Zusammenfassender Überblick zur KTT

.

Kompaktkurs Testtheorie, Mainz, SS 09
U. Mortensen

Vorbemerkung: Es werden die zentralen Begriffe der KTT zusammengefasst. Der Text soll eine Hilfestellung bei der Beschäftigung mit der KTT liefern und nicht als ein für die Klausur hinreichendes Skriptum des Skriptums aufgefasst werden.

Inhaltsverzeichnis

1 Überblick

Alle Formeln der Klassischen Testtheorie (KTT) sind Konsequenzen der von Gulliksen (1950) zuerst explizit gemachten *Axiome* der KTT.

1. **Axiome:** Für einen Testwert X gilt demnach

$$X = \tau + \varepsilon, \quad \text{mit } \mathbb{E}(X) = \tau, \ \mathbb{V}(X) = \mathbb{V}(\varepsilon) = \sigma_x^2 \tag{1}$$

d.h. X ist eine zufällige Veränderliche mit dem Erwartungswert τ und der Varianz $\sigma_x^2 = \mathbb{V}(\varepsilon)$. Die "Axiome" sind: (i) $\mathbb{E}(\varepsilon) = 0$ (eine unmittelbare Folgerung aus (1), (ii) $\mathbb{K}(\tau, \varepsilon) = 0$, d.h. die Kovarianz von wahren Werten und Fehlern ist gleich Null; hier muß berücksichtigt werden, dass auch τ eine zufällige Veränderliche ist: τ ist zwar eine Konstante für eine gegebene Person, wenn man aber zufällig eine Person aus einer Population wählt, wählt man damit auch zufällig einen Wert von τ, (iii) $\mathbb{K}(\varepsilon_x, \tau_y) = 0$, die Fehler in einem Test sind unkorreliert mit den wahren Werten in einem anderen Test, (iv) $\mathbb{K}(\varepsilon_x, \varepsilon_y) = 0$, Fehler in verschiedenen Tests sind unkorreliert.

Für eine Person $a \in \mathcal{P}$ nimmt τ den Wert τ_a an. τ_a repräsentiert sicherlich die Fähigkeit bzw. die Ausprägung des gemessenen Merkmals bei der Person a, allerdings in Abhängigkeit von den Items, mit denen das Merkmal erfasst werden soll. τ_a repräsentiert die Ausprägung des Merkmals relativ zu den Items $I_1, \ldots, I_n$; in τ_a gehen die Schwierigkeiten κ_g der Items I_g ein, vergl. Abschnitt 2.1. Andererseits assoziiert man mit dem Begriff des "wahren Wertes" einer Person einfach die Ausprägung des gemessenen Merkmals,*unabhängig von den Items, mit denen man es messen will.* Diese

Ausprägung läßt sich durch den Parameter θ repräsentieren, wie er in der Definition der Itemfunktion auftritt, und zwar separiert von der Itemschwierigkeit κ_g, vergl. Punkt 3.

2. **Reliabilität** definiert die Zuverlässigkeit eines Tests; der Reliabilitätskoeffizient ist ein Maß für die Enge des Zusammenhanges von X und τ: Generell gilt $X = \beta\tau + \alpha + \varepsilon$, und nach (1) gilt $\beta = 1$, $\alpha = 0$. Andererseits weiß man aus der Regressionsrechnung, dass die Korrelation $\rho_{x\tau}$ zwischen X und τ durch $\rho_{x\tau} = \beta\sigma_\tau/\sigma_x$ gegeben ist. Wegen $\beta = 1$ hat man dann

$$\mathrm{Rel}(X) = \rho^2_{x\tau} = \frac{\sigma^2_\tau}{\sigma^2_x} \tag{2}$$

$\mathrm{Rel}(X)$ definiert die Reliabilität eines Tests; sie ist also durch das Quadrat der Korrelationskoeffizienten $\rho_{x\tau}$ definiert.

   (a) **Bestimmung von Rel(X) durch parallele Tests:** Es seien X und X' die Testwerte zweier Tests, die beide das gleiche Merkmal erfassen: $X = \tau_x + \varepsilon_x$, $X' = \tau_{x'} + \varepsilon_{x'}$. Die beiden Tests heißen *parallel*, wenn (i) $\tau_x = \tau_{x'}$ und $\sigma^2_\varepsilon = \sigma^2_{\varepsilon'}$. Dann folgt

$$\rho_{xx'} = \rho^2_{x\tau}, \tag{3}$$

   d.h. die Korrelation zwischen zwei parallelen Tests ist gleich $\rho^2_{x\tau}$. Diese Beziehung ist der Grund, warum das Quadrat der Korrelation $\rho_{x\tau}$ zur Definition der Reliabilität benützt wird: historisch hat man mit der intuitiven Idee, dass die Korrelation zwischen parallelen Tests die Reliabilität widerspiegeln müßte, begonnen, und dann hat sich erwiesen, dass $\rho_{xx'}$ gerade dem Verhältnis $\sigma^2_\tau/\sigma^2_x = \rho^2_{x\tau}$ entspricht. Die Definition (2) erweist sich wiederum als nützlich, wenn man die Beziehung zwischen Reliabilität und Validität bzw. Trennschärfe von Items diskutieren will.

   (b) **Kongenerische Tests:** Die Forderung nach Parallelität ist restriktiv, weil es of schwierig ist, parallele Tests zu konstruieren (man denke an X: Fragebogen, X' Hormonkonzentration – die Einheiten sind verschieden und es kann eine additive Konstante ungleich Null existieren. Zwei Tests X und X' heißen demnach *kongenerisch*, wenn für die Wahren Werte die Beziehung

$$\tau_j = \nu_j + \lambda_j\eta, \quad j = 1, 2 \tag{4}$$

   gilt, mit $\tau_1 = \mathbb{E}(X)$, $\tau_2 = \mathbb{E}(X')$. Die wahren Werte sind durch eine lineare Transformation aufeinander bezogen. Gilt $\nu_j = 0$ und $\lambda_j = 1$ für $j = 1, 2$, so heißen die Tests *τ-äquivalent*. Gilt $\nu_j \neq 0$ und $\lambda_j = 1$, so heißen die Tests *essentiell τ-äquivalent*. Die wahren Werte unterscheiden sich dann nur durch eine additive Konstante. τ-äquivalente und essentiell τ-äquivalente Tests sind Spezialfälle kongenerischer Tests. Es läßt sich zeigen, dass die Reliabilität kongenerischer Tests durch

$$\rho^2_{x\tau_j} = \frac{\lambda^2_j}{\lambda_j + \sigma^2_{\varepsilon_j}} \tag{5}$$

   gegeben ist.

