.

Einführung in die Theorie psychometrischer Tests[1]

U. Mortensen

Skriptum zum Kompaktkurs Testtheorie am
Psychologischen Institut der
Johannes Gutenberg - Universität Mainz
SS 2009

Version 19. 08. 2019

[1] AVorlesungen/Testkonstruktion/TestkonstruktionNeuB.tex

Inhaltsverzeichnis

1 Möglichkeiten und Grenzen psychometrischer Tests 6

2 Grundbegriffe der Testtheorie 14

2.1 Itemparameter und Itemtypen 14
2.2 Aufgabentypen 18
2.3 Raten und Wissen 20
2.4 Testwerte, Testskalen und Normierungen 21
2.5 Testmodelle 29
2.6 Tests und latente Merkmale 31
2.6.1 Antwortfunktionen 31
2.6.2 Lokale stochastische Unabhängigkeit 38
2.6.3 Latente Variablen 41
2.6.4 *Person Reliability* und *Person characteristic curves* 47
2.6.5 Nicht-monotone Itemfunktionen 49
2.6.6 Guttman-Skalen 51
2.6.7 Thurstone-Skalen und die Positionierung von Probanden . . 56
2.6.8 Latente Merkmale und Klassifikation 61
2.6.9 Der theoretische Status latenter Variablen 65
2.6.10 Die Schätzung von Parametern 72
2.6.11 Sensitivität, Spezifität, Signalentdeckung 86
2.7 Objektivität und Urteilerübereinstimmung 96
2.7.1 Urteilerübereinstimmnung 96
2.7.2 Die Maße von Bennet et al., Scott, und Cohen 103
2.7.3 Herleitung von Cohens κ: $K = 2$ 108
2.7.4 Herleitung von Cohens κ: $K > 2$ 114
2.7.5 Beurteilungen auf Ratingskalen 115
2.7.6 Der Intraklassen-Korrelationskoeffizient 118
2.7.7 Die Hauptachsentransformation von Ratings 124
2.7.8 Korrespondenzanalyse und Urteilerübereinstimmung 127
2.7.9 Log-lineare Modelle und Urteilerübereinstimmung 133
2.7.10 Logistische Regression und die Rolle von Kovariaten 134

2.7.11 Latente Klassen und diagnostische Übereinstimmung 138

3 Klassische Testtheorie (KTT) 140

3.1 Die Annahmen der Klassischen Testtheorie 140
3.1.1 Gulliksens Axiome 140
3.1.2 Die KTT als Spezialfall einer Item-Response-Theorie 144
3.2 Reliabilität und ihre Abschätzung 147
3.2.1 Zum Begriff der Reliabilität 147
3.2.2 Kongenerische, τ-äquivalente und parallele Tests 149
3.2.3 Arten der Reliabilität 155
3.2.4 Die Spearman-Brown-Formeln für $n > 1$ Komponenten . . . 155
3.2.5 Interne Konsistenz und Cronbachs α 160
3.2.6 Gewichtung der Testwerte und die Maximierung von α . . . 168
3.2.7 Schätzung der Reliabilität 169
3.2.8 Reliabilität und Testlänge 170
3.2.9 Die Schätzung des wahren Wertes eines Probanden 174
3.3 Itemcharakteristika 177
3.3.1 Die Homogenität von Items 177
3.3.2 Die Schwierigkeit von Items 178
3.3.3 Die Trennschärfe 181
3.3.4 Itemvalidität 183
3.4 Validität 183
3.4.1 Die allgemeine Definition 183
3.4.2 Parallele und nichtparallele Messungen 185
3.4.3 Attenuierung, Reliabilität und Validität 186
3.4.4 Arten von Validitäten. 188
3.4.5 Validität bei Testverlängerung 189
3.5 Die Abschätzung von Veränderungen (Gain) 190
3.6 Die Schätzung der Item- und Testparameter 192
3.7 Gruppenheterogenität und andere Faktoren 193
3.7.1 Gruppenheterogenität und Reliabilität 193
3.7.2 Speed- und Power-Komponenten 195
3.7.3 Messbedingungen und Reliabilität 196

3.7.4 Attenuationskorrekturen . . . 198
3.8 Allgemeine Kommentare zur Klassischen Testtheorie . . . 198
3.8.1 Zur Grundannahme . . . 198
3.8.2 Dichotome Items . . . 200

4 Nichtklassische (Item Response-) Modelle 206
4.1 Das Generalisierte Lineare Modell . . . 206
4.2 Modelle auf der Basis der logistischen Funktion . . . 212
4.3 Das Rasch-Modell . . . 213
4.4 Die Schätzung der Modellparameter . . . 216
4.5 Der Test des Modells . . . 220
4.5.1 Allgemeines Vorgehen . . . 220
4.5.2 Genauigkeit der Personenparameter . . . 222
4.6 Mehrkategoriale und mehrdimensionale Verallgemeinerungen . . . 223
4.6.1 Raschs multikategoriales Modell . . . 223
4.6.2 Faktorenanalytische Ansätze . . . 228
4.6.3 Das linear-logistische Modell . . . 231
4.6.4 Tests der Eindimensionalität und Item Cluster . . . 238
4.7 Das ordinale Rasch-Modell . . . 239
4.8 Die Messung von Veränderungen . . . 242
4.9 Faktorenanalytische Modelle in der Testtheorie . . . 245
4.9.1 Überblick . . . 245
4.9.2 Das Lineare Faktorenmodell . . . 246
4.9.3 Dichotome Variable I . . . 248
4.9.4 Dichotome Variable II . . . 253
4.9.5 Faktorenanalyse und Item-Response-Theorie . . . 254
4.9.6 Nichtlineare Modelle . . . 256
4.9.7 Latent-Class-Modelle . . . 259
4.9.8 Multiple Korrespondenzanalyse . . . 262

5 Messmodelle und Prozessmodelle 271

A Anhang 281
A.1 Cohens κ für $P_r \to 0$ und $P_r \to 1$ . . . 281

A.2 Erwartungswerte, Varianzen, Kovarianzen und Korrelationen . . . 281
A.3 Äquivalenz der Definitionen von Itemfunktionen 285
A.4 Cronbachs α: Beweis von Satz 3.2.1 287

B Theorie des Messens **288**
B.1 Grundbegriffe der Messtheorie 288
B.2 Verbundmessung (Conjoint Measurement, CM) 291

Literatur **296**

Index **307**

1 Möglichkeiten und Grenzen psychometrischer Tests

Menschliches Verhalten läßt sich beobachten. Die Frage ist, wie Verhalten interpretiert werden kann. Man möchte auf Fähigkeiten, Neigungen, allgemein auf Merkmale schließen, von denen man annehmen kann, dass sie das Verhalten eines Menschen in einer allgemeinen Art bestimmen. Damit ist nicht gemeint, dass man annimmt, dass Verhalten im Detail vorhergesagt werden kann oder soll, sondern dass Klassen oder Typen von Verhaltensweisen vorhergesagt werden sollen. Man möchte etwa wissen, ob ein Mensch 'intelligent' ist, so dass ihm zugetraut werden kann, dass er bestimmte Probleme lösen kann, oder man möchte wissen, ob er bei nächster Gelegenheit wieder ein Verbrechen begeht. Man möchte wissen, ob ein Kind am ADHS[2] leidet oder ob sein Verhalten nur eine suboptimale häusliche Umgebung reflektiert, ob eine Person nur in einem Verstimmungszustand ist oder an einer Depression leidet, etc. Die Interpretation von Verhalten ist, im allgemeinsten Sinn, eine Diagnose. Diagnosen basieren auf beobachtbaren Symptomen; Verhaltensweisen sind nur ein Typ von Symptomen unter mehreren, physiologische Daten stellen einen weiteren Typ dar.

Eine Diagnose führt also von Symptomen auf Merkmale, von denen angenommen wird, dass sie sich in den Symptomen äußern. Der Rückschluß von einzelnen Symptomen auf "unterliegende" oder "latente" Merkmale ist oft nicht eindeutig: Kopfschmerzen können verschiedene Gründe haben, Schwierigkeiten mit dem Kopfrechnen müssen keinen partiellen Intelligenzdefekt bedeuten. Viele Symptome sind nicht dichotom, d.h. entweder vorhanden oder nicht vorhanden, sondern lassen sich auf einer Skala etwa von "wenig ausgeprägt" bis "stark ausgeprägt" einstufen. Dabei hängt die Bedeutung eines einzelnen Symptoms von der Gesamtmenge und Gewichtung der beobachtbaren Symptome ab. Es wird deswegen gerne gefolgert, dass die Symptome in "ganzheitlicher" ("holistischer") Weise in den diagnostischen Prozess eingehen sollen, und ebenso gerne wird daraus gefolgert, dass *Experten* besonders geeignet seien, Diagnosen zu stellen. Dabei versteht man unter Experten Menschen, die auf mindestens einem Gebiet eine Kombination von Erfahrungen und professionellem Fachwissen aufweisen derart, dass die Symptome stets in korrekter Gewichtung als Konfiguration von Symptomen gesehen werden und dass auf dieser Konfiguration beruhende Diagnosen allenfalls mit einem minimalem Fehler behaftet sind.

Andererseits legen Erfahrungen mit im eben beschriebenen Sinne als Experten auftretenden Diagnostikern nicht nur aus der Psychologie, sondern auch aus der Medizin, der Pädogagik und aus anderen Fächern nahe, dass ihre Urteile sytematische Fehler enthalten können. So haben z.B. einige Pädagogen die Neigung, die Intelligenz von Schülern abhängig von deren familiärem Hintergrund systematisch zu über- oder zu unterschätzen. Um Fehlbeurteilungen dieser Art auszuschließen könnte man *objektive Tests* mit für alle Probanden gleichen Auf-

[2] Aufmerksamkeitsdefizit-/Hyperaktivitätssyndrom

gaben und standardisierter Evaluation der Testergebnisse eine Möglichkeit sein, die Intelligenz eben objektiv, unabhängig von der "Erfahrung" des Testers zu bestimmen, obwohl auch hier der sozioökonomische Hintergrund der Getesteten eine systematisch verzerrende Rolle spielen kann. Teilresultate aus Tests, in denen insbesondere arithmetische Fähigkeiten, Raumvorstellungsvermögen, verbale Fähigkeiten etc erfaßt werden, werden in optimaler Weise gewichtet, um eben den berühmten (oder auch berüchtigten) Intelligenzquotienten zu berechnen. Der Intelligenzquotienten fasst die Leistungen in den verschiedenen Untertests eines Intelligenztests in einer einigen Zahl zusammen; aufschlußreicher kann das Intelligenzprofil eines Probanden[3] sein, das in der Angabe der Leistungen in den Untertests besteht. Man kann dann versuchen, anhand dieses Profils vorauszusagen, ob ein Proband bestimmten intellektuellen Anforderungen gerecht wird oder nicht. In analoger Weise kann man bei der Messung anderer Merkmale vorgehen. Die Gewichtung der Ergebnisse in den Untertests läßt sich durch statistische Verfahren (z.B. multiple Regression) finden. Die Evaluation von Testergebnissen wird so gewissermaßen mechanisiert; dieses Vorgehen kann gleichermaßen für die Diagnose anderer (Persönlichkeits-) Merkmale und im Prinzip auch für medizinische Diagnosen gewählt werden. Intelligenz- und Persönllichkeitstests sind Beispiele für *psychometrische Tests*. Der Ausdruck 'psychometrisch' verweist dabei auf die Annahme, dass psychische Merkmale gemessen, im allgemeinsten Sinne also quantitativ erfasst werden können. Diese Annahme ist keineswegs undiskutiert geblieben. In der Tat wird die Diskussion über die Messbarkeit, also die psychometrische Erfassung psychischer Merkmale, seit Jahrzehnten geführt. Auf der einen Seite stehen die "Experten", auch "Kliniker" genannt. Damit sind Diagnostiker gemeint, die ihre Urteile auf der Basis ihrer persönlichnen Erfahrung und oft ihres "ganzheitlichen Ansatzes" fällen, von dem sie sagen, dass er im psychometrischen Ansatz fehle. Auf der anderen Seite stehen die "Mechaniker", d.h. die Verfechter des psychometrischen Ansatzes. Ein philosophischer, oder, wenn man will, wissenschaftstheoretischer Hintergrund für die Auffassung, Experten seien am besten für das Erstellen von Diagnosen geeignet, ist die *Verstehende Psychologie*, wie sie von dem deutschen Philosophen Wilhelm Dilthey (1833 - 1911) im Rahmen seiner Versuche, die Idee einer *Lebensphilosophie* im Rahmen der Geisteswissenschaft zu etablieren, konzipiert wurde. Der 'lebendige Geist' unterliege keinen Gesetzen der Art, wie man sie etwa aus der Physik kennt. Das Verstehen eines Menschen resultiere, so Dilthey, aus einem hermeneutischen Deuten. Die Hermeneutik ist die Methode der Geisteswissenschaften schlechthin. Eine kurze Einführung in die Verstehende Psychologie und ihre Problematik findet man in Mortensen (2008)[4]. Diltheys Diktum "Die Natur erklären wir, das Seelenleben verstehen wir." (Dilthey, 1904) wird von Klinikern gern in die Debatte geworfen, um Kontrahenten in ehrfürchtiges Schweigen zu versetzen.

[3] Um den Text nicht allzu holperig werden zu lassen, wird nicht stets darauf hingewiesen, dass natürlich ebenso Expertinnen, Probandinnen etc gemeint sind.

[4] Wissenschaftstheorie III, http://www.uwe-mortensen.de/WisstheorieIII.pdf, Kapitel 4 und 5.

In der englischsprachigen Literatur war es zunächst Lundberg (1941), der den Standpunkt der 'Kliniker' im Meinungsspektrum repräsentierte. Der Gegenstandpunkt der 'Mechaniker' wurde von Sarbin (1943) vertreten, derim Rahmen einer behaviouristischen und damit (neo-)positivistischen Position argumentierte. Eine erste Zusammenfassung der Diskussion wurde 1954 von Meehl vorgelegt. Meehl führt eine Reihe von Adjektiven auf, mit denen die Kontrahenten sich und die jeweiligen Gegner belegen. So gilt die "statistische" oder "mechanische" Methode im Selbstbild ihrer Anhänger als operational, kommunizierbar, objektiv, verifizierbar, präzise, zuverlässig, wissenschaftlich, etc, während die "klinische Methode" als mystisch, transzendent, metaphysisch, unzuverlässig, grob, unverifizierbar, subjektiv, primitiv, vorwissenschaftlich, unkontrolliert bis wirrköpfig etc angesehen wird. Die Kliniker sehen ihren "verstehenden" Ansatz als dynamisch, global, bedeutungsvoll, holistisch, subtil, konfigural, einfühlsam, tief, echt, lebensnah etc an und betrachten den statistischen Ansatz als mechanisch, atomistisch, additiv, künstlich, fraktioniert, trivial, übersimplifiziert, pseudowissenschaftlich, pedantisch etc. Statt die eine oder die andere Meinung zu vertreten, diskutiert Meehl (1954) 22 empirische Arbeiten, in denen die Qualität der Vorhersage einerseits von Experten (gemeint sind damit immer die 'Kliniker'), andererseits von 'Mechanikern', die ihre Vorhersage von Verhalten auf der Basis von psychometrischen Tests und damit auf der Basis der multiplen Regression machen. Der Vergleich geht eindeutig zugunsten der Mechaniker aus.

In einer Arbeit aus dem Jahr 1986 fand Meehl seinen Standpunkt durch weitere Arbeiten dieser Art bestätigt. Dawes (1996) liefert eine Reihe zum Teil tragischer Beispiele für Diagnosefehler von Experten nicht nur aus dem Bereich der Psychologie (vergl. insbesondere das Kapitel *Prediction and Diagnosis*). Ein relativ einfaches, aber schlagendes Beispiel ist das Aufnahmeverfahren zur Texas Medical School (TMS) aus dem Jahr 1979. Die TMS nahm bis dahin jährlich 150 Studierende auf. Auch 1979 wurden 150 Studierende aus einer Menge von 2200 Bewerbern ausgesucht. Auf der Basis der Bewerbungsunterlagen der Kandidaten wurden zunächst 800 Bewerber ausgesucht, die dann eingeladen wurden, um von einem Mitglied des Zulassungskommitees und von einem Mitglied der Medizinischen Fakultät interviewt zu werden. Die Interviewer lieferten für jeden Kandidaten schriftliche Beurteilungen an ein zentrales Kommitee, und jedes Mitglied dieses Kommitees lieferte von jedem Bewerber einen Rangwert (ranking) auf einer Skala von 0 (nicht akzeptierbar) bis 7 (exzellent). Diese Rangordnungen der Kommissionsmitglieder wurden dann gemittelt, so dass für jede(n) Bewerber(in) ein einzelner Rangwert, eben der mittlere Rangplatz, bestimmt wurde, und anhand dieser Rangwerte wurden dann wiederum alle 800 Bewerber in eine Rangordnung gebracht. Nach einigen weiteren Vergleichen, auf deren Details es hier nicht weiter ankommt, zeigte sich, dass die Bewerber mit den 150 besten Rankings auch in den Interviews am besten abschnitten.

Aber nach Abschluß dieses Auswahlverfahrens geschah es, dass die Regierung von Texas die TMS aufforderte, weitere 50 Bewerber aufzunehmen. Man beschloß

nun, einfach die nächsten 50 in der Rangordnung aufzunehmen, aber die hatten bereits andere Studienplätze gefunden. Nur noch 50 Bewerber mit Rangplätzen zwischen 700 und 800 waren noch wählbar, weil sie von keiner anderen Medical School genommen worden waren. Diese wurden nun ebenfalls aufgenommen. Den Professoren der Medical School wurde aber *nicht* mitgeteilt, welche der Studierenden zu den ersten 150 und den letzten 50 gehörten. So wurde ein Vergleich der Leistungen der ersten 150 und der letzten 50 Kandidaten möglich, der nicht von möglichen Voreinstellungen der Professoren beeinflußt werden konnte.

Der Studienerfolg der ersten 150 und der letzten 50 wurde dann studiumbegleitend anslysiert. Am Ende des zweiten Jahres gab es keinerlei Unterschiede zwischen den beiden Gruppen, nach dem Ende der klinischen Ausbildung ebenfalls nicht, und nach dem ersten Berufsjahr ebenfalls nicht. Es gab nicht nur keinen *signifikanten* Unterschied, sondern überhaupt keinen Unterschied. 82% in beiden Gruppen erhielten den M.D (hier: Dr. med.), und auch in weiteren Leistungen gab es keine Unterschiede. Wie die Untersucher zusammenfassten waren die Interviews "a total waste of time". Die Befunde an anderen Medical Schools entsprachen denen an der TMS. Weitere Beispiele für die fragwürdige Validität subjektiver Einschätzungen durch Experten findet man in Dawes (1996). Die folgende Liste enthält einige Gründe, deretwegen Experten aufgrund ihres "verstehenden" Ansatzes oft fehlerhaft urteilen:

1. Situative Faktoren wie Müdigkeit, unmittelbar vorangehende Faktoren ("recency effects"), geringfügige Veränderung in der Konzeptualisierung und Verarbeitung der vorliegenden Information etc erzeugen zufällige Schwankungen im Urteilsprozess. Dies äußert sich in einer reduzierten "Zuverlässigkeit" (*Reliabilität*) und damit Genauigkeit der Urteile,

2. Ein Symptom heißt *valide*, wenn es das vorherzusagende Merkmal tatsächlich anzeigt. Die Untersuchungen zum Urteilsverhalten von Menschen legen nahe, dass auch erfahrene Personen immer Schwierigkeiten haben, *valide* und *nicht valide* Symptome oder Merkmale zu unterscheiden. Klinische Psychologen und Psychiater erhalten in der Praxis oft keine genaue Rückkopplung über die Exaktheit ihrer Diagnosen, so dass "Erfahrung" konstituierende Lernvorgänge gar nicht stattfinden, weil es eben keine Rückkopplung gibt. Dawes (1996) zitiert eine Psychotherapeutin, die angibt, ihre Erfahrung mit aufgrund sexuellen Mißbrauch psychisch belasteten Frauen erlaube es ihr, sexuell mißbrauchte Frauen *an ihrem Gang* erkennen zu können. Die Therapeutin ist unter anderem Opfer dieser mangelnden Rückkopplung.

3. Menschen scheinen eine Neigung, ihre Hypothesen und Theorien bestätigen und ihnen widersprechende Informationen abweisen zu wollen. Wason (1960) hat dafür den Begriff des *konfirmatorischen Bias* eingeführt. Er gab seinen Vpn Folgen von 3 Zahlen vor, etwa 2, 4, 6, und die Aufgabe der Vpn war, die Regel zu finden, nach der diese Folge gebildet wurde. Bei der Beispielfolge wird z.B. suggeriert, dass die Folge aus geraden Zahlen besteht.

