren. \rightarrow *Deskriptivstatistik*, \rightarrow *Inferenzstatistik*

Stichprobenfehler. Die Abweichung eines aufgrund einer Stichprobe berechneten statistischen Kennwerts vom entsprechenden Wert der Population. Die Größe zufälliger Stichprobenfehler nimmt mit steigender Stichprobengröße ab. (Dorsch)

Streuung. \rightarrow *Standardabweichung*. (Bortz)

Suppressorvariable. Variable, die den Vorhersagebeitrag einer (oder mehrerer) anderer Variablen erhöht, indem sie irrelevante Varianzen in den (der) anderen Variable(n) unterdrückt (multiple Korrelation). (Bortz).

t-Test für abhängige Stichproben. Statistischer Signifikanztest, der zwei Gruppen, die nicht unabhängig voneinander ausgewählt wurden (parallelisierte Stichproben oder Messwiederholung) auf einen Unterschied bezüglich ihrer Mittelwerte eines intervallskalierten Merkmals untersucht. (Bortz)

t-Test für unabhängige Stichproben. Statistische Signifikanztest, der zwei Gruppen, die unabhängig voneinander ausgewählt wurden, auf einen Unterschied bezüglich ihrer Mittelwerte eines intervallskalierten Merkmals untersucht. (Bortz)

Unabhängigkeit. Zwei Ereignisse sind voneinander unabhängig, wenn das Auftreten des eines Ereignisses nicht davon beeinflusst wird, ob das andere eintritt oder nicht. Mathematisch drückt sich dies darin aus, dass die Wahrscheinlichkeit für das gemeinsame Auftreten beider Ereignisse dem Produkt der Einzelwahrscheinlichkeiten der beiden Ereignisse entspricht. (Bortz).

Varianz. Summe der quadrierten Abweichungen aller Messwerte einer Verteilung vom \rightarrow *Mittelwert*, dividiert durch die Anzahl aller Messwerte (n). Maß für die Unterschiedlichkeit der einzelnen Werte einer Verteilung. (Bortz) | Statistischer Kennwert der Streuung oder Dispersion einer Verteilung. Die Varianz ist gleich dem Quadrat der Standardabweichung. (Dorsch)

Varianzanalyse. Allgemeine Bezeichnung für eine Verfahrensklasse zur Überprüfung von Unterschiedshypothesen. Man unterscheidet ein- und mehrfaktorielle Varianzanalysen, uni- und multivariate Varianzanalysen, hierarchische und nichthierarchische Varianzanalysen sowie Kovarianzanalysen. (Bortz) | Das Verfahren basiert auf einer Zerlegung der Variation (Streuungszerlegung) der abhängigen Variablen in verschiedene Varianzkomponenten, die auf bestimmte Ursachen, z.B. auf

die Variation je einer unabhängigen Variablen allein oder auf die Kombination der Klassen mehrerer unabhängiger Variablen (Wechselwirkung) oder auf Messfehler, zurückgeführt werden können. (Dorsch)

Variationsbreite („range“). Gibt an, in welchem Bereich sich die Messwerte eines Kollektivs bzw. einer Stichprobe befinden; ergibt sich als Differenz des größten und des kleinsten Werts der Verteilung. (Bortz)

z-Transformation. Ein Wert einer beliebigen Verteilung wird durch Subtraktion des Mittelwerts und anschließende Division durch die Standardabweichung der Verteilung in einen z -Wert transformiert. Eine z -transformierte Verteilung hat einen Mittelwert von 0 und eine Standardabweichung von 1. Beliebige Normalverteilungen werden durch die z -Transformation in die Standardnormalverteilung überführt. (Bortz)

z-Wert. Ein durch lineare statistische Transformation gewonnener Standardwert, der auf eine Verteilung mit einem Mittelwert $M = 100$ und einer Standardabweichung $s = 10$ bezogen ist (Dorsch).

Diagnostik

[Die im Folgenden vorgenommenen eigenen Hervorhebungen wie auch die Tilgung von ursprünglich vorhandenen Hervorhebungen sind nicht immer ausdrücklich gekennzeichnet.]

Klassische Testtheorie (KTT) und theoretische Grundlagen

Klassische Testtheorie. »Diese Theorie ist eine Fehlertheorie. Sie erlaubt daher nicht – wie die Meßmodelle – eine Aussage darüber, daß die formalen eigenschaften des Modells mit den psychologischen Gesetzmäßigkeiten (auf denen das Modell aufbaut) einhergehen. Die Klassische Testtheorie zählt daher nicht zu den Meßmodellen.« (JÄGER & PETERMANN 1992, 276).