(c) **Reliabilität und Testlänge:** Man kann die Reliabilität eines Tests bei Verdoppelung, Verdreifachung etc seiner Testlänge herleiten. Insbesondere ist ein einzelnes Item schon ein Test, – eben der Länge 1. Der Score eines solchen Tests sei durch Y gegeben, und das Item habe die Reliabilität $\rho_{yy'}$. Ein Test bestehe nun aus n Items mit der gleichen Reliabilität. Der Test hat dann die Reliabilität

$$\rho_{xx'} = \frac{n\rho_{yy'}}{1+(n-1)\rho_{yy'}}, \tag{6}$$

(*Spearman-Brown-Formel für die Reliabilität eines Tests nfacher Länge*); die Formel erlaubt die Abschätzung der Testlänge bei angenommener Itemreleabilität. Für $n = 2$ erhält man den Fall der Reliabilität eines Tests mit doppelter Länge (wichtig bei der Berechnung der Reliabilität auf der Basis der Split-Half-Methode).

(d) **Cronbachs** α ist eine Abschätzung des minimalen Wertes der Reliabilität und gilt als *Maß für die interne Konsistenz* eines Tests: mit $X = Y_1 + \cdots + Y_n$ (Y_j die Scores für die einzelnen Items, $Y_j = \tau_j + \varepsilon_j$) gilt

$$\rho_{xx'} \geq \alpha = \frac{n}{n-1}\left(1 - \frac{\sum_j \mathbb{V}(Y_j)}{\mathbb{V}(X)}\right) \tag{7}$$

$$= \frac{n}{n-1}\left(\frac{\sum_{i\neq j}\rho_{ij}\sigma_i\sigma_j}{\sum_j \sigma_j^2 + \sum_{i\neq j}\rho_{ij}\sigma_i\sigma_j}\right) \tag{8}$$

Dass α ein Maß für die interne Konsistenz ergibt sich aus der Tatsache, dass α von den Iteminterkorrelationen ρ_{ij} abhängt: gilt $\rho_{ij} = 0$ für alle $i \neq j$, wird $\alpha = 0$. α strebt stets – also auch für kleine Iteminterkorrelationen – gegen 1 für größer werdenden Wert von n und ist insofern kein gutes Maß für die interne Konsistenz. Auch für große Korrelationen mit allerdings alternierenden Vorzeichen kann α klein werden.

(e) **Schätzfehler:** Der wahre Wert τ soll aufgrund des X-Wertes "vorhergesagt" werden; dabei werden die unvermeidlichen "Vorhersagefehler" begangen:

$$\sigma_\epsilon^2 = \sigma_\tau\sqrt{1-\rho_{x\tau}^2}, \quad \text{(Standardschätzfehler)} \tag{9}$$

$$\sigma_\varepsilon^2 = \sigma_x^2\sqrt{1-\rho_{xx'}^2}, \quad \text{(Standardvorhersagefehler)} \tag{10}$$

3. **Itemparameter:**

(a) **Itemfunktion:**[1] Die Itemfunktion (*item characteristic function*) gibt die Wahrscheinlichkeit an, mit der ein Item I_g in Abhängigkeit von der Ausprägung θ des gemessenene Merkmals $\mathcal{M}$ beantwortet bzw.gelöst wird; in der KTT wird insbesondere angenommen, dass diese Wahrscheinlichkeit durch die Gauß-Funktion gegeben ist:

$$F_g(\theta) = \Phi[a_g(\theta - b_g)], \tag{11}$$

[1] Herzlichen Dank an Frau Birte Behnken, die mich auf die Verkorkstheit der ursprünglich an dieser Stelle gegebenen Definition der Itemfunktion hingewiesen hat. Ich hoffe, der folgende Text ist verständlicher!

wobei Φ die kumulative (Verteilungs-)Funktion der Standardnormalverteilung ist, d.h.

$$\Phi(z) = \int_{-\infty}^{z} f(u)\, du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du. \tag{12}$$

(Lord & Novick (1968), p. 366). Man kann, wie Lord & Novick es tun, die Parameter a_g und b_g in (11) abstrakt als Skalenparameter definieren. Andererseits läßt sich die Beziehung zwischen (11) und (12) durch Bezug auf *unterliegende Variable* herleiten: X sei eine normalverteilte zufällige Veränderliche mit dem Erwartungswert μ und der Varianz σ^2. Dann ist bekanntlich

$$P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du.$$

$P(X \leq x)$ läßt sich über die Standardnormalverteilung Φ ausdrücken. Dazu sei $Z = (X - \mu)/\sigma$. Es ist

$$P(X \leq x) = P((X - \mu)/\sigma \leq (x - \mu)/\sigma) = P(Z \leq (x - \mu)/\sigma)),$$

d.h.

$$P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-u^2/2} du = \Phi[(x - \mu)/\sigma].$$

Die zufällige Veränderliche X werde nun mit dem als in der Zeit fluktuierenden Merkmal, das mit dem Item gemessen werden soll, identifiziert; für eine gegebene Person sei $\mu = \theta$, d.h. die Person habe im Mittel die Merkmalsausprägung θ, und die Varianz für diese Person sei durch σ_g^2 gegeben, d.h. die Varianz hänge nur von dem Item I_g ab; dies ist eine vereinfachende, wenn auch nicht notwendig sehr plausible Annahme. Weiter werde angenommen, dass das Merkmal mit einer Mindestausprägung κ_g vorhanden sein muß, damit I_g beantwortet werden kann. Die Wahrscheinlichkeit, dass das Item I_g *nicht* beantwortet wird, ist dann

$$P(X \leq \kappa_g) = \Phi[(\kappa_g - \theta)/\sigma_g].$$

Man beachte, dass $P(X \leq \kappa_g) = P(X \leq \kappa_g|\theta)$ hier eine Funktion von κ_g mit θ als festem Parameter ist, da ja $P(X \leq \kappa_g)$ die *Verteilungsfunktion* von X ist. Die Wahrscheinlichkeit, dass das Item beantwortet wird, ist dann

$$P(X > \kappa_g|\theta) = 1 - P(X \leq \kappa_g|\theta) = \Phi[(\theta - \kappa_g)/\sigma_g]. \tag{13}$$

Setzt man hierin $1/\sigma_g = a_g$ und $\kappa_g = b_g$, so erhält man den Ausdruck (11), der aber *als Funktion von θ eingeführt wurde*. Der Punkt ist, dass man die Verteilungsfunktion (13) auch als Funktion von θ mit κ_g und σ_g als fixen Parametern auffassen kann, – daher der Ausdruck *Gauß-Funktion*. Diese Uminterpretation unterliegt der Definition (11). Die Itemfunktion oder *item characteristic function* wird hier also über eine Verteilungsfunktion eingeführt. Eine ausführlichere Diskussion dieser Interpretation findet man im Skriptum "Einführung in die Theorie psychometrischer Tests", Abschnitt 2.6.3, wo X in η umbenannt wird und

wo ebenfalls die Beziehung zwischen $\Phi[(\kappa_g - \theta)/\sigma_g]$ und $\Phi[(\theta - \kappa_g)/\sigma_g]$ hergeleitet wird.