Die Vp muß Folgen erzeugen, von der sie meint, sie seien mit der gesuchten Regel vereinbar, und der Versuchsleiter teilt der Vp dann mit, ob eine solche Folge tatsächlich mit der Folge kompatibel ist oder nicht. So ist etwa die Folge 2, 17, 18 mit der Regel kompatibel, ebenso 1, 13, 52. Die Folge 15, 12, 18 ist es aber nicht. Wason beobachtete, dass Vpn eine starke Tendenz haben, ihre jeweilige Hypothese zu bestätigen. Optimal wäre es gewesen, sie auf ihre Gültigkeit zu testen, indem man die Hyothese zu falsifizieren sucht. Die wahre Regel war: *Jede Folge aufsteigender Zahlen ist zulässig.*

Dieses Verhalten läßt sich nicht nur in solchen "reduzierten" Experimenten finden. Es folgt einem allgemeinen Muster, – vermutlich, weil es zunächst einmal sehr schnelle Reaktionen ermöglicht. In diagnostischen Situationen kann es aber zu schlimmen Fehldiagnosen führen, weil es Urteile auf der Basis vorgefaßter Meinungen begünstigt. Ein Schlagwort für dieses Urteilsverhalten ist *Mortons Dämon*, nach ihrem Autor Glenn R. Morton, in Anlehnung an den Maxwellschen Dämon der Thermodynamik[5]. Mortons Dämon sitzt in den Augen und Ohren und läßt nur solche Informationen ins Hirn, die mit den von diesem produzierten Hypothesen konform sind[6] Man sollte meinen, dass Wissenschaftler eher zu kritischem Denken neigen; Gegenbeispiele haben Mahoney & DeMonbreun (1977) und Mahoney (1977) geliefert, die zeigten, dass Pfarrer oft sorgfältiger urteilen als Wissenschaftler.

4. Klinische Urteile können sich selbst erfüllende Prophezeiungen erzeugen. Ein Beispiel aus Dawes (1996) mag dies erläutern: in einem Mordprozess wurde ein Angeklagter von einem Psychiater als auch in Zukunft gewalttätig beurteilt. Der Angeklagte wurde daraufhin zum Tode verurteilt. In der Todeszelle verhielt sich der Verurteilte gewalttätig und schien deshalb die Beurteilung durch den Psychiater zu bestätigen. Andererseits hatte er nichts mehr zu verlieren; bei einer anderen Beurteilung und einer dementsprechend anderen Verurteilung hätte sein Verhalten völlig anders sein können.

5. Ein Verhalten erscheint im Allgemeinen, wenn es erst einmal eingetreten ist, als vorhersagbarer als zu dem Zeitpunkt, zu dem es noch vorhergesagt werden muß. Bereits gestellte Diagnosen erscheinen daher subjektiv konsistent mit den tatsächlichen Befunden zu sein. Die tatsächlichen Befunde haben auf diese Weise keine korrigierende Wirkung; die *Erfahrung* wird also nicht vermehrt. *Vor* der Diagnose durch einen Experten kann allerdings

[5]Maxwells Dämon ist imstande, schnelle Gasmoleküle von langsamen Gasmolekülen durch bloßes Öffnen und schließen von Türen, die entweder nur schnelle Moleküle aus einem Raum A durchlassen und langsame Moleküle nicht durchläßt, und umgekehrt langsame Moleküle aus einem Raum B in den Raum A läßt, aber schnelle nicht. Da langsame Moleküle wenig Wärme und schnelle Moleküle viel Wärme bedeuten, kann er auf diese Weise ein Temperatur- und damit Energiegefälle erzeugen, das zum Antrieb von Maschinen genutzt werden kann, – wenn es denn möglich wäre, den Laplaceschen Dämon zu installieren. Man kann zeigen, dass ein solcher Dämon nicht existieren kann.

[6] vergl. http://home.entouch.net/dmd/mortonsdemon.htm.

erhebliche Unsicherheit herrschen (Arkes et al., 1981).

6. Insbesondere Kliniker werden gewissermaßen Opfer ihres Berufes, weil sie beruflich mit einer speziellen Auswahl aus der Bevölkerung, nicht mit einer insgesamt repräsentativen Stichprobe zu tun haben. Die Beziehungen zwischen Symptomen und Merkmalen und den zu diagnostizierenden bzw. vorauszusagenden Merkmalen stellen sich deshalb verzerrt dar. Spreen (1981) (zitiert nach Dawes et al., 1989) berichtet dass die Hälfte der Jugendlichen, die wegen irgendwelcher Straftaten aufgefallen sind, im EEG leichte Abweichungen von als "normal" geltenden EEG-Strukturen zeigen. Also werden diese Abweichungen als Indikator für jugendliche Delinquenz gewertet. Tatsächlich kommen diese Abweichungen aber auch bei der Hälfte der *nicht* delinquent gewordenen Jugendlichen vor, – nur werden diese eben gar nicht erst untersucht. Tatsächlich sind derartige Abweichungen bei "normalen", nicht delinquenten Kindern und Jugendlichen ganz "normal" und geben keinerlei Hinweis auf zu erwartende Delinquenz. Gleichwohl: hat man erst einmal eine Hypothese gebildet, so neigt man dazu, sie gegen widersprechende Fakten zu *immunisieren*, so dass die Konsistenz der "Erfahrung" mit der Hypothese überschätzt wird. Dementsprechend wird die Gültigkeit der widersprechenden Information *unterschätzt*. Dieses Phänomen ist allgemein als urteilsverzerrender Faktor unter dem Namen

7. *Repräsentativität* (representiveness) bekannt. Der beurteilende Experte oder Diagnostiker bezieht sich bei seiner Beurteilung auf die Übereinstimmung einiger beobachteter Merkmale (z.B. EEG-Abweichungen) mit stereotypen Kategorien in seinem Gedächtnis, ohne die jeweiligen bedingten Wahrscheinlichkeiten zu berücksichtigen. Tversky und Kahneman (1974) betrachten das folgende Beispiel: Eine Person wird als scheu und zurückgezogen mit einer Neigung zu Ordnung und Detail beschrieben. Wie groß ist die Wahrscheinlichkeit, dass sie diese Person als (a) Bauer, (b) Handelsvertreter, (c) Pilot eines Verkehrsflugzeuges, (d) Bibliothekar, oder (e) Arzt einschätzen? Das Urteil wird dann, der Repräsentativitätsstrategie entsprechend, nach Maßgabe der Ähnlichkeit der Beschreibung der Person mit den Stereotypen der Berufe getroffen (z.B. "wahrscheinlich ist der Mann Bibliothekar"), nicht aber nach Maßgabe der relativen bedingten Häufigkeit (bedingte Wahrscheinlichkeit), mit der Angehörige der verschiedenen Berufe die beobachteten Eigenschaften tatsächlich haben.

   Täuschungen dieser Art sind spezielle Beispiele für die *Grundquotentäuschung* (*base rate fallacy*). Formal läßt sich am einfachsten anhand der bekannten Beziehung zwischen bedingten Wahrscheinlichkeiten beschreiben. Sind A und B zufällige Ereignisse und sind $P(A|B)$ und $P(B|A)$ die bedingten Wahrscheinlichkeiten von A, gegeben B bzw. von B, gegeben A, so gilt bekanntlich

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}. \tag{1.0.1}$$

$P(A)$ und $P(B)$ sind die a-priori-Wahrscheinlichkeiten, im hier gegebenene Zusammenhang oft auch *Grundquoten* oder *base rates* genannt. $P(B|A)$ wird in inferenzstatistischen Zusammenhängen auch die *Likelihood* von B genannt[7]. In Bezug auf (1.0.1) ist $P(A)/P(B)$ das Verhältnis der Grundquoten. Die Grundquotentäuschung besteht darin, dass vernachlässigt wird, dass im Allgemeinen $P(A)/P(B) \neq 1$, also $P(A) \neq P(B)$ gilt und ein Urteil nur auf der Basis einer Abschätzung der Liklihood gebildet wird. Begeht ein türkischstämmiger Jugendlicher (B) eine Gewalttat (A), so wird geschlossen, dass die Eigenschaft, türkischstämmig zu sein, eine Neigung zu Gewalttätigkeit impliziert. Tatsächlich bilden die Gewalttäter eine kleine Teilmenge der türkischstämmigen Jugendlichen, d.h. $P(A)/P(B) \ll 1$, wobei $\ll$ sehr klein im Vergleich zu 1 bedeutet, im Vergleich zu $<$ für einfach 'kleiner als 1'. Zu diskutieren wäre, ob der Anteil Gewalttäter an der deutschen Bevölkerung von vergleichbarer Größenordnung ist oder nicht, bevor vom Merkmal 'türkischstämmig' auf 'gewalttätig' geschlossen wird.

Ein anderer verzerrender Einfluß auf die Urteile nicht nur von "normalen" Menschen, sondern auch von Experten ist die

8. *Verfügbarkeit* (*Availability*) (Tversky et al., 1974). Hier wird die Wahrscheinlichkeit, mit der eine Person ein Merkmal hat, in Abhängigkeit von der Anzahl von Beispielen, die man dafür im Gedächtnis hat beurteilt. Das Risiko von Herzinfarkten von Menschen im mittleren Alter wird nach Maßgabe der Häufigkeit, mit der in der eigenen Bekanntschaft solche Herzinfarkte aufgetreten sind, beurteilt. Die Schätzung hängt nun aber davon ab, wie gut man solche Beispiele erinnern kann: Herzinfarkte sind wegen ihrer drastischen Konsequenzen leichter zu erinnern als weniger saliente Merkmale. In jedem Fall kommt es hier zu Fehlabschätzungen, die zu drastischen Fehlbeurteilungen führen können.

Die Vorhersagen der 'Mechaniker' beruhen häufig auf einer Anwendung der multiplen Regression, bei der die einzelnen Symptome (Subtests) optimal gewichtet werden, wobei der Begriff 'optimal' durch die jeweils gewählte Schätzmnethode näher spezifiert wird, etwa durch die Methoded der Kleinsten Quadrate oder die Maximum-Likelighood-Methode. Die Überlegenheit des mechanischen Ansatzes scheint in erster Linie auf der korrekten Auswahl der Prädiktoren, also der Symptome, zu beruhen, denn selbst bei suboptimalen Gewichten ist die Vorhersage oft noch besser als die von Klinikern. Wie gut der Regressionsansatz funktioniert, hat unlängst Ayres (2007) noch einmal illustriert. Er beginnt sein Buch mit einem Beispiel, der Kunst des Weinkostens. Wer Weinkenner kennt oder gar selbst ei-

[7] Da gleichermaßen

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)}$$

gilt, kann auch $P(A|B)$ als Likelihood gelten, allerdings ist der Ausdruck 'Likelihood' i.a. reserviert für den Fall, dass B Daten repräsentiert und A eine Hypothese.

ne(r) ist, weiß um die komplexen Adjektive, mit denen Experten hier jonglieren, kennt die Diskussionen darüber, ob es sich lohnt, einen Wein zu lagern, damit er an Wert gewinnt, oder ob man ihn besser gleich oder gar nicht trinkt. Ayres zitiert den 'number cruncher' Orley Ashenfelter, der selbst Weinliebhaber ist. Er hat herausgefunden, dass man auf den "swishing and spitting approach" der Weingurus verzichten kann. Ashenfelter hat eine (Regressions-) Formel gefunden:

$$\mathrm{WQ} = 12.145 + .00115 \times \mathrm{RW} + .0614 \times \mathrm{TW} - .00386 \times \mathrm{RE}. \tag{1.0.2}$$

Dabei steht WQ für Weinqualität, RW für Regenfall im Winter, TW durchschnittliche Temperatur während der Wachstumsphase und RE für Regenfall in der Wachstumsphase. Ashenfelter gelang es, mit dieser Formel die Rangordnung der Weindomänen vorherzusagen, wie sie von den Franzosen geliefert werden. Das für traditionelle Weinexperten Unangenehme der Ashenfelter Formel ist ihre Einfachheit. Die Weinindustrie in Bordeaux habe dementsprechend "somewhere between violent and hysterical" reagiert. Robert Parker, Weinguru und Autor von *The Wine Advocate* nannte Ashenfelters Formel "an absolute total sham". Ashenfelters Formel "is really a Neanderthal way of looking at wine. It's so absurd to be laughable". Mathematische Formeln können nicht, so Parker, zur Beurteilung von Wein herangezogen werden, und: "I'd hate to be invited to his house to drink wine". Dass Ashenfelters Formel die Qualtität der Weine so bemerkenswert gut vorhersagen kann, ist für Parker kein Anlaß, seine Meinung zu überdenken; die emotionale Wucht seiner Reaktionen steht in umgekehrtem Verhältnis zur tatsächlichen Effizienz der Vorhersagen auf der Basis von (1.0.2). In Ayres Kapitel 'Experts versus Equations' wird eine Reihe weiterer Beispiele für die Vorhersagekraft schon einfacher Regressionsgleichungen geliefert. Ayers zitiert die Kognitionspsychologen Nisbett und Ross: "Human judges are not merely worse than optimal regression equations; they are worse than almost any regression equation." Goldberg ist bereits 1960 zum gleichen Schluß gekommen.

Die Geschichte des psychometrischen Testens verzeichnet keinen Siegeszug dieses Ansatzes von Anfang an. Die Irrungen und Wirrungen sind u.a. gut dokumentiert in Gould (1983). Wie sehr insbesondere mit dem Intelligenzquotienten interessenorientierte Politik gemacht werden kann, zeichnet Kamin (1974) nach. In neuerer Zeit haben Herrnstein & Murray (1994) noch einmal nachgelegt und anhand umfangreichen Datenmaterials nachzuweisen versucht, dass ökonomischer Erfolg in erster Linie durch Intelligenz und diese wiederum durch Gene bestimmt seien. Da Schwarze in den USA eher am Ende der ökonomischen Hackordnung stehen, meinen die Autoren nachgewiesen zu haben, dass ein Mensch im Durchschnitt um so weniger intelligent ist, je mehr "schwarze" Gene er hat. Das Buch hat Empörung ausgelöst, vergl. z. B. Blinkhorn (1994). Devlin et al. (1997) haben einen Band mit kritischen Analysen herausgebracht, die die Behauptungen Herrnstein & Murrays regelrecht zermahlen; zu empfehlen ist insbesondere Glymores Kapitel *Social statistics and general inquiry: reflections on the Bell Curve*. Der Autor zeigt, dass es (mindestens) sieben verschiedene Strukturgleichungsmodelle gibt, die die Herrnstein-Murray-Daten erklären, ohne auf deren genetische Hy-

pothese rekurrieren zu müssen. Glymours Analysen zeigen, dass anscheinend so hermetische Analysen wie die von Herrnstein & Murray keinesfalls so eindeutig sind, wie sie vorgeben zu sein. Ob Herrnstein und Murray ihre Ansichten daraufhin geändert haben, ist nicht bekannt, Mortons Dämon wird sein Werk getan haben.

Viele Kritiker der Anwendung von Statistik auf soziale und speziell psychologische Fragestellungen argumentieren, dass statistische Methoden dazu führen, dass irrige Behauptungen (wie die Herrnstein et al.s) in positivistischer Weise zementiert würden. Diese Gefahr scheint zu bestehen: Menschen haben nun einmal die Neigung, nach Bestätigungen für ihre Ansichten zu suchen, und wer zu rassistischen Ansichten neigt, wird sich über Herrnstein et al.'s Analysen freuen und sie als 'wissenschaftlich fundierte Analyse' bei Diskussionen ins Feld führen. Allerdings verdeutlicht gerade Glymours Arbeit, dass die statistische Analyse *für sich genommen* noch nicht die behauptete positivistische Zementierung impliziert, sondern dieser Zementierung entgegen wirken kann. Wo qualitativ-hermeneutische Betrachtungen eine Eindeutigkeit von Dateninterpretationen suggerieren, können statistische Analysen eben auch semantische Ambiguität von Daten und allgemein Erfahrungen offenbaren.

Auch Herrnstein und Murrays Daten waren Testergebnisse, also Daten, die aus der Anwendung objektiver, reliabler und in vieler Hinsicht valider Tests entstanden. Aber Validität erlaubt nicht so weitreichende Folgerungen, wie sie von Herrnstein und Murray vorgenommen wurden, und statistische Verfahren – in diesem Falle Strukturgleichungsmodelle – liefern nicht notwendig eindeutige Resultate. Hier ist der kritische Verstand der Experten gefordert; kritisch, indem man versucht, des konfirmatorischen Bias (Mortons Dämon) Herr zu werden. Poppers Falsifikationstheorie ist in der Wissenschaftstheorie stark kritisiert worden, aber als heuristische Grundposition ist nicht zuletzt ihr humanistischer Wert unbestritten: eine perfekte, also in allen Fällen korrekte Methode der Diagnose kann es schon deshalb nicht geben, weil beobachtete Verhaltensweisen, Testergebnisse etc stets nur Stichproben sind, deren Struktur sich oft durch verschiedene Theorien oder Modelle in gleicher Weise interpretieren läßt.

2 Grundbegriffe der Testtheorie

2.1 Itemparameter und Itemtypen

Die Betrachtungen des vorangegangenen Abschnitts motivieren die folgende Charakterisierung von Testitems, um die Wahrscheinlichkeit von Diagnosefehlern zu minimalisieren:

1. **Objektivität:** Das Item ist objektiv, wenn verschiedene Testleiter zur gleichen Bewertung der Antworten oder Reaktionen auf das Item bezüglich

des untersuchten Merkmals gelangen. Die Beurteilungen durch verschiedene Testauswerter sollten hoch miteinander korrelieren.

2. **Reliabilität:** Die Reliabilität, d.h. Zuverlässigkeit eines Tests reflektiert die Genauigkeit des Tests, mit der er dasjenige Merkmal, das er erfassen soll, mißt. Man kann das spezifizieren, indem man sagt, dass ein Test um so reliabler ist, je kleiner die Varianz der Fehler ist relativ zur Varianz des Merkmals in der Population.

3. **Validität:** Ein Test ist valide, wenn es das Merkmal, das er erfassen soll, auch tatsächlich erfasst, dh die Beantwortung des Items sollte hoch mit der Ausprägung des untersuchten Merkmals korrelieren.

4. **Trennschärfe:** Ein Test, insbesondere ein Item ist *trennscharf*, wenn es erlaubt, zwischen Personen, die das untersuchte Merkmal in unterschiedlicher Ausprägung haben, zu differenzieren.

5. **Schwierigkeit:** Ein Test bzw. ein Item kann für verschiedene Personen unterschiedlich schwierig sein, wobei "schwierig" auch auf Items angewendet wird, die sich nicht auf irgend eine Leistungsfähigkeit beziehen: für einen Gegner der Todesstrafe ist es "schwierig", einem Item der Art "Kindesmord sollte mit dem Tode bestraft werden." zuzustimmen. Die Ausprägung des gemessenen Merkmals bei einer Person kann jedenfalls relativ zu den Schwierigkeiten der Items bestimmt werden.

6. **Homogenität versus Heterogenität:** Komplexität der Items eines Tests. Items sind *homogen*, wenn stets nur ein Merkmal erfasst wird. Dies kann auch bedeuten, dass zwar mehrere Merkmale bei der Beantwortung eine Rolle spielen, diese Merkmale aber in einer bestimmten, fixen Kombination in die Antworten eingehen.

   Die Items sind heterogen, wenn verschiedene Merkmale in nicht-fixer Kombination in die Antworten eingehen. Man sagt dann auch, dass das gemessene Merkmal mehrdimensional ist.

Die Problematik des Objektivitätsbegriffs wird durch seine operationale Definition als Korrelation zwischen den Beurteilungen verschiedener Testauswerter ein wenig verwischt. Denn eine hohe Korrelation heißt ja zunächst nur, dass sich die Beurteiler hinsichtlich der Kriterien, nach denen eine Antwort beurteilt wird, einig sind, und natürlich dass sie die Antworten von Testpersonen in gleicher Weise zu diesen Kriterien in Beziehung setzen können. Die Kriterien reflektieren bestimmte Normen und theoretische Voreinstellungen, deren Subjektivität wegen der Gleichförmigkeit, mit der sie in einer Gesellschaft oder einer speziellen Gruppe innerhalb einer Gesellschaft vertreten werden, gewissermaßen verdeckt wird. Wird zum Beispiel in einer Gruppe von Psychologen die Theorie, dass das Schädelvolumen die Höhe der Intelligenz bestimmt, als wahr akzeptiert, so wird man zu sehr hohen Urteilsübereinstimmungen kommen.

Natürlich ist Reliabilität ein wünschenswertes Merkmal. Eine geringe Reliabilität reduziert die Aussagekraft einer Messung, da sie nur wenig über die "wahre" Ausprägung des gemessenen Merkmals aussagt. Andererseits ist eine hohe Reliabilität nur eine notwendige, nicht aber auch hinreichende Bedingung für eine gute Messung. So kann man Schädelvolumina sehr reliabel messen, aber ob diese Messungen tatsächlich etwas über die Intelligenz der Besitzer der Schädel aussagen ist eine ganz andere Frage.