»Die Klassische Testtheorie (...) beschäftigt sich mit der Frage, wie aus einer Anzahl von Verhaltensbeobachtungen x_{vi} von Versuchspersonen (Vp) v in bestimmten Situationen i auf die wahre Ausprägung (›true score‹) τ_v eines Persönlichkeitsmerkmals von Vp v geschlossen werden könne. Dabei geht sie von Überlegungen aus, welche sich zunächst auf den wahren Wert τ_{vi} von nur einer Person in nur einer Situation i beziehen. Diese Grundannahmen werden als Axiome der KT bezeichnet.« (JÄGER & PETERMANN 1992, 310)

Axiomatik

(weitestgehend wortgleich nach: JÄGER & PETERMANN 1992, 310ff)

Das **Existenzaxiom** besagt, daß der true-score τ_{vi} als Erwartungswert von x_{vi} existiere:

$$\tau_{vi} = E(x_{vi}) \quad (1)$$

Das **Verknüpfungssaxium** besagt, daß jede Messung x_{vi} aus einem wahren Wert τ_{vi} und einem zufälligen Fehlerwert ϵ_{vi} zusammengesetzt sei:

$$x_{vi} = \tau_{vi} + \epsilon_{vi} \quad (2)$$

Die Verbindung der Axiome (1) und (2) zeigt, dass der Zufallsfehler ϵ_{vi} den Erwartungswert Null hat:

$$E(\epsilon_{vi}) = 0 \quad (3)$$

Die zwei Axiome enthalten eine bekannte Größe, nämlich die beobachtbare Messung x_{vi} und zwei unbekannte Größen, nämlich den wahren Wert τ_{vi} und den Fehlerwert ϵ_{vi} , welche aus den Beobachtungen nicht erschlossen werden können. Schätzbar sind hingegen der ›wahre Testwert τ_v ‹ von VP v sowie die Fehlervarianz $\sigma^2(\epsilon)$.

Schätzung des ›wahren Testwertes τ_v ‹

Die Beurteilung einer Messung auf ihre Richtigkeit hin erfolgt zumeist durch eine Wiederholung der Messung und die anschließende Mittelung der Ergebnisse. Diese Vorgehensweise ist in psychologischen Untersuchungen problematisch, da Lerneffekte die Zufälligkeit der Fehlergrößen in Frage stellen. Die einzelnen Meßwerte werden deshalb aus vergleichbaren Situationen (operationalisiert in Items) erhoben. Um den Testwert der VP v zu bestimmen, summiert man die einzelnen Meßwerte x_{vi} zu einem Summenwert x_v :

$$x_v = \sum_{i=1}^m x_{vi} \quad (4)$$

Um zu erschließen, ob x_v eine geeignete Punktschätzung des ›wahren Testwertes‹ τ_v darstellt, wird der Erwartungswert von (4) untersucht:

$$E(x_v) = E\left(\sum_i x_{vi}\right) \quad (5a)$$

$$E(x_v) = \sum_i E(x_{vi}) \quad (5b)$$

Das Einsetzen von (1) in (5b)

$$E(x_v) = \sum_i \tau_{vi} \quad (5c)$$

$$E(x_v) = \tau_v \quad (5d)$$

zeigt, daß der Erwartungswert von x_v dem wahren Testwert τ_v entspricht. Es ist zu ersehen, daß die Meßwertsumme x_v als Punktschätzung $\hat{\tau}_v$ für den wahren Testwert τ_v einer bestimmten Person v verwendet werden kann:

$$\hat{x}_v = \tau_v \quad (6)$$

Die statistische Unsicherheit dieser Punktschätzung wird mit Hilfe eines geeigneten Konfidenzintervalls eingegrenzt.

Lokale stochastische Unabhängigkeit

Betrachtet man zwei Fehlervariablen ϵ_{vi} und ϵ_{vj} für zwei Items i und j (bzw. ϵ_{vi} und ϵ_{wi} für zwei Personen v und w), so läßt sich über den Zufallscharakter der einzelnen Variablen (vgl. Axiom (2)) hinausgehend auch annehmen, daß die Fehlervariablen paarweise unabhängig verteilt sind. Daraus ergibt sich, daß die Fehlerwerte sowohl zweier beliebiger Items i und j als auch zweier beliebiger Personen v und w unkorreliert sind:

$$\rho(\epsilon_{vi}, \epsilon_{vj}) = 0 \quad (7a)$$

$$\rho(\epsilon_{vi}, \epsilon_{wi}) = 0 \quad (7b)$$

Diese Eigenschaft der *lokalen stochastischen Unabhängigkeit* kann nur dann erfüllt sein, wenn

- unabhängige Items konstruiert werden, die Beantwortung des einen also keinen Einfluß auf die Beantwortung des anderen hat (7a) und
- die Itembearbeitung durch unabhängige Personen erfolgt (7b).