Das Interessante an der Herleitung von (13) ist, dass die Parameter a_g und b_g, die in der Lord & Novickschen Definition der Itemfunktion auftauchen, eine Interpretation in Bezug auf die Schwierigkeit eines Items zulassen. Denn $b_g = \kappa_g$ bedeutet ja, dass b_g nun die Mindestausprägung des Merkmals repräsentiert, die notwendig ist, um das Item I_g zu beantworten; ist das Merkmal weniger als κ_g ausgeprägt, beantwortet man das Item nicht. κ_g und also b_g repräsentiert demnach die Schwierigkeit des Items. Bei dieser Interpretation ist die Schwierigkeit des Items also ein Parameter, der in die Definition der Itemfunktion eingeht: je größer die Schwierigkeit, d.h. je größer der Wert von κ_g, desto weiter "rechts" auf der θ-Skala liegt die Itemfunktion. Die Itemfunktion ist um so steiler, je kleiner σ_g, je weniger also die Merkmalsausprägung für das g-te Item zufällig variieren. Die vereinfachende Annahme, dass diese Variation nur vom Item, nicht aber aber von der Person abhängen soll, ist allerdings diskussionsbedürftig. In Texten zur KTT werden diese Zusammenhänge i.A. nicht diskutiert.

(b) **Schwierigkeit:** Die Schwierigkeit eines Items I_g ist der Anteil π_g der Probanden in einer Population, die das Item lösen bzw. positiv beantworten. Man kann herleiten:

$$\pi_g = \frac{\kappa_g}{\sqrt{1 + \sigma_g^2}}. \tag{14}$$

Man kann also die Schwierigkeit π_g, so wie sie in der KTT definiert wird, auf die Schwierigkeit κ_g – definiert als kritische Merkmalsausprägung für ein Item – aufeinander beziehen (vergl. Skriptum "Einführung in die Theorie psychometrischer Tests").

(c) **Trennschärfe** ist eine Maßzahl, die angibt, wie gut ein Item zwischen Personen mit unterschiedlichen Merkmalsausprägungen trennt. Trennschärfe ist formal definiert als Korrelation zwischen der Beantwortung eines Items (Score Y_g) und dem Gesamtscore X:

$$\rho_{gX} = \frac{\mathbb{K}(Y_g, X)}{\sigma_g \sigma_x} = \frac{1}{\sigma_x} \sum_{h=1}^{n} \sigma_h \rho_{gh}, \tag{15}$$

(bei dichotomen Items: punkt-biseriale Korrelation). σ_h ist die Streuung für das h-te Item, $h = 1, \ldots, n$, und ρ_{gh} ist die Korrelation zwischen dem g-ten und dem h-ten Item, σ_x ist die Streuung der X-Werte. Niedrige Itemiternkorrelationen ρ_{gh} des Items I_g mit den übrigen reduzieren offenbar die Trennschärfe eines Items.

(Bei konkreter Berechnung: part-whole-Korrektur nicht vergessen!)

(d) **Itemreliabilität** ist die Reliabilität eines Items (Spezialfall der Reliabilität eines Tests der Länge 1).

(e) **Itemvalidität** ist die Korrelation zwischen der Beantwortung eines Items (Score Y_g) und der Merkmalsausprägung η:

$$\rho_{y_g \eta} = \frac{\mathbb{K}(Y_g, \eta)}{\sigma_g \sigma_\eta}. \tag{16}$$

4. **Validität:** Die Gültigkeit oder Validität eines Tests wird durch die Korrelation zwischen dem Gesamtscore X und der Merkmalsausprägung definiert:

$$\rho_{x\eta} = \frac{\mathbb{K}(X, \eta)}{\sigma_x \sigma_\eta}. \tag{17}$$

Beziehung zu Itemparametern: Man kann zeigen:

$$\rho_{x\eta} = \frac{\sum_g \sigma_g \rho_{g\eta}}{\sum_g \sigma_g \rho_{gX}} = \frac{\sum_g \sigma_g \rho_{g\eta}}{\sqrt{\sum_g \sum_h \sigma_g \sigma_h \rho_{gh}}}, \tag{18}$$

wobei $\rho_{g\eta}$ die Itemgültigkeiten, ρ_{gX} die Itemtrennschärfen und ρ_{gh} die Iteminterkorrelationen sind.

Für gegebene Itemgültigkeiten erhöhen niedrige Trennschärfen ρ_{gX} bzw. niedrige Iteminterkorrelationen ρ_{gh} die Validität, vergl. auch (15). Niedrige Trennschärfen weisen zunächst auf große σ_g-Werte hin. Niedrige Iteminterkorrelationen legen nahe, dass das Merkmal mehrdimensional ist und die Items die Mehrdimensionalität erfassen. Niedrige Trennschärfen können dann anscheinend auch mit niedrigen Interkorrelationen zusammenhängen: die Items erfassen spezielle Aspekte des Merkmals, von denen schlecht auf den Gesamtscore geschlossen werden kann.

5. **Beziehung zwischen der Reliabilität und der Validität:** Man beginnt mit der Gültigkeit eines Tests mit Messwerten X in Bezug auf einen anderen Tests für das gleiche Merkmal mit den Messwerten Y. Diese Gültigkeit ist gegeben durch

$$\rho_{xy} = \frac{\mathbb{K}(\tau_x, \tau_y)}{\sigma_x \sigma_y}, \tag{19}$$

und für den Fall, dass X und Y streng parallele Messungen repräsentieren ($\tau_x = \tau_y$, $\sigma_x = \sigma_y$) folgt

$$\rho_{xy} = \frac{\mathbb{V}(\tau)}{\mathbb{V}(X)} = \rho^2_{xx'} = \mathrm{Rel}(X), \tag{20}$$

d.h. die Gültigkeit ist gleich der Reliabilität des Tests. Gilt die Parallelität nicht streng (der Normalfall), folgt

$$\rho_{xy} < \sqrt{\rho_{xx'}}, \tag{21}$$

so dass die Wurzel aus der Reliablität eine obere Grenze für die Validität ist.