Diese Frage ist die nach der Validität des Tests. Objektive und reliable Messungen von Schädelvolumina dürften in Bezug auf die Intelligenz eine sehr geringe Validität haben, – nur in pathologischen Ausnahmen wird ein Schluß auf die Intelligenz zulässig sein. Die tatsächliche Bestimmung der Validität kann sehr schwierig sein, da sie im Grunde die Existenz eines validen Tests T_M für das Merkmal voraussetzt. Die Validität eines Tests oder Items T soll sich ja im weitesten Sinne als Korrelation mit der tatsächlichen Ausprägung des Merkmals bestimmen lassen, dh also mit den Messungen anhand von T_M. So kann man die Validität eines Intelligenztests nur dann bestimmen, wenn man für jede getestete Person schon weiß, wie hoch oder niedrig ihre tatsächliche Intelligenz ist, und dazu wiederum muß man sich auf einen Intelligenzbegriff einigen; dementsprechend ist der Intelligenzbegriff ein *Konstrukt*, wie andere Persönlichkeitsmaße, die mit Tests erfasst werden sollen, ebenfalls. Man kann dementsprechend von *Konstruktvalidität* sprechen. Die zu einem bestimmten Konstrukt führenden Debatten können sich, wie beim Intelligenzbegriff, über Jahrzehnte hinziehen, und die Resultate der Debatten entsprechen dann nicht notwendig den Intuitionen, die Nichtpsychologen zum Beispiel über die Intelligenz haben, und hier soll nur festgehalten werden, dass die Intuitionen der Laien – die von einem Laien zum anderen durchaus verschieden sein können – nicht notwendig sinnvoller sind als die mühsam erarbeiten Konstrukte der Psychologen[8]. Der Begriff des Konstrukts ist keineswegs charakteristisch für die Psychologie, auch in anderen Wissenschaften werden Begriffe vielfach als Konstrukte eingeführt – man erinnere sich an die historische Debatte über den Kraftbegriff in der Physik.

Die Trennschärfe ist sicherlich ein wichtiger Parameter eines Tests oder eines Items. Ist etwa das untersuchte Merkmal "Depressivität", so sollte ein Item, das Depressivität reflektiert, nur dann positiv beantwortet werden, wenn die be-

[8]Ein Professor für Statistik an der Universität Konstanz erklärte Intelligenztests für offenkundigen Blödsinn, da seine von ihm geschiedene Ehefrau einen höheren IQ attestiert bekam als er. Ein Historiker erklärte IQ-Tests für Blödsinn, weil z.B. bei Hermann Göring einen IQ = 138 (manche sagen, einen IQ = 142) gemessen hat, was nach heutigen Maßstäben Hochbegabung bedeutet: einen Menschen wie Göring könne man nicht als intelligent bezeichnen. Argumente dieser Art basieren auf nur implizit und oft ideosynkratisch bestimmten Begriffen (z.B.: Das Wort Intelligenz leite sich sich aus dem lateinischen intellegere ab, das so viel wie verstehen, einsehen bedeute, und IQ-Tests würden die Fähigkeit zur Einsicht doch gar nicht erfassen, etc, etc), die den empirisch gegebenen Sachverhalt, dass viele Persönlichkeitsmerkmale nicht oder nur sehr wenig miteinander korrelieren, übersehen. Schon Hegel unterschied zwischen dem "tabellarischen Verstand" und der "Vernunft".

fragte Person auch depressiv ist. Ein Item, dass sowohl von Nichtdepressiven wie auch von Depressiven positiv beantwortet wird, trennt nicht zwischen diesen Personen und ist insofern wertlos. Offenbar ist der Begriff der Trennschärfe einerseits mit dem der Validität, andererseits mit dem der Reliabilität gekoppelt. Mangelnde Trennschärfe kann einfach ein Ausdruck mangelnder Validität sein. Mißt man Intelligenz, indem man das Schädelvolumen bestimmt, so findet man für jedes Volumen sowohl hohe wie mittlere und niedrige Intelligenz, dh das Schädelvolumen trennt nicht zwischen Personen mit unterschiedlicher Intelligenz. Andererseits kann eine Denksportaufgabe ein valides Maß für einen bestimmten Inelligenzaspekt sein, weil der Lösungsprozess eine Reihe von kognitiven Fähigkeiten voraussetzt. Haben diese Fähigkeiten aber eine intrinsische Variabilität, so kann es geschehen, dass eine Person die Aufgabe in der Situation A gut lösen, in der Situation B kann aber nicht lösen kann, wobei A und B auch einfach für zwei verschiedene Zeitpunkte stehen kann. Ist die Variabilität groß relativ zu den durchschnittlichen Fähigkeiten der Personen, trennt die Aufgabe nicht gut zwischen den Personen; es ergibt sich eine geringe Trennschärfe.

Die Schwierigkeit eines Items läßt sich auf verschiedene Weise operationalisieren. In der Klassischen Testtheorie (KTT) definiert man als Schwierigkeitsindex einfach nur den Prozentsatz oder Anteil der Personen in einer Population, die die Aufgabe lösen bzw. in einem bestimmten Sinne beantworten. Andererseits kann man die Schwierigkeit eines Items auch durch einen Parameter in einem Modell repräsentieren, das das Antwortverhalten der Probanden abbilden soll. Bei der KTT-Definition ist der Wert der Schwierigkeit populationsabhängig: ein Item, das in der Population der Hauptschüler als 'schwierig' gilt, weil viele Hauptschüler die Aufgabe nicht lösen können, kann in der Population der Gymnasiasten als eher 'leicht' gelten, weil ein höherer Prozentsatz die Aufgabe löst. Die Definition der Schwierigkeit als Wert eines Parameters eines Modells, das das Antwortverhalten modelliert, wird in den Item-Response-Modellen (IRT-Modelle) gewählt. Erweist sich die Schwierigkeit als ein Parameter, der von den Personen unabhängig ist, wird man ihn zur Definition der Schwierigkeit in der KTT in Beziehung setzen können. Allerdings ist es nicht denknotwendig, die Schwierigkeit eines Items als eine personenunabhängige Größe zu sehen. So kann eine Denksportaufgabe für zwei Personen mit gleichen kognitiven Fähigkeiten unterschiedlich schwierig sein, einfach weil die eine Person z.B. gelernt hat, Aufgaben betrachteten Typs in Gleichungen zu übersetzen, deren Lösung dann trivial ist, die andere Person diese Fertigkeit aber nicht gelernt hat und deswegen nur schwer einen Zugang zur eigentlichen Struktur des Problems hat. In den IRT-Theorien wird dieser Aspekt des Schwierigkeitsbegriffs relativ schnell deutlich, während er in der KTT eher verdeckt bleibt: schätzt man die Schwierigkeit eines Items als Anteil der Personen in einer Stichprobe, die das Item lösen oder "positiv" beantworten, so hängt dieser Anteil offenkundig stark von der Zusammensetzung der Stichprobe ab.

2.2 Aufgabentypen

Das Format der Items eines Tests hängt von der Art des Tests und der Zielsetzung, mit der der Test durchgeführt wird, ab. Es kann sich um Fragen, die frei beantwortet werden, handeln, oder es wird, etwa bei einem Fragen nach einer Einstellung zu einem Thema, nur nach Zustimmung oder Ablehnung gefragt. Generell kann man ein Item durch zwei Komponenten beschreiben: der Itemstamm und das Antwortformat (Rost (2004), p. 55). Der Itemstamm kann aus einer einfachen Frage ("Werden Sie leicht rot?"), einem Bild (etwa in einem projektiven Test: ein traurig aussehender junger Mann mit Geige), eine Rechenaufgabe ("Welche Zahl ist die nächste in der Reihe $2, 3, 4, 9, 8, 27, 16, \ldots$?"), etc bestehen. Das Antwortformat definiert die jeweilige Reaktion oder eben Antwort der getesteten Person. Das Antwortformat muß so gewählt werden, dass die Antwort möglichst eindeutige Rückschlüsse auf das Merkmal, das mit dem Test erfasst werden soll, erlaubt. Insbesondere bei projektiven Verfahren können sich hier Schwierigkeiten ergeben, wenn nämlich das Antwortformat überhaupt nicht definiert ist, d.h. wenn die getestete Person völlig frei antworten und die testende Person diese Antworten wiederum frei interpretieren kann und verschiedene Testleiter oder Testinterpreten zu idiosynkratischen Deutungen der Antworten kommen können. Freie Antwortformate können gleichwohl bei bestimmten Tests angezeigt sein, etwa, wenn spontane Antworten gewünscht werden, die durch gebundene Formate unterdrückt werden können. Freie Antwortformate reduzieren u. U. die Wahrscheinlichkeit, dass die getestete Person "punktet", indem sie einfach rät, – dies wäre bei Leistungstests von Nachteil (Statistikklausuren sollten deshalb nicht als Multiple-Choice-Tests formuliert werden).

Bei gebundenen Antwortformaten werden i. A. Kategorien für die möglichen Antworten vorgegeben. Ein klarer Vorteil dieses Formats ist die höhere Auswertungsökonomie und eventuell eine höhere Objektivität, die aber den Nachteil reduzierter Validität implizieren kann, wenn nämlich die Kategorien nicht alle möglichen Reaktionsweisen erfassen. Ein Spezialfall der gebundenen Formate sind die *Ratingformate*. Ein solches Format besteht aus mehr als zwei Antwortkategorien, die hinsichtlich des getesteten Merkmals eine Rangordnung repräsentieren.

Die Antworten auf die Testitems werden im Rahmen eines bestimmten Testmodells ausgewertet; dies kann die KTT sein oder ein Modell aus der Klasse der IRT-Modelle. Die Wahl des Modells kann die Art der Kodierung der Antworten beeinflussen. Bei freien Antwortformaten lassen sich zwei Phasen der Kodierung unterscheiden: (i) eine Kategorisierung, (ii) eine Signierung (Rost (2004), p. 78). Die Signierung etwa bei Antworten in projektiven Tests setzt u. U. ein spezielles psychologisches Fachwissen voraus. Der Signierung schließt sich die Zuordnung von Messwerten an, die im Rahmen des gewählten Testmodells definiert werden, sowie die Schätzung der Parameter des Modells.

Bei der Kodierung von Antwortkategorien wird unterschieden zwischen

1. der *dichotomen* Kodierung, also einer Kodierung in zwei Kategorien, und

2. der *polynomen* Kodierung, also einer und der Einordnung in entweder geordnete (Rangordnung) oder ungeordnete Kategorien.

Eine Kodierung ist dichotom, wenn eine Antwort als entweder richtig oder falsch klassifiziert wird, oder eine Aussage als Zustimmung oder Ablehnung einer Meinungsäußerung gewertet wird. Es wird eine Kodierungsvariable X definiert, die entweder den Wert 1 ("richtig", "Zustimmung") oder 0 ("falsch", "Ablehnung") annimmt. Man kann diese Definition von X knapp in der Form $X \in \{0, 1\}$ anschreiben, d.h. X ist Elemente einer Menge, die nur aus den Zahlen 0 und 1 besteht. Man definiert dann eine Zuordnung etwa der Art: $X = 0$, wenn die getestete Person etwa eine Aufgabe nicht gelöst oder einer Meinungsaussage ("Der freie Markt regelt alles") nicht zugestimmt hat, und $X = 1$, wenn sie die Aufgabe gelöst oder dem Item zugestimmt hat. X ist eine zufällige Veränderliche, denn einerseits weiß man nicht von vorn herein, welche Fähigkeit oder Ansicht eine gegebene Person hat (die Person wird in diesem Sinne zufällig gewählt), andererseits fluktuieren auch innerhalb einer Person Fähigkeiten und Ansichten, so dass die Antworten einer Person nicht von Zeitpunkt zu Zeitpunkt deterministisch vorhergesagt werden können.

Bei einer polynomen Kodierung wird mindestens eine weitere Kategorie benützt, also etwa ausser "Ablehnung" und "Zustimmung" noch "unentschieden/neutral". Die Kategorien können durchnummeriert werden, so dass eine zufällige Variable definiert wird, deren mögliche Werte die den Kategorien zugeordneten Zahlen sind. Man betrachte noch einmal das Item "Der freie Markt regelt alles". Dem Item kann man zustimmen oder auch nicht. Man kann auch eine abgestufte Antwort zulassen, indem man verbal etikettierte Kategorien vorgibt, etwa "trifft nicht zu", "trifft ein wenig zu", trifft einigermaßen zu", "trifft voll zu". Numerische Kodierungen dienen hier eher der Auswertung als einer Benennung der Kategorien für die Befragungspersonen. Mögliche Kodierungen könnten -2, -1, 1, 2, oder 0, 1, 2, 3, oder 1, 2, 3, 4 sein. Die Relationen zwischen den Zahlen spiegeln die Abstufungen, dh die *ordinalen* Relationen, wieder. Ob sie auch eine Intervallskaleneigenschaft reflektieren, ist eine offene Frage, die explizit untersucht werden muß. In jedem Fall lassen sich die verschiedenen numerischen Repräsentationen durch lineare Transformationen ineinander überführen, – was *noch nicht* eine implizite Annahme der Intervallskaleneigenschaft bedeutet. So findet man die Transformation von -2, -1, 1, 2 in 0, 1, 2, 3, indem man $0 = -2a + b$ und $3 = 2a + b$ setzt. Es folgt $a = 3/4$ und $b = 3/2$. Bei der Repräsentation -2, -1, 1, 2 könnte man noch eine neutrale Kategorie 0 einfügen, so dass die Repräsentation -2, -1, 0, 1, 2 entsteht; die 0 repräsentiert dann eine Kategorie der Art "Ich weiß es nicht". Repräsentationen der Art -2, -1, 1, 2 oder -2, -1, 0, 1, 2 heißen auch *polar*. Neutrale Positionen, wie sie durch die 0 abgebildet werden, haben allerdings oft den Nachteil, dass sie gewählt werden, weil die befragte Person das Item eigentlich gar nicht beantworten wollen und die Interpretation der 0-Position als neutrale

Position deswegen gar nicht gerechtfertigt ist.

2.3 Raten und Wissen

Die Antwort einer Person auf eine Testfrage wird, bei der Interpretation der Antworten im Rahmen eines psychometrischen Tests, kodiert, d.h. es wird ihr eine Maßzahl zugeordnet. Im einfachsten Fall notiert man, ob die Person dem Item zugestimmt hat oder nicht, oder ob sie die Aufgabe – wenn das Item eine Aufgabe ist – gelöst hat oder nicht. Die "Maßzahl" ist dann zunächst einfach der Wert einer Indikatorvariable X, die nur zwei Werte annehmen kann, etwa 0 oder 1: $X = 1$, wenn die Person zugestimmt hat, $X = 0$, wenn sie nicht zugestimmt hat, oder $X = 1$, wenn die Aufgabe gelöst wurde, und $X = 0$ sonst. Man schreibt kurz $X \in \{0, 1\}$. Wird bei Items eine Schätzskala vorgegeben, etwa von "stimme nicht zu" über "weiß nicht" bis zu "stimme zu", kann man den Kategorien Zahlen zuordnen, etwa von 0 bis 3, oder -3 bis + 3, etc. Bei einer dritten Art von Aufgaben werden r Antwortkategorien vorgegeben und die Person muß die als "richtig" oder "korrekt" definierte Kategorie ankreuzen. Dies sind die Multiple-Choice-Aufgaben. Die Definition der Zuordnung von Antworten zu numerischen Werten besteht in der Angabe einer *item scoring rule*

Die Frage ist nun, in welcher Weise ein Gesamtwert berechnet werden soll. Im einfachsten Fall summiert man einfach die Werte, die ein Proband bei den einzelnen bekommen hat. Die Werte für die einzelnen Items können aber auch in gewichteter Form in den Gesamtwert eingehen; man muß zu diesem Zweck eine Gewichtungsformel definieren. Sind die y_j, $j = 1, \ldots, n$ die Scores für die einzelnen Items, so kann der Gesamtscore gemäß

$$x = \sum_{j=1}^{n} (a_j y_j + b_j) \tag{2.3.1}$$

definiert werden (Lord & Novick (1968), p. 303), wobei die Gewichte a_j und b_j durch noch zu spezifizierende Kriterien näher bestimmt werden.

Muß die getestete Person bei Aufgaben zwischen Antwortalternativen wählen, so kann es sein, dass sie zwar die korrekte Antwort nicht weiß, aber die korrekte Antwort rät. Die Frage ist nun, wie der Gesamtscore bewertet werden kann, wenn man davon ausgeht, dass einige der korrekten Antworten nur geraten wurden. Werden $r \geq 2$ Antwortkategorien oder mögliche Antworten vorgegeben und ist nur eine dieser Kategorien die korrekte Kategorie, so ist die Wahrscheinlichkeit, dass die Person die korrekte Antwort nur rät, gleich $1/r$, und die Wahrscheinlichkeit, dass sie die inkorrekte Antwort gibt, ist gleich $1 - 1/r = (r - 1)/r$. Der erwartete Score bei einer Aufgabe, bei der geraten wird, ist dann $\mathbb{E}(X_j) = 1p_j + 0(1 - p_j) = p_j$; mit $p_j = P(X_j = 1)$. Gibt es n Aufgaben mit jeweils r Antwortkategorien, so ist der erwartete Score, wenn nur geraten

wird,

$$x_c = \mathbb{E}(X) = \sum_{j=1}^{n} = n\mathbb{E}(X_j) = \frac{n}{r}. \tag{2.3.2}$$

Die Varianz des Scores, wenn nur geraten wird, ergibt sich gemäß

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = n\,p(1-p) = n\left(\frac{r-1}{r^2}\right), \tag{2.3.3}$$

mit $p_j = p$ für alle j. Die Varianz $\mathbb{V}(X)$ bzw. die Standardabweichung $\sqrt{\mathbb{V}(X)}$ kann eine Abschätzung des Bereiches von X-Werten liefern, der auf Raten zurückgeht.

2.4 Testwerte, Testskalen und Normierungen

Die in diesem Abschnitt eingeführten Begriffe liefern einen in der Praxis üblichen, gleichwohl intuitiven Zugang zur Interpretation von Testdaten, der im Rahmen insbesondere der Klassischen Testtheorie weiter formalisiert wird.

Für ein gegebenes Item I_g liefert oder erhält eine Testperson $a \in \mathcal{P}$ den Test- bzw. Messwert x_{ag}. x_{ag} repräsentiert die Ausprägung eines Merkmals, das im Test erfasst werden soll. Ist das Item dichotom, so wird x_{ag} den Wert 1 oder 0 annehmen: 1, wenn die Aufgabe gelöst wurde, 0, wenn sie nicht gelöst wurde, oder 1, wenn die Person dem (Meinungs-)Item zugestimmt hat, 0 sonst. $x_{ag} = 1$ signalisiert, dass die Person a die Fähigkeit hat, die g-te Aufgabe zu lösen, oder das abgefragte Merkmal (”Sind Sie gelegentlich melancholisch?”) in hinreichendem Ausmaß hat. Handelt es sich um eine Rating-Aufgabe, so ist $x_{ag} = j$, $j \in \mathbb{N}$,d.h. j ist im Allgemeinen eine natürliche Zahl, etwa $0 \leq j \leq 7$, wenn die vorgegebenen Antwortkategorien von 0 bis 7 durchnumeriert worden sind. Alternativ dazu könnten die Antwortkategorien -3, -2, -1, 0, 1, 2, 3 sein, so dass $-3 \leq j \leq 3$ gilt, etc. Natürlich kann x_{ag} auch ein Wert auf einer Verhältniskala sein, wie etwa ein Blutdruckwert, ein galvanischer Hautwiderstand, oder eine Hormonkonzentration.

Aus den Märchen wissen wir, dass es üblich war, den Bewerber um die Hand der schönen Königstochter zu testen, bevor der Bewerber in die nähere Wahl kam. Er mußte eine schwierige Aufgabe[9] lösen, oder – sozusagen zur Sicherheit – auch drei, möglicherweise unter Zeitdruck oder anderen, unangenehmen Nebenbedingungen. Das Problem bei einer solchen Versuchsanordnung sind die zufälligen

[9] So wird er etwa in einen Raum geführt, aus dem zwei Türen hinaus führen. Durch die eine gelangt er zur Königstochter, durch die andere zum Schafott, auf dem er seinen Kopf lassen muß, wenn er diese Tür wählt. Er weiß nicht, welche Tür wohin führt. Vor jeder Tür steht ein Wächter, von denen einer stets die Wahrheit spricht, während der andere stets lügt, – aber es wird ihm nicht gesagt, welcher der Wächter die Wahrheit spricht. Nun darf er einem Wächter seiner Wahl genau eine Frage stellen, die der Wächter nur mit ”ja” oder ”nein” beantworten darf, – wie lautet die Frage?