Trennschärfe und Schwierigkeitsindex

Die in (5c) vorgenommene Summierung über die Testitems setzt deren *Eindimensionalität* (Itemhomogenität, Variablenkongenerität) voraus, um zu vermeiden, dass ›Äpfel und Birnen‹ addiert werden. Eine empirische Rechtfertigung dieser Summierbarkeit ist auf der Basis der KT nicht möglich, sondern erst dann, wenn man die Annahmen der KT zu jenen des KLA-Modells erweitert.

Die KT muss sich hingegen ersatzweise damit begnügen, für jedes Einzelitem einen hinreichend engen korrelativen Zusammenhang zum Gesamtestwert x_v sicherzustellen. Man bestimmt ihn für jedes der m Items auf der Basis von n Messwertpaaren als Ite-Test-Korrelation

$$r_{it} = r(x_{vi}, x_v) \quad \text{Trennschärfe} \quad (8)$$

wobei der Summenwert x_v – vor allem bei geringer Itemanzahl – ohne das betreffende Item i gebildet werden soll (›part-whole-correction‹).

Der Koeffizient (8) wird als **Trennschärfe** der Items i bezeichnet. Liegen hohe positive Trennschärfen vor, so kann man davon ausgehen, daß die einzelnen Items Ähnliches wie der Gesamtest messen.

Hohe Trennschärfen können nur erzielt werden, wenn die Items weder allzu leicht noch allzu schwierig sind. Deshalb ist es notwendig, die **Itemschwierigkeiten** P_i zu kontrollieren: Werden die Reaktionsmöglichkeiten x_{vi} eines gestuften Items mit 0 bis k kodiert, so kann man den Schwierigkeitsindex P_i als

$$P_i = \frac{\sum_{v=1}^n x_{vi}}{[n \cdot k]} \cdot 100 \quad (9)$$

berechnen, nämlich als Quotient tatsächlicher und maximal möglicher Summe. (Die Multiplikation mit dem Faktor 100 führt zu einem Wertebereich zwischen 0 und 100.)

Im Fall von nur zweistufigen Items, vereinfacht sich der Index (9) auf den prozentualen Anteil ›richtiger‹ Lösungen:

$$P_i = \frac{N_R}{N} \quad \text{mit } N_R = \text{›Zahl der Richtiglöser‹, } N = \text{›Zahl der Probanden‹}$$

Man erkennt, dass der Schwierigkeitsindex numerisch um so größer wird, je mehr Vpn ein Item symptomatisch beantworten bzw. lösen können. Folglich kennzeichnet diese Statistik eher die ›Leichtigkeit‹ eines Items.

Schätzung der Fehlervarianz: Reliabilität

Zur Schätzung der ›Fehlervarianz‹ werden die Reaktionen auf eine Stichprobe von Items über die gesamte Stichprobe hinweg betrachtet. Die Variable der einzelnen Testwerte x_v bezeichnet man dann als x , die Variable der wahren Testwerte τ_v als τ und die Variable der Abweichungen ϵ_v zwischen den x_v und den τ_v als Fehlervariable ϵ (vgl. dazu das Verknüpfungssaxiom (2)).

Da man nur die Testwerte x_v , aber nicht die wahren Testwerte τ_v kennt, lassen sich die einzelnen Fehlerwerte ϵ_v nicht direkt bestimmen. Dennoch ist es möglich, eine pauschale Abschätzung der Varianz $\sigma^2(\epsilon)$ der Fehlervariablen ϵ vorzunehmen.

Dazu wird die Testwertvariable x gemäß dem Verknüpfungssaxiom (2) zerlegt in

$$x = \tau + \epsilon \quad (10)$$

und die Varianz von (10) untersucht. Man erhält

$$\sigma^2(x) = \sigma^2(\tau + \epsilon) \quad (11a)$$

$$\sigma^2(x) = \sigma^2(\tau) + \sigma^2(\epsilon) + 2\sigma^2(\tau, \epsilon) \quad (11b)$$

wobei $\sigma(\tau, \epsilon)$ die Kovarianz von τ und ϵ bezeichnet. Da diese Kovarianz wegen (3) gleich Null ist, folgt daß die **Testwertvarianz** $\sigma^2(x)$ aus einem ›wahren‹ Varianzanteil $\sigma^2(\tau)$ und einem Fehlervarianzanteil $\sigma^2(\epsilon)$ zusammengesetzt ist

$$\sigma^2(x) = \sigma^2(\tau) + \sigma^2(\epsilon) \quad (12)$$

Die Varianz $\sigma^2(\tau)$ bemißt diejenige Variation, welche durch die unterschiedlichen Merkmalsausprägungen der Vpn hervorgebracht wird, die Varianz $\sigma^2(\epsilon)$ jene Variation, die auf Fehler bei den Messungen zurückzuführen ist.