$$\rho(\tau_x, \tau_y) = \frac{\rho_{xy}}{\sqrt{\rho_{xx'} \rho_{yy'}}}, \tag{22}$$

d.h. die Korrelation zwischen den wahren Werten ist gleich der Korrelation ρ_{xy}, geteilt durch das geometrische Mittel der Reliabilitäten der X und Y-Werte. Allgemein gilt

$$\rho(\tau_x, \tau_y) \geq \rho_{xy}. \tag{23}$$

Paradoxe Effekte: Eine untere Grenze für die Reliabilität ist durch Cronbachs α gegeben, vergl. (7). α wird groß, wenn der Quotient $\sum_j \mathbb{V}(Y_j)/\mathbb{V}(X)$

klein wird, – dann wird $1 - \sum_j \mathbb{V}(Y_j)/\mathbb{V}(X)$ groß. Der Quotient wird klein, wenn $\sigma_x^2 = \mathbb{V}(X)$ groß wird; man erhöht die Reliabilität also, indem man die Items so auswählt, dass die Varianz der Scores X groß wird. Direkt läßt sich dies auch aus

$$\rho_{x\tau}^2 = \frac{\mathbb{V}(\tau)}{\mathbb{V}(X)} = 1 - \frac{\mathbb{V}(\varepsilon)}{\mathbb{V}(X)}$$

ersehen: $\rho_{x\tau}^2$ wird groß, wenn $\mathbb{V}(\varepsilon)/\mathbb{V}(X)$ klein wird, und dies ist der Fall, wenn $\mathbb{V}(X)$ groß relativ zu $\mathbb{V}(\varepsilon)$ wird, was wiederum bedeutet, dass die Varianz der τ-Werte groß werden muß relativ zur Varianz $\mathbb{V}(\varepsilon) = \sigma_\varepsilon^2$ der Fehler. Die Items sollte so gewählt werden, dass die systematischen Unterschiede zwischen den Probanden, also zwischen den τ-Werten, möglichst gut herauskommen relativ zum Fehler. Wegen

$$\sigma_x = \sum_g \sigma_g \rho_{gX}$$

(vergl. den Abschnitt über die Trennschärfe im Skriptum) wird $\sigma_x^2 = \mathbb{V}(X)$ groß, wenn die Trennschärfen der Items groß werden, – genau dann wird ja zwischen den Probanden hinsichtlich ihrer τ-Werte getrennt. Der Gleichung (18) entnimmt man aber, dass $\sum_g \sigma_g \rho_{gX}$ gerade umgekehrt proportional zur Gültigkeit $\rho_{x\eta}$ ist! Wenn man möchte, kann man hier mit Lord & Novick (Seite 332) von einem Reliabilitäts-Validitäts-Paradoxon sprechen. Die Beziehung (18) sagt auch, wie oben schon erwähnt, dass niedrige Iteminterkorrelationen die Validität erhöhen. Dazu kann es sogar von Vorteil sein, dass die Items negativ miteinander korrelieren. Andererseits sollen die Itemgültigkeiten groß sein, um eine hohe Gültigkeit zu erhalten. Ein großer Wert von n, der Anzahl von Items, ist ebenfalls gut für die Validität (vergl. Abschnitt 3.4.5 des Skriptums). Praktisch ist es aber schwierig, eine große Zahl von Items zu finden, die einerseits gültig sind, andererseits aber möglichst gering miteinander korrelieren. Es gibt also eine widersprüchliche Erfordernisse an die Qualitäten Reliabilität und Validität eines Tests.

6. **Homogenität versus Heterogenität:** Ein Item ist *homogen*, wenn nur ein Merkmal gemessen wird, andernfalls ist er *heterogen*. Ein Merkmal kann selbst mehrdimensional sein (Beispiel: Intelligenz) und ein Test dieses Merkmals kann in dem Sinne heterogen sein, als er die verschiedenen Dimensionen, die ja ebenfalls Merkmale repräsentieren, mißt.

   Ein wichtiger Spezialfall ergibt sich, wenn die Items dichotom sind. Die Korrelationen ρ_{gh} zwischen den Items sind dann, wenn der gewöhnliche Produkt-Moment-Korrelationskoeffizient berechnet wird, Vierfelderkoeffizienten ϕ_{gh}. Sind die Streuungen σ_g der Items unterschiedlich, ergeben sich verschiedene Randverteilungen bei den entsprechenden Vierfeldertafeln, was einen eingeschränkten Variationsbereich der ϕ_{gh}-Koeffizienten impliziert,

   $$|\phi_{gh}| \leq |\phi_{gh}^{\max}| < 1. \tag{24}$$

   Bei der Faktorenanalyse solcher Koeffizienten können dann mehr Faktoren resultieren, als zur "Erklärung" tatsächlich notwendig sind, weil man "Schwierigkeitsfaktoren" erhält, d.h. Faktoren, die eben nur die unterschiedlichen Varianzen $\sigma_g^2 = n\pi_g(1-\pi_g)$ und damit die unterschiedlichen Schwierigkeiten π_g reflektieren.

Fordert man deshalb, dass alle Items die gleichen Varianzen σ_g haben, so fordert man damit, dass sie alle die gleiche Schwierigkeit π_g haben. Die Items sind dann alle durch die gleiche Itemfunktion definiert (gleiche Steigung, gleiche Position auf der θ-Skala). Damit zwischen Probanden mit unterschiedlichen Merkmalsausprägungen differenziert werden kann, muß dann der Bereich, für den

$$0 < F(\theta|\kappa_g) < 1$$

gilt, den Bereich, in dem die wahren Werte der Probanden liegen, abdecken. Da die κ_g-Werte die π_g-Werte bestimmen (vergl. (14)), müssen alle κ_g-Werte gleich sein.

7. **Schätzung des wahren Wertes** auf der Basis eines Summenscores.

   In der KTT ist der "wahre Wert" der Erwartungswert eines Scores. In der Statistik wird der Erwartungswert – inhaltlich das arithmetische Mittel über *alle* möglichen Messungen – durch die Formel

   $$\mathbb{E}(X) = \sum_i x_i p_i$$

   definiert, wobei hier der Einfachheit halber diskrete zufällige Veränderliche angenommen werden, also Variablen, deren mögliche Werte man durchnummerieren kann: $x_1, x_2, \ldots$. Mit p_i wird die Wahrscheinlichkeit, dass X den Wert x_i annimmt, bezeichnet.

   Vom Score X wird angenommen, dass er in der Form $X = \tau + \varepsilon$ repräsentiert werden kann (vergl. (1)). Dies ist eine Modellannahme, da der Fehler ε als additiv postuliert wird; außerdem soll ε unabhängig von τ sein, so dass $\mathbb{K}(\tau, \varepsilon) = 0$. Diese Annahme wird auch im Allgemeinen Linearen Modell gemacht. Dies ist das Modell, dass der Regressionsanalyse und damit auch der Varianzanalyse zugrunde liegt. Es ist einfach, muß deswegen aber keineswegs auch psychologisch korrekt sein:

   (a) Angenommen, die Probanden sollen eine möglichst große Zahl von gleichschwierigen Aufgaben in einer vorgegebenen Zeit lösen (etwa: Arithmetiktest, $15 \times 13 =?$, $17 \times 12 =?$, etc). Werden die Aufgaben stochastisch unabhängig voneinander und jede mit gleicher Wahrscheinlichkeit korrekt gelöst, so ist die Anzahl korrekter Lösungen gleich X. Bei gegebener Fähigkeit eines Probanden ist X aber zufällig verteilt, wahrscheinlich binomialverteilt oder, in guter Näherung, Poisson-verteilt, und bei dieser (und vielen anderen) Verteilung sind Erwartungswert und Varianz gekoppelt!