Fehler. Ein wenig talentierter Bewerber kann des Rätsels Lösung bereits kennen, ohne diese selbst gefunden zu haben, oder ein talentierter Bewerber kann durch eine Schwankung der Aufmerksamkeit oder wegen einer unvorhergesehenen Unpäßlichkeit die Lösung in der vorgegebenen Zeit verfehlen. Die Antwort des Bewerbers spiegelt also nicht notwendig den wahren Wert seiner Fähigkeiten wieder.

So ist es auch in einem psychometrischen Test oder einem Fragebogen. Man wird mehrere Aufgaben oder Fragen als Stichprobe aus einer Population von Aufgaben oder Fragen stellen, um den Einfluß zufälliger Fehler zu reduzieren. Deshalb kann man sagen, dass x_{ag} die Realisierung einer zufälligen Veränderlichen ist. Ist die Aufgabe dichotom, so ist der Wertebereich dieser Variablen $\{0, 1\}$, was kurz in der Form $X \in \{0, 1\}$ geschrieben werden kann. Handelt es sich bei einer Frage um eine Ratingaufgabe, so ist $x_{ag} = j$, j eine natürliche Zahl, die eine der möglichen Antwortkategorien kennzeichnet. Handelt es sich um psychophysiologische Messungen, so nimmt x_{ag} üblicherweise einen Wert aus einem Kontinuum von Werten an.

Gesamttestwert: Enthält der Test m Items, so wird der Gesamttestwert X_a der Person a oft als die Summe der Einzeltestwerte definiert:

$$X_a = \sum_{g=1}^{m} x_{ag}. \tag{2.4.1}$$

Natürlich ist es möglich, die einzelnen Antworten noch zu gewichten, so dass der Gesamtwert durch

$$X_a = \sum_{g=1}^{m} \alpha_g x_{ag} \tag{2.4.2}$$

gegeben ist, wobei α_g irgendwelche Gewichte sind. Man kann die Gewichte wie Regressionsgewichte interpretieren, und die x_{ag} wie Werte von Prädiktorvariablen. Für dichotome Items ist X_a/m, m die Anzahl der Items, gerade die relative Häufigkeit der gelösten bzw. "positiv" beantworteten Items. Da die x_{ag} zufällige Variablen sind, ist auch ihre Summe X_a eine zufällige Variable.

Bekanntlich soll man Äpfel und Birnen nicht addieren, es sei denn, die Rede ist nicht von Äpfeln und Birnen, sondern von Früchten. In Bezug auf (2.4.1) bedeutet dies, dass die Summierung nur Sinn macht, wenn man die x_{ag} als Maße für die Ausprägung ein und desselben Merkmals auffassen kann. Man sollte also voraussetzen können, dass die Items *homogen* sind, dass sie tatsächlich nur ein Merkmal messen. Die Forderung nach Homogenität ist nicht ganz trivial: so kann man mit Adorno, Frenkel-Brunswik, Levinson und Sanford (1950) der Meinung sein, dass die Eigenschaft, autoritär zu sein, ein konstitutives Merkmal für die Eigenschaft 'faschistoid' ist. Man stellt nun Items zusammen, von denen man annimmt, dass sie eine Ausprägung des Merkmals 'autoritär' abbilden, und schließt von 'autoritären' Antworten auf eine faschistoide Persönlichkeit, und von 'nicht

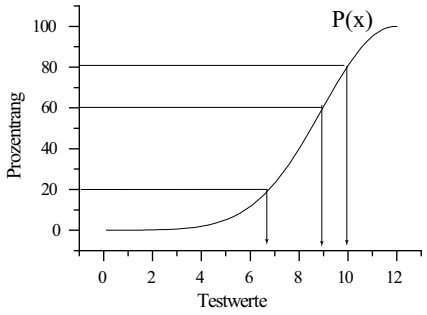
autoritären' Antworten auf eine nicht faschistoide Persönlichkeit. Nun kann es aber sein, dass auch nicht-faschistoide, aber linksradikale Personen 'autoritäre' Antworten geben, und es kann Personen geben, die 'nicht-autoritäre' Antworten geben, obwohl sie entweder faschistoid oder links-radikal sind. Die Position auf einer politischen links-rechts-Skala sind nicht oder nur wenig mit der Ausprägung auf der autoritär versus nicht-autoritär Skala korreliert. Geht man aufgrund theoretischer Vorstellungen von einer festen Kopplung von 'autoritär' und 'faschistoid' aus, so kommt es bei der Auswertung der Daten zu ausgeprägten Fehlinterpretationen. Hinzu kommt, dass Items Mischungen dieser beiden Dimensionen erfassen.

In der Tat haben faktorenanalytische Untersuchungen gezeigt, dass man autoriär sein kann, ohne faschistoid zu sein, und faschistoid kann man sein, ohne autoritär zu sein. Addiert man also einfach die Testwerte für diese Mischung von Items, so erhält man einen Gesamtwert, der irgendeine Mischung von Eigenschaften zeigt, ohne dass man wüßte, wie diese Mischung zusammengesetzt ist. Für's Erste ist also festzuhalten, dass die Definition (2.4.1) des Gesamtwerts Homogenität der Items voraussetzt, und ob eine Menge von Items homogen ist, muß mit geeigneten Methoden überprüft werden.

Rohwerteverteilung: Die Angabe eines Gesamtwerts macht eigentlich nur Sinn, wenn er in irgendeiner Weise einen Vergleich ermöglicht: ob X_a eine große, mittlere oder geringe Merkmalsausprägung ausdrückt, läßt sich nur beurteilen, wenn man diesen Wert in Beziehung setzt zu den Werten, die andere Personen erzielt haben. Dieser Sachverhalt führt dazu, dass man die Häufigkeitsverteilung der Gesamtwerte in der Population betrachtet. Man spricht auch von der Rohwerteverteilung des Tests. Sind die Aufgaben des Tests eher "leicht" relativ zur Verteilung der Merkmalsausprägungen in der Population, so wird sich eine rechtssteile Häufigkeitsverteilung ergeben, sind die Aufgaben eher "schwierig", so wird man eine linkssteile Verteilung finden. Sind die Aufgaben im Mittel von mittlerer Schwierigkeit, so wird man eher eine symmetrische Verteilung finden. Es ist auch möglich, dass die Häufigkeitsverteilung multimodal ist. Dies ist der Fall, wenn sich die Population aus Teilpopulationen zusammensetzt, die sich hinsichtlich der Merkmalsausprägung jeweils im Mittel unterscheiden. Mißt man z.B. die Fähigkeit, seine Aufmerksamkeit auf verschiedene Aspekte gleichzeitig – was auch immer das heißen mag – zu verteilen, so sollen Frauen hierin im Durchschnitt besser sein als Männer. Betrachtet man also eine Population, die als Subpopulationen Frauen und Männer enthält, und schneiden Frauen im Durchschnitt besser im Test ab als Männer, so kann es geschehen, dass man eine bimodale Verteilung erhält.

Viele Test- und Skalenkonstrukteure/innen freuen sich, wenn die Häufigkeitsverteilung Anlaß zu der Annahme gibt, dass die Punktwerte normalverteilt sind. Der Vorteil eines solchen Befundes ist, dass nun viele der inferenzstatistischen Techniken angewendet werden können, die die Normalverteilung voraussetzen. Insbesondere kann man Konfidenzintervalle berechnen, Varianzanalysen rechnen,

Abbildung 1: Prozentränge und ihre Interpretation; $P(x)$ ist ein Beispiel für eine Funktion, die *nicht* die Gauß-Funktion ist. Gleiche Differenzen von Prozenträngen (20 - 0, 80 - 60) entsprechen nicht gleichen Unterschieden zwischen Testwerten.



– etwa, um den Effekt von Interventionsmaßnahmen zu überprüfen, etc. So hat sich zm Beispiel die Annahme durchgesetzt, dass der Intelligenzquotient (IQ) in der Bevölkerung normalverteilt sei. Hat man nun einen Test entwickelt, von dem man der Ansicht ist, er messe Intelligenz, für den aber die Verteilung der Punktwerte einer hinreichend großen Stichprobe nicht normalverteilt ist, so kann man u. U. eine Normalverteilung der Punktwerte erreichen, indem man bestimmte Items des Tests durch andere ersetzt, – so lange, bis die erwünschte Verteilungsform erreicht wurde. Dieses Vorgehen ist nicht nur in der Intelligenzmessung beliebt und hat eine gewisse Ähnlichkeit mit dem Prinzip der sich selbst erfüllenden Prophezeiung.

Prozentränge: Eine Möglichkeit, eine Testleistung x_{ag} relativ zu einer Bezugspopulation zu bewerten, ist die Bestimmung des zu x_{ag} korrespondierenden Prozentrangs. Gegeben seien die Testwerte in einer Eichstichprobe vom Umfang N. Dann ist

$$p(x) = \frac{1}{N} \sum_{X_a \leq x} n(X_a)$$

die kumulative Verteilung der X_a-Werte (für alle a); $p(x)$ entspricht der Wahrscheinlichkeit, dass ein Messwert x_{ag} kleiner, höchstens gleich x ist. Dann ist $100 \times p(x)$ der entsprechende Prozentwert der Messwerte kleiner/gleich x.

Definition 2.4.1 *Der* Prozentrang $P(x) = 100 \times p(x)$ *gibt an, wieviel Prozent der Mitglieder der Eichstichprobe einen Messwert kleiner und höchstens gleich x erhalten haben.*

Man kann nun jedem möglichen Testwert X_a den Prozentrang $P(X_a)$ zuordnen; man weiß dann, wieviel Prozent der Mitglieder der Eichstichprobe einen Testwert

kleiner, höchstens gleich X_a erhalten haben; $100 - P(X_a)$ haben einen größeren Messwert erhalten. Man spricht auch von einer *Prozentrangnorm*. Hat also eine Person einen Testwert $X_a = 120$ und gilt $P(120) = 95$, so haben 95% der Eichstichprobe einen kleineren Testwert erhalten, nur 5% haben einen größeren Wert erzielt.

Über die Prozentrangnorm erfährt also ein Proband, wie seine Testleistung relativ zu den Leistungen in der Eichstichprobe einzuordnen ist. Der Vorteil dieser Norm ist, dass sie unabhängig von der speziellen Verteilung der Testwerte ist: nirgends ging eine Annahme über die Verteilung, etwa die Annahme der Normalverteilung, ein.

$P(X_a)$ ist eine nicht-lineare Funktion des Testwerts X_a; im Allgemeinen hat P einen sigmoidalen Verlauf (ähnlich der bekannten Gaußschen Verteilungsfunktion), vergl. Abbildung 1, wo $P(X)$ sigmoidal, aber nicht Gaußsch ist. Es gilt stets

$$0 \leq P(X_a) \leq 100,$$

unabhängig vom Bereich der X_a-Werte. Die Nichtlinearität von $P(X)$ impliziert, dass die Bedeutung der Differenz zweier Prozenträngе von den Prozenträngen selbst abhängt: Die Differenz 20 - 0 = 20 repräsentiert den Testwertebereich 0 bis ca 6.8, während die Differenz 80 - 60 = 20 den sehr viel kleineren Testwertebereich von ca 9 bis 10 abbildet. Differenzen von Prozenträngen können also nicht wie Differenzen von Intervallskalenwerten interpretiert werden. Dies gilt natürlich auch, wenn $P(x)$ eine Gauß-Funktion ist!

z-Normierung: Eine andere Möglichkeit, die Position eines Probanden relativ zu den Werten einer Eichstichprobe zu definieren, besteht in der Standardisierung, d.h. in der z-Transformation. Dazu sei $\bar{x}$ das arithmetische Mittel aller Testwerte in der Eichstichprobe, und $\hat{s}_x$ sei die Streuung dieser Werte.

Definition 2.4.2 *Dem Mess- oder Testwert X_a eines Probanden a kann der standardisierte Wert*

$$z(X_a) = \frac{X_a - \bar{x}}{\hat{s}_x} \tag{2.4.3}$$

*zugeordnet werden. Dies ist die z-*Normierung *für den Testwert.*

$z(X_a)$ hat offenbar die Form $z(X_a) = \alpha X_a + \beta$, mit $\alpha = 1/\hat{s}_x$, $\beta = -\bar{x}/\hat{s}_x$, d.h. die z-Transformation ist, im Gegensatz zur Prozentrangnorm, eine lineare Transformation.

Es sei $z(X_a) = 1$; dann folgt offenbar

$$X_a - \bar{x} = \hat{s}_x,$$

d.h. X_a weicht gerade um den Wert der Streuung vom Mittelwert $\bar{x}$ ab. Generell kann man also sagen, dass $z(X_a)$ die Abweichung von X_a vom Mittelwert $\bar{x}$

in Einheiten der Streuung, also in $\hat{s}_x$-Einheiten, angibt. Die z-Normierung ist ebenfalls verteilungsfrei, d.h. es muß keine bestimmte Verteilung angenommen werden. Will man keine Verteilung annehmen, erleichtert die Tchebyscheffsche Ungleichung die Interpretation. Nach dieser Ungleichung gilt

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \tag{2.4.4}$$

für $k = 2$ hat man also $P(|X-\mu| \geq 2\sigma) \leq 1/4 = .25$, d.h. 75% der Testwerte liegen innerhalb des Bereichs $\pm 2\sigma$ bzw. $\pm\hat{s}_x$. Für $k = 3$ erhält man $P(|X - \mu| \geq 3\sigma) \leq 1/9 = .11$, d.h. ca 11% der Werte liegen außerhalb des Intervalls $(\bar{x} - 3\hat{s}_x, \bar{x} + \hat{s}_x$, 89 % liegen innerhalb dieses Intervalls. Kann eine bestimmte Verteilung, etwa die Normalverteilung, angenommen werden, können natürlich schärfere Aussagen gemacht werden.

Weitere Normierungen: Handelt es sich bei dem Test um einen Intelligenztest, so ist es üblich, den Rohwert X_a in einen IQ-Wert zu transformieren. Dazu wird davon ausgegangen, dass die IQ-Werte in der Population normalverteilt sind. Man z-transformiert zunächst den Punktwert X_a in den z-Wert $z(X_a)$, und berechnet dann

$$IQ(X_a) = 100 + 15z(X_a). \tag{2.4.5}$$

Statt der IQ-Transformation wird gelegentlich auch die T-Transformation

$$T(X_a) = 50 + 10z(X_a) \tag{2.4.6}$$

gewählt. Wiederum alternativ dazu wird gelegentlich die *Stanine-Transformation* vorgenommen:

$$S(X_a) = 5 + 2z(X_a). \tag{2.4.7}$$

Diese Transformation bedeutet, dass dem X_a-Wert ein Stanine-Wert $S(X_a)$ zugeordnet wird, der zwischen 1 und 9 liegt ("standard nine").

Die Transformationen liefern bestimmte Skalen; etwas sophistiziertere Skalen werden in den Abschnitten 2.6.6 und 2.6.7 vorgestellt.

Anmerkungen:

1. **Ein- und Mehrdimensionalität:** Die Bildung eines Gesamttestwerts ("Summenscore") setzt im Allgemeinen voraus, dass die Messwerte für die einzelnen Items Ausprägungen ein und desselben Merkmals bedeuten. Repräsentieren einige Items die Merkfähigkeit, andere wiederum rein arithmetische Fähigkeiten, so ergibt sich die Frage, was die Summe der Scores aus diesen beiden (Unter-)Tests bedeuten soll. Der Intelligenzquotient (IQ) ist im Prinzip eine solche Summe über verschiedene Testleistungen. Man kann nun sagen, dass jede solche Summenbildung sinnlos ist und postulieren, dass jede Summenbildung strikte Eindimensionalität – es wird nur ein Merkmal

gemessen – voraussetzt. Man kann aber auch argumentieren, dass die verschiedenen Merkmale (Merkfähigkeit, Rechenfähigkeit) Aspekte eines komplexen Merkmals sind, in das die Messwerte der einzelnen Merkmale, möglicherweise gewichtet, additiv eingehen. Aus der Statistik ist dieses Vorgehen bekannt: bei der multiplen Regression wird z.B. der Einfluß des Geschlechts, des Alters, der Ausbildung etwa auf das Ausmaß der Zufriedenheit einer Person betrachtet.

2. **Eichstichprobe:** Die Interpretation eines Gesamtwertes X_a für eine Person a entweder in über den Prozentrang $P(X_a)$ oder die Standardisierung $z(X_a)$ geschieht durch Bezug auf eine Eichstichprobe, von der angenommen wird, dass sie eine Stichprobe aus einer bestimmten Population ist. Damit die Interpretation in Bezug auf die Population sinnvoll ist, muß die Eichstichprobe repräsentativ für diese Population sein; zur Bildung einer solchen Stichprobe gelten dann die üblichen Regeln, wie sie in der Statistik eingeführt werden.

   Dies gilt natürlich ebenso für alle Transformationen, die aus der z-Transformation abgeleitet worden sind, wie die IQ-, die T- und die Stanine-Transformation.

3. **Populationsbezug:** Der Bezug auf eine Eichstichprobe bedeutet den Bezug auf eine Population. Daraus ergibt sich die Frage, ob man die Messwerte zweier Personen, die aus verschiedenen Populationen stammen, miteinander vergleichen kann. So können Messwerte, die mit dem gleichen Test an Mitgliedern verschiedener sozialen Schichten erhoben wurden, verschiedene Bedeutungen haben.

4. **Testwert als Wert einer zufälligen Veränderlichen:** Wie oben angemerkt wurde, sind die x_{ag} und damit auch die Gesamtwerte X_a möglicherweise mit zufälligen Fehlern behaftet. Da die Transformationen, vom Prozentrang bis zum Stanine-Wert, direkt am Messwert vorgenommen werden, sind auch die transformierten Werte zufällige Veränderliche. Dies führt auf die Frage, ob der beobachtete Wert vom "wahren" Wert abweicht, und diese Frage führt auf die Frage, was denn unter einem solchen Wert verstanden werden soll. Eine Möglichkeit, diesen Begriff zu spezifizieren, besteht in der Annahme, dass die Zufälligkeit des Messwertes durch einen additiven Messfehler ε entsteht. Der Messfehler repräsentiert alle nicht weiter kontrollierbaren Effekte, die auf die Messung einwirken und eine Abweichung von dem Wert, der die tatsächliche Merkmalsausprägung repräsentiert, erzeugen.

5. **Zur Annahme der Normalverteilung:** Die Annahme eines additiven Messfehlers ε liegt vielen statistischen Analysen zugrunde, man denke an die Varianz- und Regressionsanalyse. Die übliche Annahme ist dann, dass ε normalverteilt ist. Diese Annahme wird meistens durch Hinweis auf den

Zentralen Grenzwertsatz begründet, meistens ohne Überprüfung der Bedingungen, die diesem Satz zugrundeliegen. Darüber hinaus wird oft angenommen, dass ε unabhängig vom wahren Wert ist. In der Tat hat es Vorteile, einen vom wahren Wert unabhängigen, normalverteilten Fehler annehmen zu können, denn dann steht ein elaboriertes Arsenal von inferenzstatistischen Verfahren zur weiteren Interpretation der Testwerte zur Verfügung.

Darüber hinaus wird gelegentlich postuliert, die Ausprägung des gemessenen Merkmals sei in der Population normalverteilt, z. B. die Intelligenz, wie sie sich in IQ-Werten ausdrückt. Jede Person a einer Population $\mathcal{P}$ hat also eine für sie charakteristische "wahre" Intelligenz, und die Wahrscheinlichkeit, dass eine zufällig gewählte Person einen IQ etwa zwischen 115 und 125 hat, ist durch $\Phi(125) - \Phi(115)$ gegeben, wobei $\Phi(x)$ die durch die Gauß-Verteilung gegebene Wahrscheinlichkeit ist, einen IQ kleiner/gleich x zu haben. Wählt man also eine Person zufällig aus $\mathcal{P}$ aus, so wählt man mit ihr zufällig einen wahren IQ aus, dem sich bei der Messung der Intelligenz noch ein normalverteilter Messfehler überlagern kann.

Die Annahme einer in der Population normalverteilten Intelligenz geht auf Galton[10] zurück, der sie durch Analogiebetrachtungen rechtfertigte: so wie die Körpergröße in guter Näherung normalverteilt ist, so könnte ja auch die Fähigkeit des Gehirns in guter Näherung normalverteilt sein. Zudem kann durch geeignete Wahl von Items immer approximativ normalverteilte Testwerte erreicht werden. Sicherlich kann die Annahme einer Normalverteilung für Testwerte immer nur eine Approximation sein, denn für eine normalverteilte zufällige Veränderliche X gilt $-\infty < X < \infty$, und eine negative Intelligenz ist schwer vorstellbar. Insofern könnte es sinnvoller sein, X als auf $[0, \infty)$ definiert anzunehmen und eine entsprechende Verteilung zu wählen. Dann ginge allerdings der enorme praktische Vorteil (man denke an all die Verfahren, die die Normalverteilung voraussetzen) dieser Verteilung verloren. Ob ein Merkmal in der Population approximativ normalverteilt ist oder nicht ist letztlich eine empirische Frage.