In der Klassischen Testtheorie ist die Meßgenauigkeit das wichtigste Gütekriterium eines Tests. Sie wird mit Hilfe eines Koeffizienten bemessen, der **Reliabilitätsindex** (Rel) heißt und als Verhältnis von wahrer Varianz $\sigma^2(\tau)$ und Gesamtvarianz $\sigma^2(x)$ definiert ist:

$$Rel = \frac{[\sigma^2(\tau)]}{[\sigma^2(x)]} \quad (13)$$

Ein Test ist umso meßgenauer, je größer der wahre Varianzanteil $\sigma^2(\tau)$ an der Gesamtvarianz $\sigma^2(x)$. Umgekehrt gilt (vgl. (12)) auch, dass bei zunehmender Fehlervarianz $\sigma^2(\epsilon)$ die Reliabilität abnimmt.

Paralleltest-Reliabilität

Um die unbekannte Varianz $\sigma^2(\tau)$ zu schätzen, zieht man die Testwertvariablen x_p und x_q zweier beliebiger Tests p und q heran und betrachtet deren Korrelation, welche als Kovarianz $\sigma(x_p, x_q)$ beider Tests, geteilt durch deren Standardabweichung $\sigma(x_p)$ und $\sigma(x_q)$

$$\rho(x_p, x_q) = \frac{[\sigma(x_p, x_q)]}{[\sigma(x_p) \cdot \sigma(x_q)]} \quad (14)$$

definiert ist. Da die Kovarianz zweier Tests wegen (7) und (3) gerade so groß ist wie die Kovarianz ihrer wahren Werte, kann (14) als

$$\rho(x_p, x_q) = \frac{[\sigma(\tau_p, \tau_q)]}{[\sigma(x_p) \cdot \sigma(x_q)]} \quad (15)$$

geschrieben werden. Handelt es sich bei x_p und x_q um die Testwerte zweier Paralleltests (Testes gleicher wahrer Werte und gleicher Testwertestreuungen, im Idealfall desselben Tests) mit $\tau_p = \tau_q = \tau$ und $\sigma(x_p) = \sigma(x_q) = \sigma(x)$, so bemisst die Kovarianz im Zähler von (15) gerade die gesuchte Varianz $\sigma^2(\tau)$ und das Produkt zweier gesuchter Standardabweichungen $\sigma(x)$ im Nenner die Testwertestruenanz $\sigma^2(x)$. Folglich kann man (15) für Paralleltests umformulieren als

$$\rho(x_p, x_q) = \frac{[\sigma(\tau_p, \tau_q)]}{[\sigma(x_p) \cdot \sigma(x_q)]} \quad (16)$$

Ein Vergleich mit Formel (13) zeigt, dass das Varianzverhältnis **Reliabilität** gleich der Korrelation eines Tests mit seinem Paralleltest ist.

Die auf Stichprobendaten basierende Schätzung \hat{Rel} wird häufig als r_{tt} (Test-Test-Korrelation, Paralleltest-Korrelation)

$$\hat{Rel} = r_{tt} = r(x_p, x_q) \quad (17)$$

berechnet.

Die Reliabilität hat einen Wertebereich zwischen 0 und 1, wobei der Maximalwert 1 bedeutet, dass der Test fehlerfrei mißt, der Minimalwert 0, dass die Testwertestruenanz nur aus Fehlervarianz besteht.

Interne Konsistenz

Mit der Reliabilitätsbestimmung als Paralleltest-Korrelation r_{tt} nach (17) gehen einige Probleme einher. Entweder man verwendet denselben Test zweimal (Test-Retest) und gewinnt das Paralleltest-Ergebnis durch Meßwiederholung. Dieser *ideale* Paralleltest kann zu einer überhöhten Retest-Reliabilitätsschätzung führen, sofern das zweite Testergebnis nicht als unabhängige Messung, sondern als schlichte Erinnerungsreplikation des ersten Testergebnisses aufgefaßt wird, was auf der Basis des bereits erwähnten Gedächtnisspeichers nicht unrealistisch erscheint.

Mehr Informationen über die Korrekte Zusammensetzung eines Tests erhält man, wenn die Reliabilität als interne Konsistenz des Tests bestimmt wird. Hatte die Retest- bzw. Paralleltest-Reliabilität zwei Meßzeitpunkte erfordert, um die Übereinstimmung der Testpunktwerte überprüfen zu können, so benötigt man für die konsistenzanalytische Reliabilitätsbestimmung nur einen Meßzeitpunkt.