   (b) Der Messwert X kann auch eine Reaktionszeit sein. Gerade bei einfachen Reaktionszeitaufgaben zeigt sich, dass X *nicht* normalverteilt ist, sondern durch Verteilungen beschrieben wird, bei denen der Erwartungswert und die Varianz ebenfalls gekoppelt sind (Exponentialverteilung, oder Gamma-Verteilung, die sich als Summe unabhängiger Exponentialverteilungen ergibt).

   Bei den genannten Beispielen ist es im Prinzip gar nicht sinnvoll, eine additive Zerlegung (X ist Summe eines wahren Wertes τ und eines Fehlers ε) anzunehmen, – auch wenn, wie bei der oft vorgenommenen Varianzanalyse

von Reaktionszeiten, eine solche Zerlegung in erster Näherung funktionieren kann.

Wenn das Modell $X = \tau + \varepsilon$ angemessen ist, ist die inhaltliche Bedeutung des Parameters τ gleichwohl unklar, da er von den Schwierigkeiten κ_g der Items abhängt, – vergl. Abschnitt 2.

2 Schätzung des wahren Wertes

2.1 Direkte Schätzung

In der KTT wird $X = \tau + \varepsilon$ angenommen, wobei für eine Person $a \in \mathcal{P}$

$$X_a = \sum_{g=1}^{n} Y_{ag}, \quad Y_{ag} = \{0, 1\}. \tag{25}$$

Formal gilt nun

$$\mathbb{E}(X_a) = \tau_a = \sum_{g=1}^{n} \mathbb{E}(Y_{ag}). \tag{26}$$

Aber $\mathbb{E}(Y_{ag}) = 1 \cdot p_{ag} + 0 \cdot (1 - p_{ag}) = p_{ag}$, so dass

$$\tau_a = \sum_{g=1}^{n} p_{ag}. \tag{27}$$

Man könnte nun τ_a schätzen, hätte man Schätzungen für p_{ag}. Die Wahrscheinlichkeiten p_{ag} sind aber Parameter, die für die Person a und für das Item I_g spezifisch sind. Deswegen kann man sie nicht über relative Häufigkeiten schätzen, denn die Person a reagiert nur einmal auf das Item I_g. Nun ist das arithmetische Mittel eine Schätzung für den Erwartungswert, so dass

$$\bar{y}_a = \frac{1}{n} \sum_{g=1}^{n} Y_{ag} = \frac{n_a}{n}, \tag{28}$$

n_a die von der Person a beantworteten oder gelösten Items. Definiert man das arithmetische Mittel der p_{ag}, also $\bar{p}_a = \sum_g p_{ag}/n$, so hat man $n\bar{p}_a = \sum_g p_{ag}$, so erhält man wegen (27)

$$\hat{\tau}_a = \bar{y}_a = n\bar{p}_a. \tag{29}$$

An dieser Gleichung wird die Problematik des Begriffs des wahren Werts einer Person, wie er in der KTT eingeführt wird, deutlich. Denn p_{ag} ist ja durch die Itemfunktion definiert: $p_{ag} = F(\theta_a|\kappa_g)$, und in $\bar{p}_a$ gehen die verschiedenen κ_g ein, d.h. τ_a und dementsprechend auch $\hat{\tau}_a$ hängen von den Items I_g ab. Man kann von einem wahren Wert der Person nur in Bezug auf eine bestimmte Menge von Items sprechen. Daraus folgt, dass τ_a nicht als Transformation von θ_a, unabhängig von den κ_g-Werten, aufgefasst werden kann.

Man kann nun den Fall betrachten, dass die Items alle durch den gleichen Schwierigkeitsparameter $\kappa_g = \kappa$ charakterisiert sind. Dieser Fall ist ja wünschenswert, wenn die Iteminterkorrelationen in einer Faktorenanalyse auf eine mögliche

Mehrdimensionalität des gemessenen Merkmals untersucht werden soll. Nimmt man das Ogiven-Modell an, so soll gelten

$$p_{ag} = \Phi\left[\frac{\theta_a - \kappa}{\sigma}\right]. \tag{30}$$

Dann liefert die inverse Transformation

$$\Phi^{-1}(p_{ag}) = \frac{\theta_a - \kappa}{\sigma} = \alpha\theta_a + \beta, \tag{31}$$

mit $\alpha = 1/\sigma$, $\beta = -\kappa/\sigma$. Damit hat man eine Schätzung von θ_a, die eindeutig bis auf eine lineare Transformation ist. Diese ist natürlich wiederum abhängig von der Itemschwierigkeit. Da diese aber für alle Personen die gleiche ist, bekommt man auf jeden Fall eine Ordnung der Personen, d.h. man kann die Personen relativ zueinander in Beziehung hinsichtlich der θ-Skala setzen.

2.2 Maximum-Likelihood-Schätzungen

Im Falle unterschiedlicher Schwierigkeiten ist es kaum möglich, die einzelnen Wahrscheinlichkeiten p_{ag} zu schätzen, – hätte man Schätzungen, so könnte man die inverse Transformation

$$\Phi^{-1}(p_{ag}) = (\theta_a - \kappa_g)/\sigma_g$$

bestimmen und die θ_a ausrechnen, vorausgesetzt, man hätte aus vorangegangenen Messungen Schätzungen für κ_g und σ_g.

Ein Ausweg ergibt sich, wenn man bedenkt, dass für fixen Wert von θ die Antworten stochastisch unabhängig sind (lokale stochastische Unabhängigkeit). Für die Person a seien nun die Itemscores $Y_1, \ldots, Y_n$ gegeben; dies ist eine Folge von Nullen und Einsen. Die Wahrscheinlichkeit für diese Folge ist

$$P(Y_1, \ldots, Y_n) = \prod_{g=1}^{n} p_{ag}^{y_{ag}}(1 - p_{ag})^{1-y_{ag}}.$$

Für $Y_{ag} = 1$ ist $p_{ag}^{y_{ag}}(1 - p_{ag})^{1-y_{ag}} = p_{ag}$, und für $Y_{ag} = 0$ ist $p_{ag}^{y_{ag}}(1 - p_{ag})^{1-y_{ag}} = 1 - p_{ag}$. Macht man darüber hinaus die plausible Annahme, dass die einzelnen Probanden einer Stichprobe stochastisch unabhängig voneinander die Aufgaben bzw. Items bearbeiten, so erhält man für die Gesamtheit der Folgen $Y_{a1}, \ldots, Y_{an}$, $a = 1, \ldots, m$ die Wahrscheinlichkeit

$$L = \prod_{a=1}^{m} \prod_{g=1}^{n} p_{ag}^{y_{ag}}(1 - p_{ag})^{1-y_{ag}}. \tag{32}$$