Natürlich ist man am "wahren" Wert der Merkmalsausprägung einer Person interessiert. In den vorangegangenen Betrachtungen wurde von der naheliegenden Konzeption eines "wahren Testwerts" ausgegangen; es muß nun elaboriert werden, wie man von den beobachteten Werten zu den in diesem Sinne wahren Werten gelangt. Diese Elaboration findet in der *Klassischen Testtheorie* (KTT) statt. Für viele Arten von Messungen kann der Ansatz der KTT sinnvoll sein. Ob er generell für psychometrische Messungen sinnvoll ist, ist eine andere Frage. Man denke etwa an Tests, deren Items dichotom sind, wenn etwa die Antworten auf Fragen nur als entweder "richtig" oder "falsch" gewertet werden. X_a ist dann gleich der Anzahl der als "richtig" gewerteten Antworten. Ist die Anzahl der Item des Tests hinreichend groß, kann man wieder davon ausgehen, dass wegen des Zentra-

[10] Francis Galton (1822 – 1911), (Halb-)Cousin von Charles Darwin.

len Grenzwertsatzes X_a approximativ normalverteilt ist. Die Annahme eines vom wahren Wert unabhängigen Fehlers ist aber kaum noch zu rechtfertigen; schon im Falle einer Binomialverteilung, die ja für hinreichend großen Wert von n durch eine Normalverteilung approximiert werden kann, sind ja Erwartungswert $\mu_a = np$ und die Varianz $\sigma_a^2 = np(1-p)$ nicht unabhängig voneinander, d.h. man kann X_a nicht in der Form $X_a = \mu_a + \varepsilon$, ε unabhängig von μ_a, konzipieren, denn die Unabhängigkeit widerspricht dem Sachverhalt, dass $\mathbb{V}(\varepsilon) = np(1-p) = \mu_a(1-p)$ ist. Analoge Betrachtungen lassen sich für Tests anstellen, bei denen Antworten auf Ratingskalen eingefordert werden. Es wird sich zeigen, dass gerade die dichotomen Items eine Analyse erfordern, die von dem oben skiziierten intuitiven Ansatz nicht entsprechen.

Die Interpretation von Testdaten setzt offenbar ein bestimmtes Modell voraus, das den Rahmen für die Interpretation definiert. Im folgenden Abschnitt werden die wichtigsten Modelle vorgestellt.

2.5 Testmodelle

Grundlage für die Konstruktion eines psychometrischen Tests ist eine Theorie, anhand der die Antworten eines Probanden auf das zu messende Merkmal bezogen werden. Es gibt zwei verschiedene Ansätze: (i) die Klassische Testtheorie (KTT), und (ii) die Probabilistische oder Item-Response-Theorie (IRT). Beide sollen hier kurz vorgestellt werden, damit bestimmte Begriffsbildungen vorausgesetzt werden können. Ausführlicher diskutiert werden die beiden Theorien in späteren Abschnitten.

Im Folgenden wird weitgehend von der durch Lord & Novicks (1968) eingeführten Notation Gebrauch gemacht: $\mathcal{P}$ ist die Population, in der der Test angewendet werden soll, $a \in \mathcal{P}$ bezeichnet eine bestimmte Person aus dieser Population, g ist der Index für ein Item, d.h. I_g bezeichnet das g-te Item eines Tests, $g = 1, 2, \ldots, m$.

Allerdings wird die in neueren Texten üblich werdende Bezeichnung $\mathbb{E}$ für den Erwartungswert einer zufälligen Veränderlichen eingeführt, damit eine Unterscheidung zwischen der Bezeichung E für "Fehler", wie sie in Lord & Novick und in vielen anderen Texten zur Testtheorie üblich ist, und dem Erwartungswert, der in Statistiktexten ebenfalls mit E bezeichnet wird, möglich ist (Lord & Novick benutzen $\mathcal{E}$). Die Bezeichnung $\mathbb{V}$ für die Varianz ist kürzer als die oft verwendete Schreibweise Var und korrespondiert im Erscheinungsbild zu $\mathbb{E}$, ebenso $\mathbb{K}$ für die Kovarianz (Kov in vielen Statistiktexten).

1. **Die Klassische Testtheorie, KTT:** Die KTT ist eine Ausformulierung der im vorangegangenen Abschnitt 2.4 vorgestellten intuitiven Betrachtungen zum Gesamttestwert. Für eine Person a wird der Testscore X_a, überlicherweise die Summe der Scores (Punktwerte) für die einzelnen Aufgaben oder Fragen, berechnet. Für jede Person gibt es einen für sie charakte-

ristischen "wahren" Wert; dies ist der Erwartungswert $\tau_a = \mathbb{E}(X_a)$. Der tatsächliche Wert X_a weicht im Allgemeinen von diesem wahren Wert ab: die Abweichung $\varepsilon_a = X_a - \mathbb{E}(X_a)$ ist der Messfehler. Die Abweichung ε_a ist eine zufällige Veränderliche. Diese Begriffe werden schon für einzelne Items eingeführt: für den Messwert x_{ag} der a-ten Person beim g-ten Item wird die Zerlegung $x_{ag} = \tau_{ag} + \varepsilon_{ag}$ eingeführt, mit analoger Deutung von τ_{ag} und ε_{ag}. τ_{ag} bzw. τ_a kann nicht direkt beobachtet, sondern muß geschätzt werden. Man spricht auch von *Messfehlertheorie* (Steyer und Eid, 1991). Eine ausführlichere Diskussion des Begriffs des wahren Wertes wird in Abschnitt 3.1, Seite 140, gegeben, in dem die KTT dargestellt wird. Es wird postuliert, dass ε_{ag} bzw. ε_a von τ_{ag} bzw. τ_a unabhängig ist.

Im Falle dichotomer Items ist x_{ag} mehr Indikatorfunktion als Messwert im üblichen Sinne: $x_{ag} = 1$ zeigt an, dass die Frage "richtig" beantwortet wurde. Ist p_{ag} die Wahrscheinlichkeit, dass die Person a die g-te Frage richtig beantwortet, so ist $\mathbb{E}(x_{ag}) = \tau_{ag} = p_{ag}1 + (1 - p_{ag})0 = p_{ag}$. Man wird sinnvollerweise vermuten, dass p_{ag} von der gemessenen Merkmalsausprägung abhängt. Diese habe bei der Person a den Wert θ_a, so dass $p_{ag} = p_{ag}(\theta)$ angenommen werden kann, d.h. die Wahrscheinlichkeit p_{ag} ist eine Funktion von θ. Es liegt nun nahe, den Begriff des wahren Wertes mit dem Parameter θ bzw. θ_a in Beziehung zu setzen. Damit hat man dann aber den Übergang zu der folgenden Theorie vollzogen:

2. **Die Probabilistische oder Item-Response-Theorie (IRT)** Bei dieser Theorie wird die Wahrscheinlichkeit einer Antwort zu einem Fähigkeitsparameter θ_a, der die Person $a \in \mathcal{P}$ repräsentiert, und einem Schwierigkeitsparameter κ_g für das g-te Item direkt in Beziehung gesetzt. Beide Parameter müssen aus den Daten geschätzt werden, wobei θ_a und κ_g unabhängig voneinander geschätzt werden können, wenn bestimmte Bedingungen erfüllt sind. Aufgrund dieser Unabhängigkeit werden die Fähigkeitsparameter τ_a und θ_a sowie die Schwierigkeitsparameter π_g und κ_g nicht nur jeweils verschiedene Werte haben, sondern müssen auch verschieden interpretiert werden. Eine weitergehende Diskussion der Bedeutung der Parameter θ_a und κ_g erfolgt in Abschnitt 2.6.1.

Die KTT wird gelegentlich eine deterministische Theorie genannt, wohl weil der Personenparameter μ_a eine feste Größe ist. Andererseits ist der entsprechende Personparameter θ_a in der probabilistischen Theorie ebenfalls eine Konstante, – zufällige Fluktuationen des zu messenden Merkmals werden in beiden Theorien *nicht* angenommen. Zumindest werden zufällige Prozesse im Antwortverhalten nicht explizit diskutiert.

Lord und Novick (1968) haben die Charakterisierung der KTT als eine deterministische Theorie als irreführend bezeichnet. Der tatsächlich erhobene Testwert X weicht ja zufällig vom "wahren" Wert μ_x ab. Insbesondere wird angenommen, dass der "Fehler" ε Gauß-verteilt ist. Tatsache ist aber, dass die Beziehung zwi-

schen X, τ_x und der Wahrscheinlichkeit, dass X einen bestimmten Wert annimmt, einigermaßen indirekt ist, während in den probabilistischen Theorien die Beziehung zwischen dem Personenparameter, der Schwierigkeit eines Items und der Wahrscheinlichkeit einer Antwort (Lösung einer Aufgabe, Zustimmung zu einem Meinungsitem, etc) sehr direkt erscheint.

2.6 Tests und latente Merkmale

2.6.1 Antwortfunktionen

Die Messungen, also die Antworten der Personen und die Kodierung der Antworten sind die *manifesten Daten*, oder auch Realisierungen *manifester Variablen*. Das zu messende Merkmal, das oft nicht direkt beobachtet werden kann, heißt dementsprechend *latente Variable* (*latent trait*). Das Ziel einer Testtheorie ist, von den manifesten auf die latenten Variablen schließen zu können. Ein zentraler Begriff beider Theorien, der KTT wie der IRT, ist desbalb der der *Antwortfunktion*, auch *Itemfunktion*, *Item Characteristic Function* (Lord & Novick (1968), p. 360) oder *Item Response Function* bzw. einfach *Response Function*. Intuitiv gesprochen kann man sagen, dass die Itemfunktion dem Personenparameter θ, der die Ausprägung des gemessenen Merkmals repräsentiert, eine Wahrscheinlichkeit zuordnet, mit der eine Aufgabe (Item) gelöst oder einer in einem Item geäußtern Meinung zugestimmt wird; der Einfachheit halber wird im Folgenden nur von Aufgaben, die gelöst oder nicht gelöst werden, gesprochen; dies sind die *dichotomen Items*. Für diese Items ist ja der "wahre Wert" τ_{ag} einer Person für das g-te Item gerade gleich der Wahrscheinlichkeit p_{ag}, mit der das Item beantwortet wird, wie im vorangegangenen Abschnitt bereits ausgeführt worden ist. Die Verallgemeinerung für den Fall polynomer Items ist dann nicht weiter schwierig. Eine eher formale Definition der Itemfunktion ist die

Definition 2.6.1 *Es sei $X_g = \{0, 1\}$ eine Indikatorfunktion für das Item I_g, mit $X_g = 1$, wenn I_g gelöst wurde, und $X_g = 0$ sonst. θ repräsentiere die Ausprägung des gemessenen Merkmals. Die Funktion*

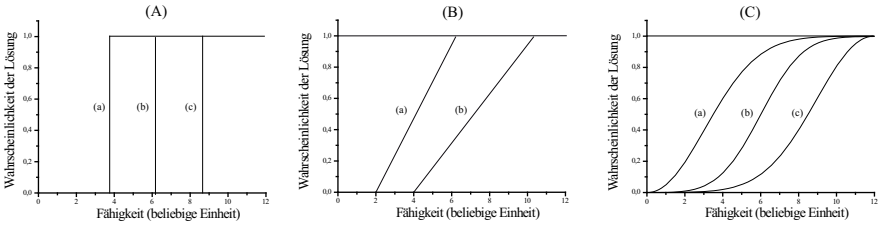
$$F_g(\theta) = P(X_g = 1|\theta), \quad 0 \leq F_g(\theta) \leq 1 \tag{2.6.1}$$

heißt Itemfunktion *für das Item I_g. Gilt $F_g(\theta) \leq F_g(\theta')$ für alle (θ, θ') mit $\theta \leq \theta'$, so heißt F_g* mononoton*, anderfalls heißt F_g* nicht-monoton*.*

Anmerkungen:

1. Da $F_g(\theta)$ eine Wahrscheinlichkeit ist, nämlich die des zufälligen Ereignisses $X_g = 1$, folgt, dass in jedem Fall $0 \leq F_g(\theta) \leq 1$. Monotone Itemfunktionen ergeben sich z.B. in Leistungstests, wenn die Wahrscheinlichkeit, eine Aufgabe zu lösen, mit dem Merkmal, das θ repräsentiert, wächst; die Wahrscheinlichkeit, die Aufgabe $17.352 \times \sqrt{\pi} =?$ innerhalb von fünf Sekunden

Abbildung 2: Mögliche monotone Itemfunktionen; zur weiteren Erläuterung siehe den Kommentar 2 im Text.



im Kopf zu lösen, steigt mit der rechnerischen Begabung. Auch bei Einstellungsfragen finden sich monotone Itemfunktionen: die Wahrscheinlichkeit, dem Item "Die Lebensaufgabe der Frauen ist, sich der Wiederherstellung der Arbeitskraft des Mannes zu widmen" steigt mit der Ausprägung einer talibanesken Grundeinstellung. Andererseits ergeben sich nicht-monotone Itemfunktionen gerade auch bei Einstellungsfragen. So stimmen dem Item "Kindesmißhandlungen sollten mit fünf Jahren Gefängnis bestraft werden" nur diejenigen zu, die der Ansicht sind, dass fünf Jahre Gefängnis das richtige Strafmaß sind, während alle diejenigen, die diese Strafe entweder zu hoch oder zu niedrig sehen, nicht zustimmen.

2. Die Abb. 2 zeigt drei mögliche Klassen von Itemfunktionen. In (A) sind die Itemfunktionen Schrittfunktionen; im Rahmen der Testtheorie spricht man auch von Guttman-Funktionen, nach Guttman (1950). Solche Funktionen sind gemäß

$$F_g(\theta) = \begin{cases} 0, & \theta < \kappa_g \\ 1, & \theta \geq \kappa_g \end{cases} \tag{2.6.2}$$

definiert: Ist die Fähigkeit kleiner als ein bestimmter θ-Wert, etwa $\theta = \kappa_g$ für das g-te Item I_g, so ist die Wahrscheinlichkeit einer Antwort gleich Null, ist sie größer als κ_g, so ist die Wahrscheinlichkeit einer Antwort gleich 1. Bei der Konstruktion von Guttman-Skalen wird diese spezielle und sicherlich nicht sehr plausible Annahme gemacht. (B) zeigt lineare Itemfunktionen,

$$F_g(\theta) = \begin{cases} 0, & \theta < \theta_{g1} \\ \alpha_g\theta + \beta_g, & \theta_{g1} \leq \theta \leq \theta_{g2} \\ 1, & \theta > \theta_{g2} \end{cases} \tag{2.6.3}$$

Da $\alpha_g\theta_{g1} + \beta_g = 0$ folgt $\beta_g = -\alpha_g\theta_{g1} = -\kappa_g$. Dieses Modell wird ergibt sich u.a. im Zusammenhang mit der Klassischen Testtheorie (KTT), wenn dichotome Items betrachtet und bestimmte Zusatzannahmen über die Beziehung zwischen dem Erwartungswert des Scores und der Ausprägung θ des gemessenen Merkmals gemacht werden. Dieser Fall wird in Abschnitt 3.8.2, Seite 200, ausführlich diskutiert.

Die Itemfunktionen (a), (b) und (c) in (C) haben verschiedene Formen. Die Funktion (a) steigt erst schnell mit der Fähigkeit θ, dann aber langsam. Bei der Funktion (c) ist es umgekehrt: die Funktion steigt erst langsam, um dann schnell gegen 1 zu streben. Die Funktion (b) ist symmetrisch in dem Sinne, dass in der gleichen Weise, wie Wahrscheinlichkeit einer Lösung bis zum Wendepunkt beschleunigt wächst, vom Wendepunkt an entschleunigend gegen 1 strebt. Bei (a) und (c) handelt es sich um Weibull-Funktionen, bei (b) um eine Gauß-Funktion (also um die kumulierte Normalverteilung). Die genaue mathematische Definition der Funktionen ist an dieser Stelle aber nicht wesentlich, es kommt darauf an, zu illustrieren, dass man zwischen verschiedenen Klassen von Funktionen wählen kann. Welche man wählt, sollte von psychologischen Betrachtungen abhängen. Betrachtungen dieser Art werden aber in der Testtheorie üblicherweise nicht angestellt. Statt dessen werden entweder ad-hoc-Argumente benützt: die Gauß-Funktion wird mit Hinweis auf den Zentralen Grenzwertsatz eingeführt und akzeptiert, weil sie als "Normal"verteilung eben die Standardverteilung ist, die bei vielen statistischen Analysen unterstellt wird. Die logistische Verteilung wurde zunächst gewählt, weil sie der Gauß-Funktion sehr ähnlich ist, mathematisch aber leichter zu handhaben ist. Später zeigte sich dann, dass die logistische Verteilung Eigenschaften hat, die sie für einige Autoren als die einzig sinnvolle Itemfunktion erscheinen läßt. Darauf wird weiter unten noch eingegangen.

3. Testtheorien werden im Allgemeinen als Meßtheorien konzipiert. Dies bedeutet, dass in erster Linie nach einer numerischen Repräsentation der Ausprägungen eines Merkmals gesucht wird. Sind $a \in \mathcal{P}$ und $a' \in \mathcal{P}$ zwei Personen aus der untersuchten Population, so werden also Zahlen θ_a und $\theta_{a'}$ mit $a \mapsto \theta_a$, $a' \mapsto \theta_{a'}$ gesucht derart, dass $\theta_a \neq \theta_{a'}$, wenn sich a und a' hinsichtlich des untersuchten Merkmals unterscheiden; ist das Merkmal quantitativ, so soll insbesondere $\theta_a < \theta_{a'}$ gelten, wenn bei a das Merkmal weniger ausgeprägt ist als bei a'. Worauf es nun ankommt, ist, dass die Itemfunktion F_g diese Relationen korrekt abbildet. Dazu wird, um F_g zu spezifizieren, im Allgemeinen kein explizites Modell des Antwortprozesses vorgelegt. Jede Funktion, die einerseits den in Definition 2.6.1 genannten Forderungen genügt und andererseits eine numerische Repräsentation $a \mapsto \theta_a$ liefert, die die Merkmalsausprägung für beliebige $a \in \mathcal{P}$ korrekt abbildet, ist eine zulässige Spezifikation der Itemfunktion F_g. Durch die Wahl der Gauß- bzw. der logistischen Funktion werden bestimmte Einschränkungen dieser allgemeinen Betrachtungen induziert.

4. Die im Folgenden präsentierten Itemfunktionen werden durch Verteilungsfunktionen von zufälligen Veränderlichen, sagen wir η, definiert, die aber dann aber nicht als Verteilungsfunktion, sondern als Funktion eines Parameters betrachtet werden. Ist η eine $N(\mu, \sigma^2)$-verteilte zufällige Veränder-

liche, so ist die *Verteilungsfunktion*

$$F(\eta_0|\mu,\sigma^2) = P(\eta \leq \eta_0|\mu,\sigma^2)$$

eine Funktion von η_0 für gegebene Parameterwerte von μ und σ^2, wobei für alle endlichen Werte von μ und σ^2 gilt, dass die Bedingung $0 \leq F(\eta_0|\mu,\sigma^2) \leq 1$ erfüllt ist. F kann nun zu einer Itemfunktion werden, wenn man den Wert η_0 fest hält und F etwa als Funktion von $\theta = \mu$ betrachtet. Die zufällige Variable η tritt bei diesen Betrachtungen dann nicht mehr auf. Die Itemfunktion gibt die Wahrscheinlichkeit an, dass $\eta \leq \eta_0$ ist für fixen Wert von η_0, wenn μ einen bestimmten Wert hat. η ist eine *unterliegende Variable* (underlying variable) (Bartholomew (1980, 1987)), auf die, wie oben schon angemerkt, im Rahmen der Testtheorie im Allgemeinen nicht explizit rekurriert wird; es läßt sich aber zeigen, dass die interessierenden Parameter der Itemfunktion, also der Personen- und der Itemparameter, über die Erwartungswerte bzw. Momente der unterliegenden Variablen definiert werden können. Bartholomew (1987), p. 103, zeigt, dass dieser Ansatz für binäre (dichotome) Items dem üblichen Ansatz, Itemfunktionen einzuführen, äquivalent ist. Auf die Implikationen dieser Äquivalenz wird weiter unten noch zurückgekommen.

Die folgende Betrachtung motiviert die in der Literatur (Lord & Novick (1967) gegebene Definition der Itemfunktion (s. unten).