Zunächst hatte man die Idee, einen einzigen Test in zwei parallele Halbttests (split-half) zu zerlegen, für diese die Halbttest-Reliabilität zu berechnen und nach der klassischen Formel von Spearman & Brown zur Gesamtest-Reliabilität aufzuwerten. Cronbach (1951) schlug vor, den Test nicht nur in zwei, sondern in mehrere parallele Testteile zu zerlegen und die Gesamtest-Reliabilität mit Hilfe des Alpha-Koeffizienten, welcher auf Guttman (1945) basiert, zu kalkulieren.

Am meisten Information wird hingegen verwertet, wenn man einen Test entsprechend seiner Itemzahl nicht nur in zwei oder mehrere, sondern in so viele – nämlich m – Testteile zerlegt, wie der Test Items enthält. Diese im engeren Sinn konsistenzanalytische Reliabilitätsschätzung, welche auf einen varianzanalytischen Ansatz von Hoyt (1941) zurückgeht, untersucht, wie gut die m Teiltestergebnisse einander replizieren. Sie soll im folgenden näher beschrieben werden.

Die konsistenzanalytische Reliabilitätsschätzung geht wieder von Definition (13) aus: Das Verhältnis Rel aus wahrer Varianz und Gesamtvarianz (18a) wird unter Verwendung der Varianzzerlegung (12) umgeformt in (18b) und weiter durch Kürzen von $\sigma^2(x)$ in (18c).

$$Rel = \frac{[\sigma^2(\tau)]}{[\sigma^2(x)]} \quad (18a)$$

$$Rel = \frac{[\sigma^2(x) - \sigma^2(\epsilon)]}{[\sigma^2(x)]} \quad (18b)$$

$$Rel = 1 - \frac{[\sigma^2(\epsilon)]}{[\sigma^2(x)]} \quad (18c)$$

Aus (18c) ist zu erkennen, dass eine hohe interne Konsistenz Rel dann vorliegt, wenn der Fehlervarianzanteil $\sigma^2(\epsilon)$ an der Testwertvarianz $\sigma^2(x)$ gering ist.

Anhand eines Beispiels soll die praktische Schätzung der einzelnen Varianzkomponenten gezeigt werden – das Beispiel wird hier nur in Form der verwendeten Formeln wiedergegeben.

Zur varianzanalytischen Auswertung sind vier Quadratsummen Q zu berechnen:

1. Die totale Quadratsumme Q_{total} als Summe der Abweichungsquadrate aller Meßwerte x_{vi} von dem Gesamtmittelwert \bar{x}

$$Q_{total} = \sum_{v=1}^n \sum_{i=1}^m (x_{vi} - \bar{x})^2 \quad (19)$$

2. Die Vpn-Quadratsumme Q_{Vpn} als m-fache Summe der Abweichungsquadrate aller Zeilenmittelwerte \bar{x}_v vom Gesamtmittelwert \bar{x}

$$Q_{Vpn} = m \cdot \sum_{v=1}^n (\bar{x}_v - \bar{x})^2 \quad (20)$$

3. Die Itemquadratsumme Q_{Items} als n-faches der Summe der Abweichungsquadrate aller Spaltenmittelwerte \bar{x}_i vom Gesamtmittelwert \bar{x}

$$Q_{Items} = n \cdot \sum_{i=1}^m (\bar{x}_i - \bar{x})^2 \quad (21)$$

4. Die Fehler- oder Restquadratsumme Q_{Rest} , nämlich der nicht auf Vpn und Items rückführbare Teil an der totalen Quadratsumme, als Differenz von Q_{total} , Q_{Vpn} und Q_{Items}

$$Q_{Rest} = Q_{total} - Q_{Vpn} - Q_{Items} \quad (22)$$

Das in (18c) benötigte Varianzverhältnis $[\sigma^2(\epsilon)]/[\sigma^2(x)]$ wird geschätzt, indem man als Zähler Q_{Rest} durch die Fehlerfreiheitsgrade $df_{Rest} = (n - 1) \cdot (m - 1)$ und als Nenner Q_{Vpn} durch die Personenfreiheitsgrade $df_{Vpn} = n - 1$ dividiert wird. Die zwei Ausdrücke schätzen jeweils das m-fache von $\sigma^2(\epsilon)$ bzw. $\sigma^2(x)$. Wegen $m/m = 1$ kann das Notieren von m im Zähler und Nenner entfallen. Somit resultiert Rel dann nach (18c) als

$$\hat{Rel} = 1 - \frac{\left[\frac{Q_{Rest}}{df_{Rest}} \right]}{\left[\frac{Q_{Vpn}}{df_{Vpn}} \right]} \quad (23)$$