L ist die *Likelihood* der Daten (vergl. Abschnitt 4.3 im Skriptum; die Tabelle D zeigt beispielhaft, wie die Antwortmuster der einzelnen Personen zusammengefasst werden). Setzt man für p_{ag} nun die entsprechenden Ausdrücke $\Phi[(\theta_a - \kappa_g)/\sigma_g]$ ein, so sieht man, das L eine Funktion der unbekannten Parameter θ_a, κ_g und σ_g ist. Diese werden nun so bestimmt, dass L ein Maximum annimmt, d.h. man bestimmt die Maximum-Likelihood-Schätzungen für diese Parameter. Statt des Ogive-Modells kann auch das logistische Modell gewählt werden. Für dieses Modell erweisen sich die θ_a- und κ_g-Schätzungen als unabhängig voneinander, und dieser Sachverhalt (spezifische Objektivität) macht den Reiz des logistischen Modells aus. Man wird damit zu der Frage geführt, welches Modell denn adäquat ist. Die Frage wird in der Fortsetzung des Kurses diskutiert.

3 Aufgaben

1. Was versteht man unter einer latenten Variablen?

   Fähigkeit bzw Ausprägung des gemessenen Merkmals; bestimmt die Wahrscheinlichkeit der Beantwortung eines Items. Aber auch: Faktorielle Dimensionen - mehrdimensionale Struktur des gemessenen Merkmals (Beispiel: Intelligenz und Thurstones 7 primary mental abilities).

2. Welche Begründung würden Sie für die Wahl der Gauss-Funktion als Itemfunktion geben?

   Der übliche Verdächtige: Zentraler Grenzwertsatz. Der muß aber gar nicht gelten (Summe unabhängiger zufälliger Veränderlicher, etc)

3. Welche Beziehung besteht zwischen der Verteilungsfunktion der *unterliegenden Variablen* des als zufällig fluktuierend angenommenen Merkmals η und der Itemfunktion?

   Verteilungsfunktion: $P(\eta \leq \kappa_g|\theta)$, $\theta = \mathbb{E}(\eta)$ als Funktion von κ_g (entspricht dem x in $P(X \leq x)$). Dies ist die Wahrscheinlichkeit, dass das Item nicht beantwortet bzw. gelöst wird. Dann ist $P(\eta > \kappa_g|\theta)$ die Wahrscheinlichkeit, dass es beantwortet bzw gelöst wird, – gegeben den Personenparameter θ. $F(\theta) = 1-P(\eta > \kappa_g|\theta)$ *als Funktion von* θ, nicht von κ_g, ist die Itemfunktion.

4. Axiome der Klass. Testtheorie (KTT). Welchen Algemeinheitsgrad können sie beanspruchen?

   Die Axiome entsprechen in gewisser Weise den Annahmen, die üblicherweise bei Regressions- und damit auch bei Varianzanalysen gemacht werden: man nimmt "deterministische" Effekte der unabhängigen Variablen – hier τ – an, auf die "Fehler" ε additiv und unabhängig einwirken. In der KTT ist die Hauptannahme, dass der Score $X = \tau + \varepsilon$ ist. Die Annahme $\mathbb{E}(X) = \tau$ und damit $\mathbb{E}(\varepsilon) = 0$ ist stets möglich, und damit ist die Annahme eines additiven Fehlers stets möglich, – er ist einfach die Abweichung des Messwerts vom "wahren Wert" τ. Die Annahme, dass $\mathbb{V}(\varepsilon)$ unabhängig von τ und damit für alle τ-Werte gleich groß ist, ist dagegen nich trivial, – sie muß keineswegs korrekt sein. Ein einfaches Beispiel ist durch den Fall gegeben, dass eine Person die Items unabhängig voneinander beantwortet und dass die Items alle gleich schwierig sind (vereinfachende Annahmen). Die Anzahl der Items, die korrekt beantwortet werden, ist dann binomial-verteilt mit Erwartungswert $\mathbb{E}(X) = np$ und Varianz $np(1-p)$ – Erwartungswert und Varianz sind also nicht unabhängig voneinander. Deshalb ist die Annahme der Normalverteilung wichtig: sie ist eine der wenigen Verteilungen, bei denen σ^2 unabhängig von $\mathbb{E}(X) = \tau$ gewählt werden kann. Aber selbst im Falle normalverteilter X-Werte muß σ^2 nicht notwendig unabhängig von τ sein; ob diese Unabhängigkeit zutrifft oder nicht, ist eine empirische und keine theoretische Frage.

5. Was versteht man in der KTT unter dem wahren Wert eines Probanden? Was bedeutet es, wenn zwei Tests, die das gleiche Merkmal messen, verschiedene wahre Werte ergeben?

   Der wahre Wert ist einfach der Erwartungswert, dh der Mittelwert über alle möglichen Werte von X, für einen Probanden. Das Wort "wahr" hat in diesem

Zusammenhang also eine eingeschränkte Bedeutung; der wahre Wert hängt ja von der Anzahl der Items, der Art des Scoring etc ab. Deshalb ist der Begriff des kongenerischen Tests wichtig: hier ist der "wahre" Wert durch $\tau = \nu + \lambda\eta$ gegeben. η sind Repräsentationen der unterliegenden gemessenen Variablen, die aber nicht direkt, sondern eben nur vermittels eines Tests gemessen werden können. Es wird aber angenommen, dass τ eine lineare Transformation von η oder θ ist, wobei die Parameter der Transformation, also ν und λ, implizit durch die Wahl des Tests bestimmt werden (Ratings, Blutdruckwerte, Hormonkonzentrationen, Zeitangaben in Sekunden oder Millisekunden, etc). Eine Skala für η oder θ wird nicht explizit angegeben, es wird aber angenommen, dass diese Variable existiert und systematisch auf die Testleistung einwirkt.

6. Was versteht man unter der Reliabilität eines Tests?

   Reliabilität meint Meßgenauigkeit ist durch die Korrelation zwischen X- und τ-Werten gegeben. Die Korrelation ist hoch, wenn die Varianz der τ-Werte groß ist im Vergleich zur Varianz der ε-Werte. Diese Festlegung der Begriffs der Zuverlässigkeit ergibt sich aus der Annahme $X = \tau + \varepsilon$, die ja formal eine Regressionsgleichung ist. Daraus folgt $\rho^2_{x\tau} = \mathbb{V}(\tau)/\mathbb{V}(X) = \rho_{xx'}$, wobei die Größe $\rho_{xx'}$ nur im Falle paralleler Tests mit den Scores X und X' eine Schätzung für die Reliabilität liefert.