Diesem Begriff der Itemfunktionen unterliegt die folgende Betrachtung: Es sei η die zufällige Veränderliche, die die bei einer Person Merkmalsausprägung zum Zeitpunkt der Messung repräsentiert. Um das Item I_g "positiv" beantworten zu können(die Aufgabe zu lösen, der Meinungsaussage zuzustiommen, etc), muß $\eta > \kappa_g$ sein, wobei κ_g ein für I_g kritischer Wert ist. η habe den für die Person charakteristischen Erwartungswert $\mathbb{E}(\eta) = \theta$, die Varianz von η sei σ. Die Wahrscheinlichkeit einer positiven Antwort ist dann $P(\eta > \kappa_g|\theta,\sigma)$, wobei jetzt der Einfachheit halber nur $P(\eta > \kappa_g)$ geschrieben werden soll. Man beachte, dass $P(\eta > \kappa_g)$ und nicht $P(\eta \leq \kappa_g)$ betrachtet wird: die Itemfunktion soll ja die Wahrscheinlichkeit angeben, mit der eine *positive* Antwort gegeben wird. Dann folgt

$$P(\eta > \kappa_g) = P(\eta - \theta > \kappa_g - \theta) = P\left(\frac{\eta-\theta}{\sigma} > \frac{\kappa_g-\theta}{\sigma}\right) = P\left(Z > \frac{\kappa_g-\theta}{\sigma}\right).$$

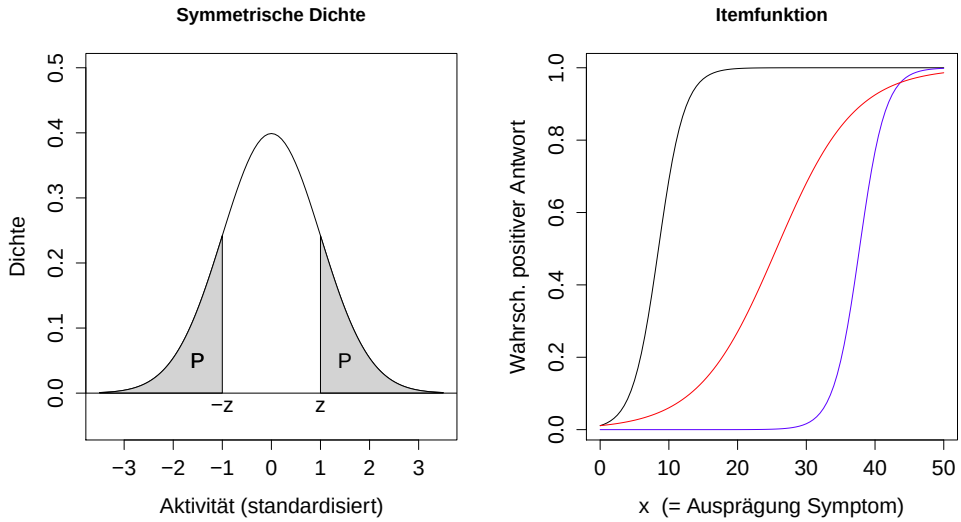
Z habe die Verteilungsfunktion F und die Dichte f. Dann ist

$$F((\kappa_g-\theta)/\sigma) = P\left(Z \leq \frac{\kappa_g-\theta}{\sigma}\right).$$

Nun werde angenommen, dass f symmetrisch ist. Dann gilt $P(Z > z_0) = P(Z \leq -z_0)$ (vergl. Abbildung 3). Dann folgt mit $z_0 = (\kappa_g - \theta)/\sigma$

$$P(Z > z_0) = P(Z \leq -z_0) = P\left(Z \leq \frac{\theta-\kappa_g}{\sigma}\right) = F((\theta-\kappa_g)/\sigma). \quad (2.6.4)$$

Abbildung 3: Wahrscheinlichkeiten bei symmetrischen Verteilungen, sowie Itemfunktionen



Obwohl also eigentlich $P(Z > z_0)$ betrachtet wird, kann man den Wert dieser Wahrscheinlichkeit über die Verteilungsfunktion an der Stelle $-z_0$ ausdrücken. Die schwarze und die blaue Itemfunktion in Abb. 3 unterscheiden sich nur hinsichtlich des Parameters κ_g, während sich die rote von der schwarzen hinsichtlich des Wertes von σ unterscheiden. Man beachte, dass die Itemfunktionen *Funktionen von θ sind.* □

Die Itemfunktionen, die in den Testtheorien eine zentrale Rolle spielen, sind das Ogiven-Modell sowie das logistische Modell:

1. **Das Ogiven-Modell:** Mit Φ sei die Verteilungsfunktion einer $N(0,1)$-verteilten zufälligen Veränderlichen gemeint, d.h es gelte $\Phi(z) = P(Z \leq z)$, $Z \sim N(0,1)$. Die Itemfunktion ist definiert durch

$$F_g(\theta) = \Phi(\alpha_g(\theta - b_g)) = \int_{-\infty}^{\alpha_g(\theta - b_g)} \varphi(z)\, dz, \quad \alpha_g > 0 \tag{2.6.5}$$

wobei $\varphi(z)$ die Dichte der Standardnormalverteilung ist, also

$$\varphi(z) = \exp(-z^2/2)/\sqrt{2\pi}.$$

$\alpha_g 1/\sigma_g > 0$ und b_g sind freie Parameter, auf deren Interpretation weiter unten noch eingegangen wird; im Allgemeinen ist $b_g = \kappa_g$.

Das Ogive-Modell ist im Rahmen der KTT konzipiert worden (Lord & Novick (1968), p. 366).

2. **Das logistische Modell:** Die Itemfunktion ist definiert durch

$$F_g(\theta) = \frac{e^{\alpha_g(\theta-\kappa_g)}}{1+e^{\alpha_g(\theta-b_g)}}, \quad \alpha_g > 0. \tag{2.6.6}$$

α_g und b_g sind wieder freie Parameter.

Das logistische Modell läßt sich wieder über den Ansatz $P(\eta > \kappa_g)$ motivieren, wobei dann nur die logistische Verteilung statt der Gauß-Verteilung betrachtet wird.

Das logistische Modell ist das Standardmodell der probabilistischen Testtheorien. Ein Spezialfall dieses Modells ist die Itemfunktion, wie sie im Rasch-Modell[11] angenommen wird:

$$F_g(\theta) = \frac{e^{\theta-\kappa_g}}{1+e^{\theta-\kappa_g}}, \tag{2.6.7}$$

Formal ergibt sich diese Funktion aus (2.6.6), indem man $\alpha_g = 1$ und $b_g = \kappa_g$ setzt. Eine alternative Interpretation bzw. Herleitung des Rasch-Modells wird in Abschnitt 2.6.3 gegeben.

Anmerkungen: Da Φ eine Verteilungsfunktion ist, nämlich die der standardisierten Gauß-Variablen, ist für das Ogive-Modell die Bedingung $0 \leq F_g(\theta) \leq 1$ sicherlich für alle θ erfüllt. Darüber hinaus gilt $\Phi(\alpha_g(\theta-\kappa_g)) \to 1$ für wachsenden Wert von θ, so dass im Ogive-Modell eine monotone Itemfunktion definiert wird. Eine analoge Ausage gilt für das logistische Modell, denn hinter dem Ausdruck auf der rechten Seite von (2.6.6) verbirgt sich die logistische Verteilungsfunktion, auf die in Abschnitt 2.6.3 noch genauer eingegangen wird. Auch hier wird eine monotone Itemfunktion definiert. Wichtig ist, dass in meßtheoretisch motivierten Testmodellen sowohl Φ als auch die logistische Funktion nur jeweils als Funktion insbesondere von θ eingeführt werden; dass beide Funktionen auch Verteilungsfunktionen irgendwelcher zufälliger Veränderlicher sind, spielt bei der Definition von F_g zunächst keine Rolle. Das Rasch-Modell erlaubt unter bestimmten Bedingungen, die Parameter $\theta = \theta_a$ einer Person $a \in \mathcal{P}$ und den itemspezifischen Parameter κ_g für ein Item I_g unabhängig voneinander zu bestimmen. Unter diesen Bedingungen sind die Schätzungen der Merkmalsparameter θ für verschiedene Personen *spezifisch objektiv*. Die spezifische Objektivität der Schätzungen begründet den besonderen Reiz des Rasch-Modells. □

Zur Interpretation der Parameter α_g und b_g: Sowohl im Ogive- wie auch im logistischen Modell definiert α_g die Trennschärfe eines Items I_g, und b_g die Schwierigkeit von I_g.

Um diese Interpretation einzusehen, betrachte man den Extremfall $\alpha_g = 0$. Dann ist $\alpha_g(\theta - b_g) = 0$ für alle θ, und da $\Phi(0) = 1/2$ folgt für das Ogive-Modell

[11] Georg Rasch (1901 - 1980), dänischer Statistiker

$F_g(\theta) = 1/2$ für alle θ. Es ist, als würde die Aufgabe durch Münzwurf gelöst, unabhängig von der Ausprägung des zu messenden Merkmals. Dies bedeutet, dass zwischen zwei Personen a und a' anhand der Beantwortung oder Lösung des Items I_g nicht getrennt werden kann, denn die Antwort ist ja unabhängig von den möglicherweise verschiedenen θ-Werten. Nun sei α_g groß, z.B. $\alpha_g = 100$. Für so große Werte von α_g nähert sich Φ einer Schrittfunktion an, d.h. es existiert ein θ_0 derart, dass $\Phi(\alpha_g(\theta - b_g)) \approx 0$ für $\theta < \theta_0$ und $\Phi(\alpha_g(\theta - b_g)) \approx 1$ für $\theta > \theta_0$. F_g trennt nun scharf zwischen allen Personen mit $\theta < \theta_0$ und $\theta > \theta_0$.

Die Betrachtung für die logistische Funktion ist vollkommen analog: für $\alpha_g = 0$ folgt

$$F_g(\theta) = \frac{e^0}{1 + e^0} = \frac{1}{2}$$

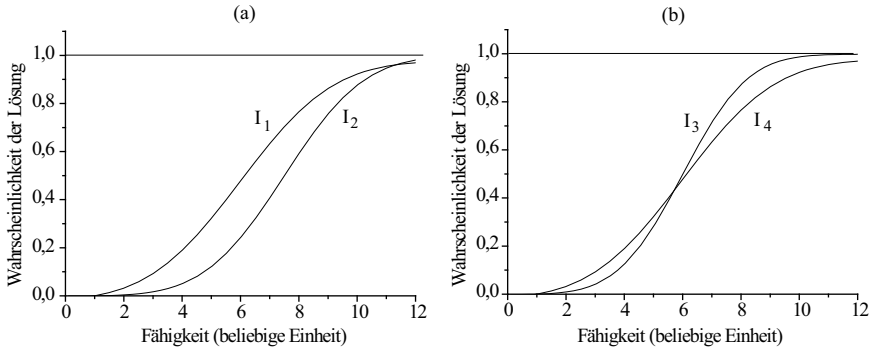
für alle θ. Dividiert man in (2.6.6) auf der rechten Seite Zähler und Nenner durch $\exp(\alpha_g(\theta - b_g))$, so ergibt sich der Ausdruck

$$F_g(\theta) = \frac{1}{1 + e^{-\alpha_g(\theta - b_g)}}.$$

Für $\theta - b_g > 0$ folgt $F_g(\theta) \to 1$, wenn $\alpha_g \to \infty$, und für $\theta - b_g < 0$ folgt $F_g(\theta) \to 0$, wenn $\alpha_g \to \infty$, d.h. I_g trennt für großen Wert von α_g scharf zwischen Personen mit einem Parameter $\theta < b_g$ und Personen mit einem Parameter $\theta > b_g$. Deswegen kann man sowohl im Ogive- wie auch im logistischen Modell α_g als Trennschärfeparameter interpretieren. Es muß aber angefügt werden, dass für große Werte von α_g die Trennschärfe zwar zunimmt, was die Unterscheidung zwischen Personen mit entweder $\theta < b_g$ oder $\theta > b_g$ angeht, dass dafür aber zwischen Personen mit $\theta < b_g$ immer weniger getrennt wird. Analog dazu wird immer weniger zwischen Personen, für die $\theta > b_g$ gilt, getrennt. Deshalb sind große α_g-Werte nur dann von Interesse, wenn es eben nur auf die Unterscheidung zwischen $\theta < b_g$ und $\theta > b_g$ ankommt. Will man möglichst gut zwischen allen Personen unterscheiden, sind mittlere Werte von α_g besser. Es sei angemerkt, dass für den Spezialfall des Rasch-Modells die Trennschärfe für alle Items I_g identisch ist, da $\alpha_g = 1$. Nur für den Fall $\alpha_g \neq 1$ können sich unterschiedliche Trennschärfen ergeben, da dann F_g von der Differenz $\alpha_g(\theta - b_g) = \alpha_g\theta - a_g b_g$ abhängt, d.h. wird dann durch den Faktor α_g gewissermaßen itemspezifisch skaliert. Abb. 4 zeigt Itemfunktionen für verschiedene α_g-Werte. Die Verschiedenheit der α_g-Werte impliziert, dass sich die Funktionen schneiden. Am Beispiel (b) wird dies besonders deutlich: Item I_3 ist für kleinere Werte von θ schwieriger ist als Item I_4, während für größere Werte von θ I_4 schwieriger ist als I_3. Das bedeutet, dass Personen mit einer geringeren Merkmalsausprägung das Item I_3 mit größerer Wahrscheinlichkeit beantworten als Personen mit einer größeren Merkmalsausprägung.

Der Parameter b_g bestimmt die Lokation der Itemfunktion, d.h. den Wertebereich von θ, innerhalb dessen F_g von 0 gegen 1 strebt. Deckt dieser Bereich eher kleine Werte von θ ab, so genügen eben kleinere Werte von θ, um die Aufgabe

Abbildung 4: Itemfunktionen für verschiedene Items mit verschiedenen α_g-Werten (Ogive- oder logistisches Modell)



zu lösen oder die Frage zu beantworten; die Aufgabe ist "leicht". Deckt der Wertebereich eher größere Werte von θ ab, so benötigt man eine größere Fähigkeit oder allgemein eine größere Ausprägung des gemessenen Merkmals, um das Item zu beantworten; das Item I_g ist "schwierig". Der Wert von b_g bestimmt also das Schwierigkeitsniveau des Items. Diese Betrachtung gilt für das Ogive- wie das logistische Modell gleichermaßen.

2.6.2 Lokale stochastische Unabhängigkeit

Der Begriff der lokalen oder bedingten stochastischen Unabhängigkeit ist von großer Wichtigkeit für die Möglichkeit, die Parameter θ und κ zu schätzen. Es wird zunächst versucht, die Bedeutung des Begriffs zu veranschaulichen.

Gegeben seien die drei Variablen X, Y und Z, und es gelte $X = b_{xz}Z + a_{xz} + e_x$, $Y = b_{yz}Z + y_{yz} + e_y$, wobei $b_{xz} \neq 0$, $b_{yz} \neq 0$ und $\mathbb{K}(e_x, e_y) = 0$ gelte. Dann wird $\mathbb{K}(X, Y) \neq 0$, aber

$$\mathbb{K}(X - b_{xz}Z, Y - b_{yz}Z) = \mathbb{K}(e_x, e_y) = 0$$

sein, d.h. für die partielle Korrelation wird $\rho_{xy.z} = 0$ gelten. Ebenso wird für die Menge $\{(X, Y)|Z = z\}$, d.h. für die Menge der X- und Y-Werte, für die zugehörigen Z-Werte gleich der konstanten z sind, $\rho(X, Y|Z = z) = 0$ gelten. Denn diese X- und Y-Werte unterscheiden sich ja nur durch die voneinander unabhängig variierenden e_x- und e_y-Werte. Man kann sagen, dass die X- und Y-Werte aus $\{(X, Y)|Z = z\}$ *bedingt* oder *lokal stochastisch unabhängig* sind.

Man kann diese Betrachtungen auch für dichotome Variablen durchführen.

Man betrachte etwa die Kontingenztabelle[12] für zwei dichotome Items I_g und $I_{g'}$:

	$I_{g'}$		
I_g	+	-	Σ
+	350	200	550
-	150	300	450
Σ	500	500	1000

(2.6.8)

Diese Häufigkeiten zeigen eine Assoziation zwischen den Items I_g und $I_{g'}$; der Vierfelderkorrelationskoeffizient beträgt $\phi_{gg'} = .307$. Dabei sei aber zwischen weiblichen und männlichen Probanden nicht unterschieden worden, man habe nur nachgesehen, ob ein Proband oder eine Probandin die beiden Items jeweils gelöst oder nicht gelöst haben. Man zählt nun die Daten für die Geschlechter (w - weiblich, m - männlich) aus und erhält die beiden Kontingenztabellen

w	$I_{g'}$		
I_g	+	-	Σ
+	320	80	400
-	80	20	100
Σ	400	100	500

m	$I_{g'}$		
I_g	+	-	Σ
+	30	120	150
-	70	280	350
Σ	100	400	500

(2.6.9)

In beiden Tabellen findet man nun keinerlei Assoziation mehr[13]. Die Vierfelderkorrelationen betragen $\phi_{gg'.w} = \phi_{gg'.m} = 0$. In der Tabelle (2.6.8) konnte das Geschlecht von einer Person zur nächsten variieren, aber in den beiden Tabellen in (2.6.9) war es jeweils konstant. Die Items I_g und $I_{g'}$ werden lokal oder bedingt stochastisch unabhängig voneinander beantwortet, die Assoziation zwischen I_g und $I_{g'}$ in der Tabelle (2.6.8) wurde durch das Geschlecht "erzeugt". Die beiden Tabellen in (2.6.9) unterscheiden sich in Bezug auf die *Grundquoten*, mit denen die Items gelöst werden: Für die w-Stichprobe lösen 400 das Item I_g, 100 lösen es nicht, und das Item $I_{g'}$ wird ebenfalls von 400 gelöst und 100 nicht gelöst. Das Verhältnis der Anzahl von "'Lösern" zu "Nichtlösern" ist die jeweilige Grundquote. Bei der m-Stichprobe sind sind die Grundquoten nachgerade reziprok zu denen der w-Stichprobe.

Es ist wichtig, zu sehen, dass die Korrelation zwischen den Antworten auf I_g und $I_{g'}$ in (2.6.8) bedeutet, dass für eine beliebig gewählte Person von einer Antwort auf I_g auf eine Antwort auf $I_{g'}$ geschlossen werden kann, – nur probabilistisch, nicht deterministisch, versteht sich. Bei den Teilpopulationen in den Tabellen (2.6.9) ist dies nicht mehr möglich, denn hier werden die Items ja unabhängig voneinander beantwortet. Allgemein läßt sich sagen, dass für die

[12] Die Tabellen sind Bartholomew (1987) entnommen.

[13] Bei Unabhängigkeit gilt, bis auf Stichprobenfehler, $n_{ij} = n_{i+}n_{+j}/N$, wobei n_{i+} die Randsumme für die i-te Zeile ist und n_{+j} die Randsumme für die $j$te Spalte; N ist die Gesamtzahl der Fälle. Man rechnet leicht nach, dass diese Bedingung für alle n_{ij} exakt erfüllt ist, nicht aber bei der Tabelle (2.6.8).

Teilpopulation von Probanden, die durch einen bestimmten Wert θ charakterisiert wird, die Beantwortung nicht nur zweier, sondern aller Items paarweise stochastisch unabhängig erfolgt. Die Konzeptualisierung des Begriffs der lokalen Unabhängigkeit kann verallgemeinert werden: angenommen, man betrachtet nur eine latente Variable, etwa θ_1, findet aber, dass für einen konstanten Wert von θ_1 die Items nicht stochastisch unabhängig voneinander beantwortet werden. Dann läßt sich vielleicht eine zweite latente Variable θ_2 finden derart, dass für konstante θ_1 und θ_2 die Items unabhängig voneinander beantwortet werden. Werden die Items nicht unabhängig voneinander beantwortet, läßt sich vielleicht eine dritte latente Variable θ_3 finden derart, dass für konstante θ_1, θ_2 und θ_3 stochastische Unabhängigkeit gefunden wird. Allgemein läßt sich sagen,

Definition 2.6.2 *Die Beantwortung der Items ist* lokal (stochastisch) unabhängig*, wenn die Verteilungen der Punktwerte (Scores) für die Items für alle Personen mit bestimmten Werten $\theta = (\theta_1, \ldots, \theta_r)$ unabhängig voneinander sind. (Lord & Novick (1968), p. 361)*

Der Begriff der lokalen Unabhängigkeit läßt sich formal spezifizieren. Für jedes Item I_g sei die Score-Variable X_g definiert, im dichotomen Fall ist $X_g = \{0, 1\}$. Man kann die gemeinsame Verteilung (Dichte) der X_g betrachten; sie sei f. Für den Fall, dass die X_g, $g = 1, \ldots, n$ stochastisch unabhängig voneinander sind, die Beziehung

$$f(X_1, \ldots, X_n) = f_1(X_1) f_2(X_2) \cdots f_n(X_n) = \prod_{g=1}^{n} f_g(X_g),$$

wobei f_g die Dichte- oer Wahrscheinlichkeitsfunktion von X_g ist[14]. Beim Begriff der lokalen Unabhängigkeit wird diese Wahrscheinlichkeitsverteilung als *bedingte Wahrscheinlichkeitsverteilung* aufgefasst;

$$f(X_1, \ldots, X_n|\theta) = \prod_{g=1}^{n} f_g(X_g|\theta), \quad \theta = (\theta_1, \ldots, \theta_r). \tag{2.6.10}$$

Für die Teilmenge der Items $I_2, \ldots, I_r$ erhält man dann den Ausdruck

$$f(X_2, \ldots, X_n|\theta) = \prod_{g=2}^{n} f_g(X_g|\theta), \quad \theta = (\theta_1, \ldots, \theta_r). \tag{2.6.11}$$

Man kann nun die Verteilung von X_1 unter der Bedingung θ und den Werten $X_2, \ldots, X_r$ betrachten. Nach dem Satz für bedingte Wahrscheinlichkeiten[15] gilt

[14] Diese Beziehung entspricht dem Produktsatz für unabhängige zufällige Ereignisse A und B: $p(A \cap B) = p(A)p(B)$. Für paarweise stochastische unabhängige Ereignisse $A_1, \ldots, A_n$ gilt

$$P(A_1 \cap \cdots \cap A_n) = p(A_1)p(A_2) \cdots p(A_n).$$

[15] $P(A|B) = p(A \cap B)/p(B)$.

dann

$$f(X_1|\theta; X_2, \ldots, X_r) = \frac{f(X_1, \ldots, X_n|\theta)}{f(X_2, \ldots, X_n|\theta)} = f_1(X_1|\theta). \tag{2.6.12}$$

Lokale Unabhängigkeit bedeutet also, dass die Verteilung von X_1 unabhängig von der Verteilung der $X_2, \ldots, X_n$ ist. Dies gilt natürlich für beliebiges X_g.