Standardmessfehler und Konfidenzintervall

Ist der Reliabilitätskoeffizient Rel berechnet oder einem Testmanual entnommen worden, so kann die Varianzzerlegung (12) bei bekannter Testwertvarianz $\sigma^2(x)$ sehr einfach als

$$\hat{\sigma}^2(x) = Rel \hat{\sigma}^2(x) + (1 - Rel) \cdot \hat{\sigma}^2(x) \quad (24a)$$

$$\hat{\sigma}^2(x) = \hat{\sigma}^2(\tau) + \hat{\sigma}^2(\epsilon) \quad (24b)$$

vorgenommen werden. Löst man (24) nach der Fehlervarianz $\hat{\sigma}^2(\epsilon)$ auf als

$$\hat{\sigma}^2(\epsilon) = \hat{\sigma}^2(x) - Rel \hat{\sigma}^2(x) \quad (25a)$$

$$\hat{\sigma}^2(\epsilon) = \hat{\sigma}^2(x) \cdot (1 - Rel) \quad (25b)$$

und zieht man aus (25b) die Wurzel, so findet man den **Standardmessfehler** $\sigma(\epsilon)$ als

$$\sigma(\epsilon) = \sigma(x) \cdot \sqrt{1 - Rel} \quad (26)$$

Der Standardmessfehler ermöglicht es, das weiter oben erwähnte **Konfidenzintervall um die Punktschätzung (6) des wahren Testwertes τ_v zu bilden**: Bei Vorliegen großer Vpn-Stichproben findet man das Konfidenzintervall unter der Annahme normalverteilter Fehler mit Hilfe der Standardnormalverteilung (z-Verteilung) als

$$\tau_v - z_{\alpha/2} \cdot \sigma(\epsilon) \leq \tau_v \leq \tau_v + z_{\alpha/2} \cdot \sigma(\epsilon) \quad (27)$$

wobei der wahre Testwert τ_v mit einer statistischen Sicherheit von $(1 - \alpha)$ in diesem Intervall zu liegen kommt.

Objektivität – Reliabilität – Validität

Objektivität

»Objektivität bezeichnet das Maß, wie weit in der diagnostischen Situation eine **Standardisierung** des gesamten Testvorgangs gelingt.« (FISSENI 1997, 66) Ziel der Standardisierung ist es vor allem, die einzelnen Schritte von der jeweils untersuchenden Person unabhängig zu machen. Unterschieden werden drei Arten von Objektivität:

1. **Durchführungsobjektivität.** »Die Objektivität der Durchführung betrifft Raum und Zeit der diagnostischen Situation, die kognitiv-emotionale Verfassung des Probanden, darüber hinaus die Instruktion, welche die Testvorgabe und den Verlauf der Anwendung regelt.« (FISSENI 1997, 67)
2. **Auswertungsobjektivität.** »Auswertungsobjektivität besteht darin, daß gleichen Itemantworten gleiche numerische Werte (Scores) zugeordnet werden.« (FISSENI 1997, 68) »Die Auswertungsobjektivität läßt sich überprüfen, indem verschiedene Auswerter dasselbe Antwortprotokoll kodieren. Die Übereinstimmung kann korrelativ oder varianzanalytisch geschätzt werden.« (a.a.O.)
3. **Interpretationsobjektivität.** »Die Objektivität der Interpretation betrifft den Grad der Eindeutigkeit, mit der verschiedene Anwender dem gleichen numerischen Wert (dem Test-Score) die gleiche Merkmalsausprägung zuordnen. Wenn der Testautor Merkmalsbezeichnungen vorgibt und der Auswerter sie übernimmt, ist formal die Objektivität der Interpretation gegeben – es handelt sich um eine Sprachregelung.« (FISSENI 1997, 68)

Reliabilität

»Reliabilität charakterisiert das Meßinstrument Test unter dem Aspekt der Präzision. Implizit sind damit zwei Anteile angesprochen: wahrer Wert und Fehlerwert. Von beiden Anteilen her ist Reliabilität definiert worden:

- Reliabilität ist die *Meßgenauigkeit* des Instruments unter Absehung vom Inhalt.
- Reliabilität gilt als Bestimmung des *Meßfehlers*, mit dem die Testwerte behaftet sind, unabhängig davon, für welchen Inhalt die Werte stehen.« (FISSENI 1997, 70)

Validität

Objektivität – Reliabilität – Validität: logische Beziehungen zwischen den drei Gütekriterien (Rost 1996, 39f)

»Zwischen den drei Gütekriterien eines Tests bestehen verschiedene *logische Beziehungen*, die sich unter bestimmten mathematischen Annahmen sogar in Formeln beschreiben lassen. Und zwar ist die Objektivität eine logische Voraussetzung für die Reliabilität und diese wiederum ist die logische Voraussetzung für die externe Validität.