7. Welche Beziehung zwischen der Parallelität von Tests und den zugehörigen Itemfunktionen würden Sie vermuten?

   Parallelität bedeutet, dass $\tau_x = \tau_{x'}$ und $\sigma_x = \sigma_{x'}$. Handelt es sich um einzelen Items, folgt, dass die zugeörigen Itemfunktionen gleiche Steigungen haben, dh sie sind parallel. Dies gilt auch für essentiell τ-äquivalente Tests. Für dichotome Items ist die Varianz durch $\sigma_g = \pi_g(1 - \pi_g)$ gegeben, wobei π_g die Schwierigkeit des Items I_g ist. Gleiche Varianz heißt dann gleiche Schwierigkeit. Eine Itemfunktion hat die Form $\Phi[a_g(\theta - b_g)]$, mit $a_g b_g = \kappa_g/\sigma_g$. In Abschnitt 3.3.1 (Schwierigkeit von Items) wird die Beziehung

   $$\pi_g = \frac{\kappa_g}{\sqrt{1 + \sigma_g^2}}.$$

   angegeben. Also erhält man die Beziehung

   $$a_g b_g = \frac{\kappa_g}{\sigma_g} = \frac{\pi_g\sqrt{1 + \sigma_g^2}}{\sigma_g}$$

   gleiche π_g-Werte bedeuten gleiche σ_g-Werte bedeuten gleiche κ_g-Werte bedeuten $a_g b_g$-Werte, die wiederum die Position auf der θ-Skala bedeuten, – also haben die Itemfunktionen auch die gleiche Position.

8. Wie ist die Gültigkeit eines Tests allgemein definiert, und welche Arten von Gültigkeit kennen Sie?

   Ist η das gemessene Merkmal, so ist die Gültigkeit durch die Korrelation $\rho_{x\eta}$ definiert. Im Allgemeinen wird η durch einen bereits existierenden Test mit Messwerten Y angegeben, so dass die Gültigkeit durch die Korrelation $\rho_{xy} = \mathbb{K}(\tau_x, \tau_y)/\sigma_x\sigma_y$ geschätzt wird.

   Arten von Validität: Konstrukt- etc, jeweils erläutern.

9. Sind die Begriffe der Validität und der Reliabilität unabhängig voneinander?

   Auf der rein begrifflichen Ebene *erscheinen* sie als unabhängig: der Test ist reliabel, wenn er das Merkmal, das er mißt genau mißt, und er ist valide, wenn er das Merkmal, das er messen soll, tatsächlich auch mißt.

   Aber es der Begriff des "genauen Messens", über den sie miteinander verbunden sind, denn dabei geht es um die Messfehler, die mit jeder Messung verbunden sind. Die operationale Definition[2] der Begriffe geschieht durch Bezug auf den Korrelationsbegriff: wenn ein Test genau mißt, müssen die Messwerte mit dem, was er misst, gut übereinstimmen, d.h. die Fehler müssen gering sein. Die Charakterisierung des Begriffs der Gültigkeit ist analog. In beiden Korrelationskoeffizienten tritt die Fehlervarianz der Messungen auf. Deswegen resultiert schließlich die Aussage, dass $\rho_{xy} \leq \sqrt{\rho_{xx'}}$, dh die Gültigkeit kann nie größer als die Reliabilität sein. Die operationalen Definitionen zeigen die Beziehung zwischen den Begriffen auf.

10. Die Trennschärfe eines Items I_g ist groß, wenn die Scores Y_g hoch mit dem Gesamtscore X korrelieren. Sind diese Korrelationen nahe bei Null für alle Items, so trennen die Items nicht zwischen den Probanden. Ist die Varianz σ_x^2 der Scores nun groß, weil ja wohl zufällig geantwortet wird?

    Nein, wegen $\sigma_x = \sum_g \sigma_g \rho_{gx}$ (vergl. Abschnitt über Trennschärfe) folgt, dass σ_x klein wird für $\rho_{gx} \approx 0$. Große Trennschärfen sorgen für eine "große" Varianz der Scores X.

    Das ist plausibel: hohe Trennschärfen bedeuten ja, dass sich die Beantwortung eines Items im Gesamtscore X ausdrückt, Probanden, die ein Item nicht lösen können, haben einen dementsprechend erniedrigten Score, und umgekehrt. Unterschiedliche Merkmalsausprägungen drücken sich also in hohen Trennschärfen aus, dh wegen $\mathbb{V}(X) = \mathbb{V}(\tau) + \mathbb{V}(\varepsilon)$ ist dann $\mathbb{V}(\tau) \neq 0$, dh die Unterschiedlichkeit der τ-werte drückt sich wiederum in den in den Trennschärfen aus. Wäre dies nicht so, wäre also $\mathbb{V}(\tau) \approx 0$, so wären auch die Trennschärfen klein und $\mathbb{V}(X)$ wäre nur durch $\mathbb{V}(\varepsilon)$ bestimmt. Bei großen Trennschärfen geht auch noch $\mathbb{V}(\tau) > 0$ in $\mathbb{V}(X)$ ein.

11. Was bedeutet es, wenn ein Test entweder homogen oder heterogen ist? Welchen Einfluß kann die strenge Homogenität auf die Validität eines Tests haben?

    Viele psychische Merkmale, die durch ein Eigenschaftswort gekennzeichnet werden (zB "depressiv", "intelligent", "kreativ", etc), sind letztlich Kombinationen von Merkmalen. Ist M ein solches "komplexes" Merkmal, so ist etwa

    $$M = M_1 \oplus M_2 \oplus \cdots \oplus M_n,$$

    wobei $\oplus$ irgendeine Verknüpfung bedeutet. Hier ist angenommen worden, dass es eine feste Anzahl n von Sub-Merkmalen $M_1, \ldots, M_n$ gibt, die M in bestimmter Weise, wie sie durch die $\oplus$-Verknüpfung gegeben ist, näher spezifiziert wird. Das muß nicht so sein: wie man bei Begriffsexplikationen merkt (Was heißt es, dass ein Mensch "aggressiv" ist?), sind die M_j und die Art der

[2] Operationale Definitionen sind, salopp gesagt, Definitionen, eine Operation zum Erfassen eines Begriffs in einem empirischen Zusammenhang spezifizieren. Dieses Erfassen ist üblicherweise eine Art von Messen.