Man betrachte nun den Fall, dass für eine Teilmenge von Probanden, die alle identische $\theta_1, \ldots, \theta_r$-Werte haben (was im Fall kontinuierlicher θ_g-Werte sicher eine Idealisierung darstellt), für die aber die Beziehung (2.6.12) nicht gilt. Dann ist für diese Gruppe die Verteilung der X_1-Werte anders als in anderen Gruppen mit zwar den gleichen $\theta_1, \ldots, \theta_r$-Werten, aber anderen $X_2, \ldots, X_n$-Werten. Dies heißt aber, dass der durch die Koordinaten $\theta_1, \ldots, \theta_r$ definierte Raum nicht vollständig ist, denn für einen vollständigen Raum impliziert ein Satz von Koordinaten $\theta_1, \ldots, \theta_r$ lokale Unabhängigkeit. Man muß dann mindestens eine Dimension θ_{r+1} hinzunehmen, um insgesamt lokale Unabhängigkeit zu erhalten.

Eine sehr allgemeine Diskussion des Begriffs der latenten Variablen bis hinein in Fragestellungen der Physik findet man in Suppes & Zanotti (1981).

2.6.3 Latente Variablen

Die Zuordnung eines Meßwertes zu einer Person, $a \to \theta_a$, suggeriert, dass das gemessene Merkmal bei einer Person a in konstanter Ausprägung, die eben durch θ_a repräsentiert wird, vorhanden ist. Aber diese Vorstellung ist nicht zwingend. Psychologisch plausibler ist die Vorstellung, dass die Ausprägung psychischer Merkmale in der Zeit variiert. In Abschnitt 2.6.4 werden verschiedene Vorstellungen über die Art der Variation der Merkmale vorgestellt. Die allgemeinere Repräsentation der Ausprägung eines Merkmals wäre dann die als einer Funktion der Zeit, etwa $\eta(t)$. η kann insbesondere als eine *zufällige Funktion*[16] aufgefasst werden, einfach, weil der genaue Verlauf im Allgemeinen nicht vorausgesehen werden kann, da er von einer Reihe von Ereignissen abhängen kann, die die Person selbst nicht herbeigeführt hat, etwa ein Wetterumschwung, der bei einer Person Kopfschmerz hervoruft, der wiederum ihre Fähigkeit zum Kopfrechnen reduzieren kann, etc. Dieser Konzeptualisierung von Probanden mit zufällig variierender Merkmalsausprägung entspricht der von Holland (1990) explizierte Begriff des *stochastischen Subjekts*. Nimmt man der Einfachheit halber an, dass die Variation von η langsam und deswegen während der Dauer des Testens angenähert konstant ist, so kann man sagen, dass das gemessene Merkmal zum Zeitpunkt des Tests durch eine zufällige Veränderliche η repäsentiert werden kann, deren Erwartungswert $\mathbb{E}(\eta) = \theta$ und deren Varianz $\mathbb{V}(\eta) = \sigma^2$ ist. Der Personenparameter θ tritt also wieder als eine Konstante auf, bedeutet aber eben nicht mehr, dass das Merkmal selbst konstant ist. Das Ogive-Modell ebenso wie das logistische Modell können dann über die Verteilungsfunktionen von η hergeleitet werden.

[16] Formal entspricht η dann einer Trajektorie eines stochastischen Prozesses. Auf die Theorie stochastischer Prozesse muß hier aber nicht eingegangen werden.

Die Reaktion auf eine Aufgabe oder Frage kann sehr spezifisch für eine Person und eventuell noch für den Zeitpunkt sein. Es ist im Prinzip denkbar, dass ein Item eine personspezifische Merkmalskombination anspricht. Dem Begriff der Itemfunktion scheint implizit die Annahme zugrunde zu liegen, dass das gemessene Merkmal wohl aus einer Kombination von Merkmalen bestehen kann, diese Kombination aber für die verschiedenen Items konstant ist; die Itemschwierigkeit κ_g hängt ja nur von I_g ab, nicht auch noch von der Person a. Diese sicherlich vereinfachende Annahme wird im Folgenden unterstellt.

Annahmen:

1. Es sei $a \in \mathcal{P}$ eine beliebige Person. Für ein gegebenes Item I_g, $g = 1, \ldots, n$ existiert eine zufällige Veränderliche η_a, die die Ausprägung des gemessenen Merkmals bei dieser Person zum Zeitpunkt der Messung repräsentiert. Es sei $\mathbb{E}(\eta_a) = \theta_a$ der Erwartungswert von η_a, und die Varianz sei $\mathbb{V}(\eta_a) = \sigma_a^2$.

2. Für ein Item I_g existiert eine Konstante κ_g derart, dass für $\eta_a > \kappa_g$ die Person $a \in \mathcal{P}$ die Aufgabe löst. Für $\eta_a < \kappa_g$ löst die Person die Aufgabe I_g nicht.

Spezielle Modelle ergeben sich durch Wahl einer Verteilungsfunktion für η_a.

1. **Das Ogiven-Modell:** Es wird die Annahme gemacht, dass die Dichte von η_a durch
$$f(\eta_a) = \frac{1}{\sigma_a\sqrt{2\pi}} \exp\left[-\frac{(\eta_a - \theta_a)^2}{2\sigma_a^2}\right] \tag{2.6.13}$$
gegeben ist, d.h. es wird postuliert, dass η_a normalverteilt mit dem Erwartungswert θ_a und der Varianz σ_a^2 ist. Mit Φ werde die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Dann gilt
$$P(\eta_a > \kappa_g | \theta_a) = \Phi\left(\frac{\theta_a - \kappa_g}{\sigma_a}\right) = \Phi[\alpha_a(\theta_a - b_g)], \tag{2.6.14}$$
folgt, mit
$$\alpha_a = 1/\sigma_a, \quad b_g = \kappa_g. \tag{2.6.15}$$

   Beweis: Es ist
$$\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt. \tag{2.6.16}$$
   Wegen der bekannten Symmetrie der Normalverteilung gilt dann bekanntlich[17]
$$P(Z > z) = 1 - \Phi(z) = \Phi(-z). \tag{2.6.17}$$

[17] Für die Dichte von Z gilt $f_Z(z) = f_Z(-z)$, d.h. f_Z ist eine gerade Funktion. Es sei $V = -Z$; dann folgt $P(V \leq v) = P(-Z \leq v) = P(Z > -v) = 1 - P(Z \leq -v)$. Für die Dichte von V folgt dann $f_V(v) = f_Z(-v) = f_Z(v)$, und $P(V \leq v) = 1 - \Phi(-v)$. Andererseits hat man, wegen

Gesucht ist $P(\eta_a \leq \kappa_g|\theta_a)$, mit $\theta_a = \mathbb{E}(\eta_a)$, $\sigma_a^2 = \mathbb{V}(\eta_a)$. Standardisiert man η_a, so hat man $Z_a = (\eta_a - \theta_a)/\sigma_a$, d.h. $\eta_a = Z_a\sigma_a + \theta_a$. Dann folgt

$$P(\eta_a \leq \kappa_g) = P(Z_a\sigma_a + \theta_a \leq \kappa_g) = P\left(Z_a \leq \frac{\kappa_g - \theta_a}{\sigma_a}\right) = \Phi\left[\frac{\kappa_g - \theta_a}{\sigma_a}\right],$$

und damit wegen (2.6.17)

$$P(\eta_a > \kappa_g|\theta_a) = 1 - \Phi[(\kappa_g - \theta_a)/\sigma_a] = \Phi[(\theta_a - \kappa_g)/\sigma_a].$$

□

2. **Das logistische Modell:** Es wird die Annahme gemacht, dass die Verteilungsfunktion der η_a durch die logistische Verteilung gegeben ist, d.h. es soll gelten

$$P(\eta_a \leq \kappa_g|\theta_a) = \frac{1}{1 + \exp\left(-\frac{\kappa_g - \theta_a}{\sigma_a}\frac{\pi}{\sqrt{3}}\right)} \tag{2.6.18}$$

gegeben ist, wobei $\theta_a = \mathbb{E}(\eta_a)$ und $\sigma_a^2 = \mathbb{V}(\eta_a)$ ist. Eine Erläuterung zur Wahl dieser Verteilungsfunktion wird unten in den Anmerkungen gegeben. Für die Definition einer Itemfunktion muß die Wahrscheinlichkeit $P(\eta_a \leq \kappa_g)$ bestimmt werden. Man findet leicht

$$P(\eta_a > \kappa_g|\theta_a) = \frac{\exp\left(-\frac{\kappa_g - \theta_a}{\sigma_a^*}\right)}{1 + \exp\left(-\frac{\kappa_g - \theta_a}{\sigma_a^*}\right)} = \frac{\exp\left(\frac{\theta_a - \kappa_g}{\sigma_a^*}\right)}{1 + \exp\left(\frac{\theta_a - \kappa_g}{\sigma_a^*}\right)}, \tag{2.6.19}$$

mit der Reparametrisierung $\sigma_a^* = \sigma_a\sqrt{3}/\pi$.

Verteilungs- und Itemfunktionen: Als Verteilungsfunktionen sind (2.6.14) und (2.6.19) Funktionen von κ_g. Die Verteilungsfunktionen werden zu Itemfunktionen, wenn man κ_g konstant lässt und die Ausdrücke für die Verteilungsfunktion als Funktionen von θ_a bzw. von θ auffasst – der Index a muß fallengelassen werden, da ja θ_a ein spezieller, für die Person a charakteristischer Wert ist. Aus dem gleichen Grund können die individuellen Varianzen bzw. Streuungen σ_a nicht beibehalten werden.

Der Übergang von θ zu θ' entspricht ein Übergang von einer Person $a \in \mathcal{P}$ zu einer anderen Person $a' \in \mathcal{P}$, also von θ_a zu $\theta_{a'}$. Natürlich sind in einer Population $\mathcal{P}$ immer nur höchstens abzählbar viele Personen, während $F_g(\theta)$ als stetige Funktion von θ aufgefasst wird. Es ist also durchaus möglich, dass einem bestimmten Wert von θ keine Person a mit $\theta = \theta_a$ existiert. Dies invalidiert aber

---

$f_V = f_Z$,

$$P(V \leq v) = \int_{-\infty}^{v} f_V(u)du = \int_{-\infty}^{v} f_Z(u)du = \Phi(v).$$

Damit ergibt sich $1 - \Phi(-v) = \Phi(v)$ bzw. $\Phi(-v) = 1 - \Phi(v)$. Die Beziehung (2.6.17) gilt mithin nicht nur für die Normalverteilung, sondern für alle Verteilungen mit gerader Dichtefunktion.

die Definition von F_g als stetige Funktion von θ nicht, die Idee ist ja, dass jedem *möglichen* Wert von θ eine Wahrscheinlichkeit $F_g(\theta)$ zugeordnet werden soll.

Es sind nun zwei Ansätze denkbar.

1. Es wird $\sigma_a = \sigma_g$ für alle $a \in \mathcal{P}$ gesetzt, d.h. es wird postuliert, dass die Varianz bzw. Streuung von η_a nicht von der speziellen Person a abhängt, bzw. dass η bei allen Personen in gleicher Weise streut, aber in einer für das Item I_g charakteristischen Weise. Die Motivation für diese Interpretation besteht in der Möglichkeit, dass die Komplexität der Prozesse, die bei der Beantwortung eines Items ablaufen, für dieses Item spezifisch sein können; eine gößere Komplexität bedeutet eine größere Anzahl von Teilprozessen und damit eine größere Varianz im Antwortverhalten, was sich in einer größeren Varianz von η äußern könnte. Insbesondere könnte gelten, dass σ_g^2 proportional zur Schwierigkeit κ_g des Items I_g ist, so dass man etwa $\sigma_g^2 = c\kappa_g$ hätte, c eine Konstante. Ob eine solche Annahme psychologisch sinnvoll ist, muß im Einzelfall diskutiert werden. Jedenfalls erhält man für $\sigma_a = \sigma_g$ das Birnbaum-Modell und für $\sigma_a = \sigma$ identisch für alle $a \in \mathcal{P}$ und alle I_g das Rasch-Modell.

2. Es wird angenommen, dass die Itemfunktion eine Mischung von Funktionen über alle $a \in \mathcal{P}$ ist. Man könnte sie in der Form

$$F_g(\theta) = \sum_{a\in\mathcal{P}} \delta_a F_g(\theta), \quad \sum_a \delta_a = 1 \tag{2.6.20}$$

anschreiben; die Nebenbedingung $\sum_a \delta_a = 1$ sichert, dass $F_g(\theta_a)$ eine Wahrscheinlichkeit ist, also einen Wert zwischen 0 und 1 annimmt. Damit $F_g(\theta)$ berechenbar wird, müssen weitere Annahmen über (i) die σ_a und (ii) die δ_a gemacht werden. Es ist klar, dass hier diese Annahmen eine Beliebigkeit bezüglich der Definition von $F_g(\theta)$ implizieren, die den Begriff der Itemfunktion letztlich sinnlos machen.

Akzeptiert man den Ansatz, erhält man die Itemfunktionen

$$F_g^G(\theta) = \Phi(\alpha_g(\theta - b_g)), \qquad \text{Gauß-Verteilung} \tag{2.6.21}$$

$$F_g^L(\theta) = \frac{\exp(\alpha_g(\theta - b_g))}{1 + \exp(\alpha_g(\theta - b_g))}, \quad \text{logist. Verteilung} \tag{2.6.22}$$

wobei, analog zu (2.6.15), $\alpha_g = 1/\sigma_a^*$ und $b_g = \kappa_g$.

Es läßt sich zeigen, dass der messtheoretische Ansatz im Prinzip im Rahmen des hier vorgestellten Ansatzes eines stochastischen Subjekts, d.h. eines zufällig fluktuierenden Merkmals interpretiert werden kann. Der Beweis wurde von Bartholomew (1987) (p. 104) geliefert, vergl. Abschnitt A.3, Seite 285 in diesem Skript. Wird das Rasch-Modell gewählt, muß allerdings, um die Interpretation

eines stochastischen Subjekts beibehalten zu können, die unplausible Nebenbedingung $\sigma_{ag} = \sigma_g$ eingeführt werden. Der messtheoretische Ansatz wird deshalb überlicherweise auch nicht mit Bezug auf ein stochastisches Subjekt interpretiert. In Abschnitt 2.6.9 wird die Beziehung zwischen dem messtheoretischen Ansatz und dem des stochastischen Subjekts ausführlicher diskutiert,

Trennschärfe und Schwierigkeit: Nach (2.6.15) gilt $\alpha_g = 1/\sigma_g$ und $b_g = \kappa_g$. Der Trennschärfeindex α_g ist demnach der Reziprokwert der Streuung σ_g des gemessenen Merkmals. Die Itemfunktion F_g soll ja die Beziehung zwischen einem Wert θ und der Wahrscheinlichkeit der Beantwortung eines Items I_g, unabhängig von einer speziellen Person a angeben. Die Unplausiblität der Annahme $\sigma_g^2 =$ konstant für alle a läßt sich eventuell beheben, indem man σ_g als den Mittelwert der σ_{ga} ansieht.

Der Schwierigkeitsindex b_g ist gleich dem Schwellenwert κ_g, der die Mindestausprägung des Merkmals definiert, die zum Beantworten der Aufgabe nötig ist.

Zur logistischen Verteilung: Verteilungen vom Typ

$$P(X \leq x) = F(x) = \frac{1}{1 + \exp(-(x - \mu)/\sigma^*)}, \tag{2.6.23}$$

mit $\sigma^* = (\pi/\sqrt{3})\sigma$, wurden von dem belgischen Mathematiker Verhulst[18] 1836 zur Diskussion von Wachstumsprozessen, insbesondere des Wachstums der Bevölkerung von Paris, verwendet, um Fragen der Planung des Wohnungs- und Straßenbaus beantworten zu können (das Prädikat *logistisch* leitet sich vom französischen Wort für Wohnung, *logis*, her). In der Tat läßt sich zeigen, dass (2.6.23) aus bestimmten Annahmen über Wachstumsprozesse folgt. Es gilt

$$\frac{dF(x)}{dx} = f(x) = \alpha F(x)(1 - F(x)), \tag{2.6.24}$$

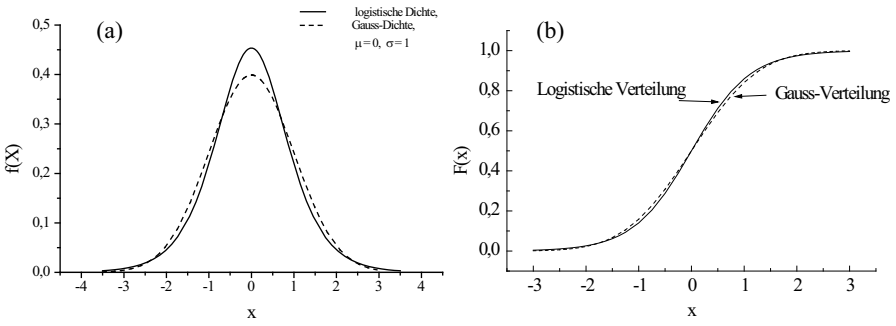
d.h. für kleine Werte von x, für die $F(x) \approx 0$, hat man

$$\frac{dF(x)}{dx} = \alpha F(x)(1 - F(x)) = \alpha F(x) - \alpha F^2(x) \approx \alpha F(x),$$

weil $F^2(x)$ wiederum kleiner als $F(x)$ und für hinreichend kleine F deshalb vernachlässigbar ist. Da dF/dx die Veränderung von F mit x ist, sieht man, dass für hinreichend kleine Werte von x die Funktion F in guter Näherung exponentiell wächst, denn für $F(x) = xe^{\alpha x}$ ist $dF(x)/dx = \alpha c e^{\alpha x} = \alpha F(x)$. Die Beschleunigung des Wachstums wird dann geringer, bis $F(x) = 1/2$, um dann für $F(x) \to 1$ wieder gegen Null zu streben. Eine zunächst wachsende und dann kleiner werdende Beschleunigung ist typisch für Wachstumsprozesse bei beschränkten Ressourcen. Wie weit ein derartiges Wachstumsmodell auf psychologische Prozesse, die beim Lösen von Aufgaben etwa in einem Intelligenztest ablaufen, übertragen

[18] Pierre-François Verhulst (1804 – 1849)

Abbildung 5: Vergleich der logistischen Verteilung und der Normalverteilung;(a) Dichtefunktionen, (b) Verteilungsfunktionen. Die maximaleDifferenz zwischen logistischer und Gaußscher Verteilungsfunktion beträgt bei geeigneter Parametrisierung weniger als .001, d.h. $\max_x |\Phi(x) - \Psi(1.7x)| < .01$ (Haley, 1952). vergl. auch Lord & Novick (1968, p. 399



werden kann, ist eine Frage, die noch nicht diskutiert worden zu sein scheint. Ein Grund dafür ist vermutlich, dass bei der Konstruktion von Testmodellen im Allgemeinen keine Prozessmodelle betrachtet werden und Itemfunktionen eher nach messtheoretischen Gesichtspunkten gewählt werden. In Abschnitt 5 wird darauf näher eingegangen. Die Beziehung (2.6.24) bedeutet u. a., dass die Dichtefunktion wie die Gauß-Dichte glockenförmig ist. Diese Eigenschaften spielen aber im hier gegebenen Zusammenhang keine Rolle, es sind vielmehr einerseits die Ähnlichkeit mit der Gauß-Verteilung, deren Rolle insbesondere in der KTT durch den zentralen Grenzwertsatz motiviert ist, dann die größere mathematische Einfachheit (die Verteilungsfunktion einer Gaußschen Variablen ist, im Unterschied zur logistischen Verteilung, nicht in geschlossener Form darstellbar), sowie eine Reihe weiterer Eigenschaften, die bei der Diskussion des Rasch-Modells erwähnt werden.