Ein Test, der bei einem anderen Testleiter oder in einem anderen Raum bei denselben Personen gänzlich andere Resultate erbringt, also nicht objektiv ist, kann auch keine hohe Meßgenauigkeit haben, d.h. nicht reliabel sein.

Ebenso kann ein Test mit einer sehr geringen Meßgenauigkeit (Reliabilität) keine besonders hohe *externe Validität* erreichen. Soll z.B. ein Test entwickelt werden, der die Schulleistung vorherzusagen gestattet, so kann diese Vorhersage nicht besonders gut ausfallen, wenn der Test nur sehr ungenau mißt.

Eine solche Voraussetzung besteht nicht zwischen der Meßgenauigkeit und der *internen Validität*. Auch ein ungenau messender Test kann intern valide sein.

Andererseits besteht zwischen Meßgenauigkeit, interner und externer Validität auch ein *kontradiktorisches Verhältnis*:

Das Streben nach einer möglichst hohen Meßgenauigkeit bei der Testentwicklung kann in einem Widerspruch stehen zum Ziel einer möglichst hohen Validität. Dieser Widerspruch ergibt sich daraus, daß sich die Meßgenauigkeit im allgemeinen dadurch steigern läßt, daß man den *Test verlängert*, d.h. zusätzliche Items aufnimmt (...).

Durch eine *Testverlängerung*, die den Test rein theoretisch beliebig genau machen könnte, können Items hineinkommen, die einen etwas anderen Aspekt der latenten Variablen ansprechen, es können Bearbeitungseffekt wie Ermüdung, Konzentrationsmangel, Wechsel der Antwortstrategie, Erinnerungseffekte, Lerneffekte und ähnliches eintreten. Diese Effekte können sowohl die präexperimentelle Theorie über das Antwortverhalten, d.h. das Testmodell, in seiner Gültigkeit einschränken, als auch die Korrelation mit dem Validitätskriterium, also die externe Validität, beeinträchtigen.

Auch die Ziele einer möglichst hohen internen und externen Validität können bei der Testentwicklung miteinander in Konflikt stehen. So läßt sich die interne Validität im allgemeinen dadurch steigern, daß man den Test *homogener* macht, d.h. möglichst ähnliche Aufgaben auswählt. Damit erfaßt man aber eine sehr eng gefaßte, spezielle Personeneigenschaft, die nur noch sehr geringe Korrelationen mit einem Validitätskriterium aufweist.

Die immanenten Widersprüche zwischen Reliabilität und Validität werden auch als *Reliabilitäts-Validitäts-Dilemma* der Testtheorie bezeichnet (...). Dieses Dilemma ist letztlich die Ursache für den weiterverbreiteten Argwohn, daß Tests entweder mit einer hohen Präzision etwas völlig Irrelevantes messen, oder eine Personeneigenschaft in ihrer ganzen Breite, aber völlig unzuverlässig erfassen.«

Validität (ext.)

•

Reliabilität

•

Objektivität

Testbatterien vs. Testprofil

Entscheidungstheorie.

»Der Entscheidungstheoretische Ansatz erweitert den Rahmen der klassischen Testtheorie. (...)

- Die klassische Testtheorie bemißt den Wert von Verfahren nach dem Grade ihrer *Objektivität, Reliabilität und Validität*.
- Dagegen bewertet der entscheidungstheoretische Ansatz ein Verfahren unter dem Aspekt von *Kosten und Nutzen*: »Das Kriterium für den Wert eines Tests ... ist nicht so sehr ein Grad der Ge-

nauigkeit, die er selber hat, vielmehr der Beitrag, den er für das Urteil leistet« (Cronbach & Gleser, 1965, 148).« (FISSENI 1997, 373; Herv.H JG)

Bayes-Theorem.

»Mit der Bayes-Statistik möchte man aus der Kenntnis früherer Fälle eine Vorhersage über die Erfolgswahrscheinlichkeit eines aktuellen Falles treffen. Das Bayes-Theorem bildet einen Algorithmus, mit dem man die a-posteriori-Wahrscheinlichkeit auf der Grundlage von Informationen über die Gesamtgruppe bestimmen kann. Dies ist erforderlich, da es sich gezeigt hat, daß Personen, die über alle relevanten Informationen verfügen, in ihrem intuitiven Urteil die a-posteriori-Wahrscheinlichkeit überschätzen (...).« (JÄGER & PETERMANN 1992, 306).