Verknüpfung $\oplus$ nicht immer explizit und aufzählbar gegeben. Bei Persönlichkeitstests hat es Jahrzehnte zum Teil kontroverser Diskussionen gedauert, bis man sich auf die "Big Five" geeinigt hat, bei der Intelligenz gibt es analoge Diskussionen. Testtheoretisch wird $\oplus$ i. A. als additive Verknüpfung

$$X = b_1 X_1 + \cdots + b_n X_n + e$$

postuliert, mit b_j als geeignet gewählten Gewichten (bestimmt durch multiple Regression bzw. Faktorenanalyse). Homogen ist ein Test dann, wenn $n = 1$ angenommen werden kann, was aber nicht ausschließt, dass X_1 ebenfalls wieder ein komplexes Merkmal sein kann, dessen Struktur aber mit den gegebenen Items nicht weiter aufgeschlüsselt wird. Für $n > 1$ ist der Test heterogen, insbesondere, wenn X_1, X_2 etc unabhängig voneinander sind, was i. A. unkorreliert heißt. Im homogenen Fall werden die Iteminterkorrelationen zwischen allen Items eher hoch sein, im heterogenen Fall wird es auch niedrige Iteminterkorrelationen geben, – was sich positiv auf die Validität des Tests auswirkt.

12. Der Ansatz $X = \tau + \varepsilon$ impliziert, dass die Messwerte X objektiv ggeben sind, dh verschiedene Tester gelangen zu den gleichen X-Werten. Was bedeutet es, wenn diese Objektivität nicht gegeben ist, verschiedene Tester also verschiedene X-Werte bestimmen?

    Formal kann man die Frage angehen, indem man eine weitere zufällige Größe τ_0 einführt, die den Effekt eines Testers repräsentiert, so dass man nun

    $$X = \tau + \tau_0 + \varepsilon$$

    hat. Dann ist der Erwartungswert von X durch

    $$\mathbb{E}(X) = \mathbb{E}(\tau) + \mathbb{E}(\tau_0)$$

    gegeben, und die Varianz[3] durch

    $$\mathbb{V}(X) = \mathbb{V}(\tau) + \mathbb{V}(\tau_0) + \mathbb{K}(\tau, \tau_0) + \mathbb{K}(\tau_0, \varepsilon) + \mathbb{V}(\varepsilon),$$

    wobei die Möglichkeit berücksichtigt wurde, dass der Tester sein Urteil von der getesteten Person in der Weise abhängig macht, dass eine Kovarianz zwischen τ-Werten und τ_0-Werten existiert (intelligente Personen werden als besonders intelligent eingestuft, weniger intelligente als besonders dumm, etc). Auch kann es zu Kovarianzen zwischen Fehlern ε und τ_0-Werten kommen. Natürlich gelten jetzt alle Formeln für die Reliabilität, die Validität, etc nicht mehr. Objektivität ist eine notwendige Voraussetzung für eine sinnvolle Anwendung der KTT.

13. Sie wollen den Effekt einer psychologischen Intervention überprüfen, indem Sie die Messwerte eines Merkmals vor und nach der Intervention betrachten. Wie gehen sie dabei vor, wenn sie davon ausgehen können, dass die Anfangswerte X_1 und die Differenzwerte $G = X_2 - X_1$ korrelieren?

    Man betrachtet die Regression von G auf die Werte X_1 und X_2, vergl. Abschnitt 3.5 des Skriptums; die Regressionsgewichte weisen eine Verwandschaft zu Partialkorrelationen auf (dies gilt übrigens generell für die Regressionskoeffizienten bei einer multiplen Regression!).

[3] Die Varianz einer Summe ist nur gleich der Summe der Varianzen, wenn alle Kovarianzen verschwinden, – was aber nicht notwendig so sein muß!

14. Im Zusammenhang mit der Interpretation der Personparameter τ bzw θ ist die Frage nach einer Interpretation dieser Parameter aufgekommen. Eine Klasse von Theoretikern argumentiert, es handele sich um personenspezifische Werte, die eben spezifisch für eine Person und konstant für sie seien, andere argumentieren, der Antwortprozess selbst sei ein zufälliger Prozess; sie sprechen von einem "stochastischen Subjekt". Wie können Sie den Begriff des stochastischen Subjekts mit dem der Itemfunktion in Zusammenhang bringen, und welche Argumente fallen Ihnen als Psychologin bzw Psychologen ein, wenn es um die messtheoretisch begründete Annahme eines festen, konstanten Messwertes θ für eine Person geht?

    Der Begriff des stochastischen Subjekts wird in Abschnitt 2.5.8 des Skriptums behandelt. In Abschnitt 2.5.3 werden die 'unterliegenden Variablen' diskutiert, über die der Bezug zu den Itemfunktionen hergestellt wird. Die von Messtheoretikern bevorzugte Theorie konstanter Werte θ ist psychologisch eher unplausibel, weil kognitive Prozesse stets einen stochastischen Anteil haben. Man kann aber θ als Erwartungswert $\mathbb{E}(\eta)$ eines zufällig variierenden Merkmals interpretieren, dann nähert man sich dem messtheoretischen Standpunkt an. Die Annahme eines konstanten θ ist natürlich nur relativ, denn systematische Veränderungen der Merkmalsausprägung kann man i. A. nicht ausschließen.

15. Hat die Annahme eines stochastischen Subjekts irgendeinen Einfluß auf die Interpretation der Testscores eines Probanden?

    Eigentlich nicht. Die Idee des stochastischen Subjekts dient eher der Einbettung der Testtheorie in die Psychologie, sofern sie sich um psychologische Prozesse bemüht. Die Messtheorie ist eher eine fundamentalistische Bewegung, die aus bestimmten Ecken der Wissenschaftstheorie kommt, deswegen aber dennoch auch in der Wissenschaftstheorie nicht unumstritten ist.

16. Die Faktorwerte sind Personenparameter, die sich bei der Faktorenanalyse von Testwerten; die Faktorladungen charakerisieren dagegen die Items. Für den standardisierten[4] Messwert z_{ag} (a-te Person, g-tes Item) gilt ja im Falle von r latenten Dimensionen (Merkmalen)

$$z_{ag} = \alpha_{g1}\theta_{a1} + \cdots + \alpha_{gr}\theta_{ar} + \varepsilon_{ag}, \tag{33}$$

    α_{gj} die Faktorladungen, θ_{aj} die Personenparameter (Faktorwerte) der a-ten Person auf den $j = 1, \ldots, r$ latenten Merkmalen.

    Vergl. hierzu Abschnitt 2.5.8 – Modelle und ihre Semantik. (Die Notation ist dort ein wenig anders, aber das macht nichts!)

17. Es sind nur monotone Itemfunktionen diskutiert worden. Damit ist die Diskussion auf eine bestimmte Klasse von Items beschränkt worden. Liefern Sie ein Beispiel (oder mehrere) für Items, nicht nicht durch eine monotne Itemfunktion charakterisiert werden können.

[4] Es müssen nicht standardisierte Scores verwendet werden; für den Fall unterschiedlicher Skalen bei verschiedenen Items haben standardisierte Skalen aber den Vorteil, dass die Kovarianzen zwischen ihnen vom Effekt verschiedener Maßeinheiten befreit sind.