Konfidenzintervall (Vertrauensintervall) & Standardmeßfehler, Standardschätzfehler

»Das Vertrauensintervall eines beobachteten Meßwerts informiert darüber, *innerhalb welcher Grenzen der tatsächliche Meßwert bei festgelegter Irrtumswahrscheinlichkeit unter Berücksichtigung der an das jeweilige diagnostische Verfahren gebundenen Meßfehler variieren kann.*« (JÄGER & PETERMANN 1992, 253)

»Bei deterministischer Gleichsetzung von gemessenem Merkmal (*numerischem Relativ*) und Ausprägung des in Frage stehenden Merkmals (*empirischem Relativ*) im Sinne der Klassischen Testtheorie, wird das durch Meßfehler verursachte Vertrauensintervall durch den Standardmeßfehler s_e festgelegt. Der Standardmeßfehler variiert in Abhängigkeit von der Standardabweichung der beobachteten Meßwerte s_t und der Reliabilität r_{tt} . Er errechnet sich nach der Formel

$$s_e = s_t \cdot \sqrt{1 - r_{tt}}$$

Wird ein Test eingesetzt, um die Ausprägung eines in Frage stehenden Kriterienmerkmals zu erfassen (bzw. zu prognostizieren), läßt sich das Vertrauensintervall eines korrelativ erschlossenen Kriterienwerts c durch den Standardschätzfehler s_{ct} eingrenzen. Der Standardschätzfehler steigt nach der Formel

$$s_{ct} = s_c \cdot \sqrt{1 - r_{tc}^2}$$

mit zunehmender Standardabweichung der Kriterienwerte s_c . Er fällt mit zunehmender Validität, definiert als Korrelation zwischen Test und Kriterium (r_{tc}).« (JÄGER & PETERMANN 1992, 253f)

Für jeden Testwert X ist das **Vertrauensintervall** bei einer Irrtumswahrscheinlichkeit von $p = 0,05$ bzw. $p = 0,01$ festzulegen

$$CL_X (\text{confidential limits}) = X \pm z_{\alpha/2} \cdot s_e$$

$$CL_X = X \pm 1.96 \cdot s_e \quad (\text{für } p = 5\%)$$

$$CL_X = X \pm 2.58 \cdot s_e \quad (\text{für } p = 1\%)$$

CL_X = Unsicherheitsbereich eines Testpunktwerts bei $p = 0,05$ bzw. $p = 0,01$ Irrtumswahrscheinlichkeit

X = individueller Testpunktwert eines Pb

s_e = Standardmessfehler (Messfehler) eines Tests

Rechnung.

$$CL = 120 \pm 1.96 \cdot s_e \quad s_e = 6.7 \quad CL = 106 ; 133$$

- Der wahre IQ-Wert liegt zwischen 106 und 133.

Stichprobenunabhängigkeit und »spezifische Objektivität« bei RASCH

»spezifische Objektivität«. »Eine besondere Eigenheit des Modells von Rasch ist, daß es *spezifisch objektive* Vergleiche ermöglicht; d.h. der Unterschied der Fähigkeit ξ_v und ξ_w je zweier Per-

sonen kann unabhängig davon bestimmt werden, welche Aufgaben eines modellkonformen Itempools dafür herangezogen werden, bzw. umgekehrt, der Vergleich je zweier Aufgaben bezüglich ihrer Schwierigkeiten σ_i und σ_j ist unabhängig davon, welche Personenstichprobe dafür verwendet wird – die Schätzungen der Parameter sind stichprobenunabhängig, weil die Wahl der Stichprobe aus einer bestimmten Population für die statistische Inferenz dieser Parameter keine Rolle spielt.« (JÄGER & PETERMANN 1992, 324)

Stichprobenunabhängigkeit. »Diese Stichprobenunabhängigkeit bzw. die dem Modell von Rasch inwohnende *Spezifische Objektivität* ist die wesentliche Grundlage von Modelltests: Bei Modellgültigkeit müssen die Parameterschätzungen $\hat{\sigma}$ in verschiedenen Teilstichproben statistisch gleich sein. Sind sie es wenigstens für ein Item nicht, dann gilt das Modell von Rasch nicht, d.h. irgendeine seiner Voraussetzungen ist verletzt – in jedem Fall ist die Anzahl in bestimmter Richtung beantworteter Items keine erschöpfende Statistik für die gesuchte Eigenschaft einer Person, sondern sinnlos: Gleiche Scores drücken nicht gleiche Eigenschaften aus.« (JÄGER & PETERMANN 1992, 325)

Literatur

- Fisseni, H.-J. (1997): Lehrbuch der psychologischen Diagnostik. Göttingen: Hogrefe
Jäger, R. S. & Petermann, F. (1992): Psychologische Diagnostik: ein Lehrbuch. Weinheim: Psychologie-Verlags-Union