□

Obwohl die Interpretation der Itemfunktionen über die Verteilungsfunktionen zufälliger Veränderlicher plausible Interpretationen für die Itemfunktionen implizieren, ist die Annahme gleicher Varianzen $\mathbb{V}(\eta_{ga}) = \sigma_g^2$ sicherlich keine besonders plausible Randbedingung, und die noch stärkere Einschränkung $\mathbb{V}(\eta_{ga}) = \sigma^2$, die für das Rasch-Modell gelten muß, ist noch weniger plausibel. Man hat zwei Möglichkeiten:

1. Man wählt den meßtheoretischen Ansatz mit der Möglichkeit, den Ausdruck für die Wahrscheinlichkeit einer korrekten oder "positiven" Antwort in Bezug auf eine Wahrscheinlichkeitsverteilung zu interpretieren. Dies bedeutet, eine möglichst einfache Repräsentation ohne irgendeine Modellierung des

Antwortprozesses zu finden. Die Parameter θ_a für getestete Personen $a \in \mathcal{P}$ werden als Meßwerte für das gemessene Merkmal interpretiert, und fluktuierende Merkmalsausprägungen werden nicht weiter in Betracht gezogen. Die Präferenz für entweder das Ogive-Modell oder das logistische Modell wird über die Eigenschaften der Schätzungen für die θ_a und κ_g begründet.

2. Man sucht nach anderen Modellen für das Antwortverhalten, die auf die Itemfunktionen führen. Es müßte sich dabei um Modelle handeln, die nicht einfach nur zufällig variierende Merkmalsausprägungen annehmen, die sich durch zufällige Veränderliche repräsentieren lassen. Es ist nicht notwendig, dass diese Modelle Gaußsche oder logistische Verteilungsfunktionen implizieren, um Itemfunktionen wie die hier betrachteten zu erhalten. Wichtig ist dabei, dass die Anzahl der freien, aus den Daten zu schätzenden Parameter möglichst nicht die übersteigt, die in den hier behandelten Modellen auftreten.

Diese Alternativen repräsentieren ein Dilemma. Der meßtheoretische Ansatz definiert nur einen funktionalen Zusammenhang zwischen einem Satz von Parametern und der Wahrscheinlichkeit einer "positiven" Antwort, ohne dass die Struktur der psychologischen Mechanismen, die den Antwortprozess ausmachen, näher spezifiziert wird. Explizite Modelle für das jeweilige Antwortverhalten lassen sich andererseits kaum anhand einer Reihe von Testfragen überprüfen. In Abschnitt 5, Seite 271, wird der Anspruch einiger Theoretiker, dass Messmodellen ein erkenntnistheoretisches Primat zukommt, kritisch diskutiert.

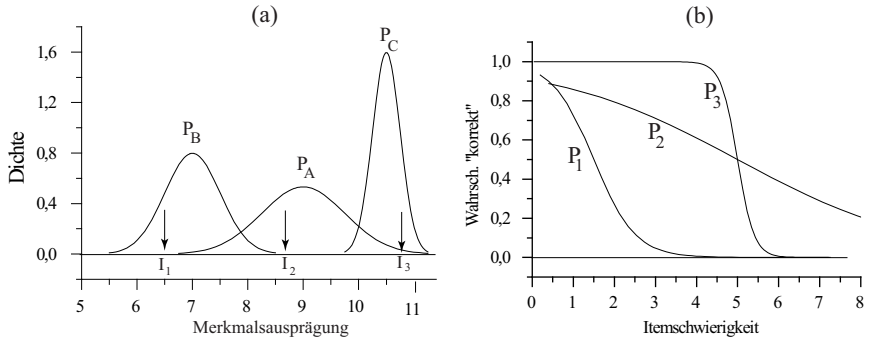
2.6.4 *Person Reliability* und *Person characteristic curves*

Der Ansatz der "unterliegenden Variablen" ist nicht Standard in der Testtheorie, sowohl die KTT wie auch die IRT-Modelle gehen im Allgemeinen von der Annahme aus, dass die Zufälligkeiten in den Testwerten ein Resultat der zufälligen Auswahl der Probanden aus der Population möglicher Probanden sind; in Abschnitt 2.6.9 wird diese Interpretation weiter diskutiert. An dieser Stelle soll der Begriff der *person reliability* eingeführt werden; gemeint ist die Stabilität des Testverhaltens einer Person. Bei Testwiederholungen fällt eben auf, dass Personen sich nicht identisch verhalten, d.h. nicht exakt das gleiche Leistungsniveau oder die gleiche Ausprägung einer Meinung zeigen.

Lumsden (1977) hat drei Arten von Variabilität in einer Person unterschieden:

1. **Entwicklungsmäßiger Trend** (*developmental trend*): Diese Trends äußern sich in individuellen Wachstumskurven bzw. Verfallskurven (für das Abnehmen von Leistungsfähigkeit oder das Abklingen radikaler Ansichten. Reifungsprozesse gehören zu dieser Klasse von Veränderungen. Diese Veränderungen können längere Zeit dauern.

Abbildung 6: (a) Attributkontinuum für drei Personen P_A, P_B und P_C sowie drei Items I_1, I_2 und I_3, (b) Personencharakteristika für drei weitere Personen



2. **Swells** Mit "Swell" bezeichnet Lumsden Schwankungen um das durcschnittliche Leistungs- oder Meinungsniveau, die die Dauer von Stunden oder Tagen haben. Sie können mit emotionalen Zuständen zu tun haben, oder mit nicht permanenten situativen Faktoren. Auch Tagesschwankungen zählen zu den "wells".

3. **Tremors** Mit *Tremors* bezeichnet Lumsden schnelle Fluktuationen der Leistungsfähigkeit. Tremors können sich Swells und Trends überlagern.

Die Frage ist, wie sich diese drei Variabilitätskomponenten in der Reliabilität äußern. Lumsden argumentiert, dass Trends hier kaum eine Rolle spielen dürften, Swells jedoch sehr wohl. Am wichtigsten seien aber die Tremoreffekte. Die Abbildung 6 illustriert ein mögliches Modell (Lumsden, 1977). Ein Item korrespondiert zu einer Position auf dem Kontinuum; hier werden drei betrachtet,I_1, I_2 und I_3. Zu einer Person korrespondieren jeweils ganze Verteilungen möglicher Attributswerte, gezeigt werden drei Verteilungen, die den Personen P_A,P_B und P_C entsprechen. Die Verteilungen charakterisieren die Schwankungen durch Tremor. Lumsden nimmt insbesondere an, dass die Verteilungen Gauß-Dichten entsprechen. Die Varianz-Parameter der Dichten dürfen verschieden sein. Das Modell entspricht dem Thurstone-Modell, das in Abbildung 10 illustriert wird. Der Unterschied besteht darin, dass hier die *discriminal dispersions* alle gleich Null sind, und die Kategorien für die Scores wurden durch Personenlokationen ersetzt. Eine Person beantwortet oder löst ein Item, wenn die momentane Lokation höher als die Lokation des Items ist. Dementsprechend ist es möglich, dass eine Person z.B. eine Aufgabe an einer bestimmten Itemlokation löst, ein anderes an der gleichen Lokation aber nicht, – weil wegen des Tremors die momentane Ausprägung der Leistungsfähigkeit nicht hinreicht. Lumsdens Modell ist also ein spezielles Modell für unterliegende Variablen.

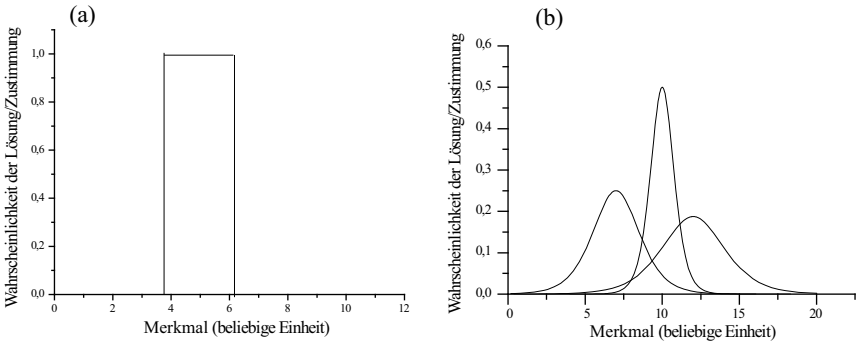
Entsprechend der Itemfunktion kann nur eine *person charakteristic function*, oder einfach eine Personenfunktion, eingeführt werden. Die Personenfunktion gibt, für eine gegebene Person, die Wahrscheinlichkeit einer korrekten Antwort als Funktion der Schwierigkeit eines Item an; die x-Achse ist also eine κ-Achse. Lumsden weist darauf hin, dass diese Funktion wohl zuerst von Weiss (1973) explizit diskutiert wurde. Abbildung 6 (b) zeigt drei Personenfunktionen für die Personen P_1, P_2 und P_3 (die von den Personen in (a) verschieden sind). Die Personen P_2 und P_3 unterscheiden sich nicht in Bezug auf die Position ihres mittleren Wertes auf der Merkmalsausprägung, aber in Bezug auf die Varian ihres Tremors. P_3 löst einen großen Teil der leichteren Aufgabe mit großer Zuverlässigkeit, d.h. die Wahrscheinlichkeit,sie zu beantwortgen, liegt bei 1, und Aufgaben mit einer Schwierigkeit größer als ein bestimmter Wert mit Wahrscheinlichkeit Null. Person P_2 löst auch leichtere Aufgaben gelegentlich nicht, dafür gelegentlich aber auch schwierige Items. Die Personenreliabilität von P_3 ist größer als die von P_1. Eine nach Lumsden auf Spearman (1927) zurückgehende Interpretation ist, dass Personen wie P_2 gut sind als Designer von Flugzeugen, aber katastrophal sind als Flugzeugpiloten.

Es ist keine Frage, dass Unterschiede in der *person reliability* existieren. Sowohl in der KTT wie auch in den probabilistischen Modellen zielt die Interpretation aber nur auf den wahren Wert τ_a bzw. auf den Personenparameter θ_a einer Person $a \in \mathcal{P}$. Ferrando (2004) und (2007) hat eine Verallgemeinerung dieser Ansätze vorgeschlagen.

2.6.5 Nicht-monotone Itemfunktionen

Während für Items, die eine Leistung oder Fähigkeit testen, eine monotone Itemfunktion vermutet werden kann, sind nicht-monotone Itemfunktionen für andere Persönlichkeitsmerkmale oder Einstellungen durchaus denkbar. Man nehme etwa eine Frage, die sich auf die Selbsteinschätzung der psychischen Belastbarkeit bezieht. Repräsentiert η die Wahrnehmung bezüglich der eigenen Belastbarkeit, so würde etwa $\eta = 0$ bedeuten, dass sich eine Person für gar nicht belastbar hält, und für $\theta > \theta_{\max}$ für "sehr belastbar". Die meisten Menschen werden sich bei einem Wert von η zwischen 0 und $\theta_{\max}$ verorten. Nimmt eine Person a sich mit einem Erwartungswert $\theta_{ao} = \mathbb{E}(\theta_a)$ mit $0 \leq \eta_{ao} < \theta_{\max}$ wahr, so wird ihr η-Wert zu einem Befragungszeitpunkt irgendwo in der Nachbarschaft von θ_{ao} liegen. Die Verteilung der Werte von θ_a um den Wert θ_{ao} ist dann eine nicht-monotone Itemfunktion. Bei monotonen Itemfunktionen war die Schwierigkeit eines Items über einen Parameter, der die Lokation der Itemfunktion bestimmt, definiert worden. Eine analoge Definition von Schwierigkeit könnte auch hier vorgenommen werden, indem man einen Lokationsparameter (Modus, Median, Erwartungswert) als Schwierigkeitsparamter definiert. Wie sinnvoll eine solche Definition im Falle nicht-monotoner Itemfunktionen ist, muß dann im jeweiligen Fall entschieden werden.

Abbildung 7: Itemfunktionen für nichtmonotome Items. (a) Guttman-Item, (b) der allgemeine Fall.



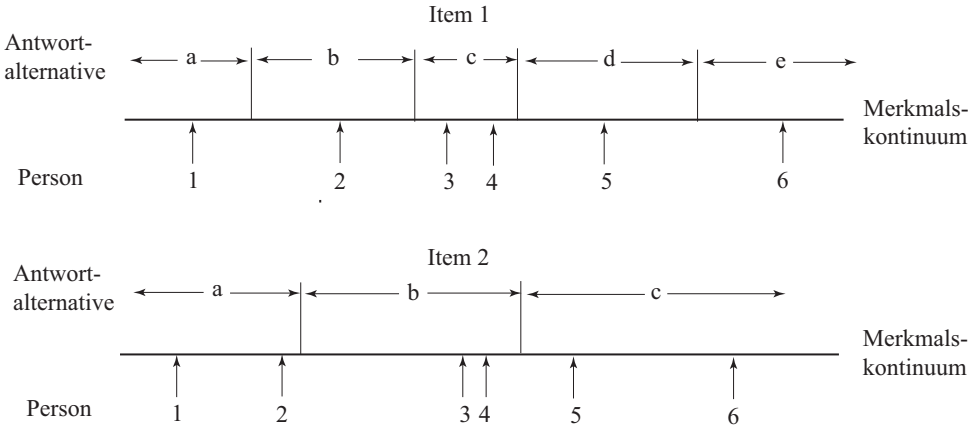
Die Trennschärfe eines Items bei nicht-monotoner Itemfunktion läßt sich über Verteilungsparameter wie die Streuung oder Varianz definieren; generell kann man sagen, dass das Item um so trennschärfer ist, je mehr die Itemfunktion einer Dirac-δ-Funktion ähnelt. Denn dann werden nur die Personen eine Antwort auf das Item liefern, deren θ-Wert in einem sehr schmalen Bereich liegen.

Insbesondere bei Items, mit denen Einstellungen erhoben werden sollen, müssen die Itemfunktionen nicht notwendig homogen sein. Will man z. B. etwas über die Einschätzung der Schwere von Verbrechen in einer Population erfahren, könnte man Item der Art: "Wer eine Million Euro Steuern hinterzieht

1. " sollte mit mindestens 2 Jahren Gefängnis bestraft werden".
2. " sollte mit 2 Jahren Gefängnis bestraft werden".
3. " sollte mit maximal zwei Jahren Gefängnis bestraft werden."

Dem Item 1. können alle zustimmen, die der Ansicht sind, dass Steuerhinterziehung in dieser Höhe mit 2 oder 5 oder 7 etc Jahren bestraft werden sollte, dem Item 3. stimmen alle zu, die mehr als Null Jahre, aber eben weniger als 2 Jahre für angemessen halten. Dem Item 2. stimmen aber eben nur diejenigen zu, die exakt zwei Jahre für die richtige Bestrafung halten. Man spricht deswegen auch von einem Punkt-Item. Eine Verallgemeinerung ergibt sich, wenn man statt einer exakten Jahreszahl ein Intervall vorgibt, etwa 2.5 bis 2.5 Jahre. Abbildung 7, (a) deckt diesen Fall ab, wenn auch in deterministischer Weise. Der Fall (b) beschreibt den Fall von Punkt-Items, für die aber angenommen wird, dass die Einstelllung um einen Punkt auf der Skala fluktuiert. So kann eine Person im Durchschnitt 2 Jahre für angemessen halten, aber ja nach dem Zustand, in dem sie sich bei der Befragung befindet, eine geringere oder eine höhere Strafe für angemessen halten.

Abbildung 8: Aufteilung einer Skala in Bereiche für ein gegebenes Item. Item 1: Person 1 stimmt der Antwortalternative a zu, Person 2 stimmt der Alternative b zu, Personen 3 und 4 stimmen c zu, Person 5 stimmt der Alternative d zu und Person 6 der Alternative e. Für Item 2 ist die Betrachtung analog.



Die im folgenden Abschnitt beschriebene Guttman-Skalierung, auch als Skalogrammanalyse bekannt, behandelt u.A. einen Spezialfall für nichthomogne Itemfunktionen.

2.6.6 Guttman-Skalen

Guttman (1950) schlug eine Methode vor, die es im Prinzip gestattet, aus qualitativen Items zumindest eine Ordinalskala zu gewinnen. Die Annahme ist, dass das von den Items erfasste Merkmal eindimensional ist, und sein Verfahren erlaubt einen Test dieser Forderung. Guttman geht dabei von deterministischen Itemfunktionen aus; Abb. 7 illustriert Itemfunktionen (A) für ein Punkt-Item, (B) für ein dichotomes Item. Ein Punkt-Item charakterisiert eine bestimmte Position auf dem Merkmalskontinuum, während ein dichotomes Item das Kontinuum in zwei Bereiche teilt. Ein Beispiel für ein Punkt-Item ist: "Eine Vergewaltigung sollte mit 5 Jahren Gefängnis bestraft werden". Personen, die eine geringere oder eine höhere Strafe für angemessen halten, werden dem Item nicht zustimmen, und nur Personen, die genau dieses Strafmaß für adäquat halten, werden zustimmen. Ein dichtomes Item wäre "Eine Vergewaltigung sollte mit mindestens 5 Jahren Gefängnis bestraft werden." Personen, die ein geringeres Strafmaß als 5 jahre Gefängnis für hinreichend halten, werden dem Item nicht zustimmen, und alle, die fünf oder mehr Jahre für angemessen halten, werden zustimmen. Eine Variante des Punkt-Items ist das Mehrkategorienitem. Ein Beispiel wäre:

Eine Vergewaltigung sollte bestraft werden mit

(a) höchstens einem Jahr Gefängnis.

(b) einem bis zu drei Jahren Gefängnis.

(c) mindestens drei, maximal mit fünf Jahren Gefängnis.

(d) mindestens fünf Jahren Gefängnis.

Jeder der Antwortalternativen (a) bis (d) entspricht die *trace line* eines Punkt-Items. Verschiedene Items müssen nicht gleich viele Antwortalternativen haben, und die Grenzen der Alternativen auf dem latenten Merkmal können von Item zu Item verschieden sein. Das Verfahren kann am einfachsten illustriert werden, wenn man Items wählt, bei denen die jeweiligen Grenzen von vornherein bekannt sind (Torgerson (1962), p. 307):

Item 1: Welche der folgenden Aussagen über Ihre Körpergröße ist wahr?

(a) Ich bin größer als 1.83 m
(b) Ich bin zwischen 1.83 m und 1.67 m
(c) Ich bin zwischen 1.67 m und 1.52 m
(d) Ich bin kleiner als 1.52 m

Item 2: Sind Sie

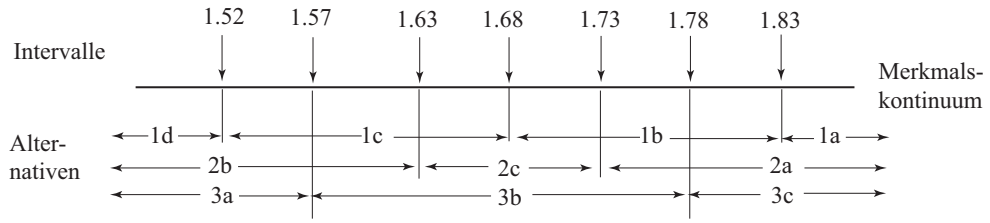
(a) Über 1.73 m groß?
(b) Unter 1.63 m groß?
(c) Weder größer als 1.73 m noch kleiner als 1.63 m?

Item 3: Ich glaube, dass meine Größe

(a) kleiner als 1.57 m ist.
(b) zwischen 1.57 m und 1.78 m ist.
(c) größer als 1.78 m ist.

Es gibt 7 Größenangaben und damit insgesamt 8 Intervalle, in die die Größe einer Person fallen kann: ein Intervall für alle, die kleiner/gleich der kleinsten Größe sind, ein Intervall für alle, die größer/gleich der größten Angabe sind, und 6 Intervalle zwischen den Größen. Es wird angenommen, dass befragte Personen ihre wahre Größe kennen, sorgfältig antworten und nicht lügen. Wird einer Alternative zugestimmt, wird dies durch eine 1 dokumentiert, andernfalls mit einer 0. Zunächst kann man sich einen Überblick über die Gesamtzahl möglicher Antwortmuster verschaffen. Im Prinzip kann jede Alternative eines Items mit jeder Alternative eines anderen kombiniert werden; dementsprechend gibt es $4 \times 3 \times 3 = 36$ mögliche Antwortmuster. Antworten die Personen in der beschriebenen Weise, kann aber nur eine Teilmenge davon tatsächlich beobachtet werden. Man überlegt sich (leicht), dass bei konsistentem Antwortverhalten nur 8 Antwortmuster auftreten können. Welchen Alternativen eine befragte Person an-

Abbildung 9: Intervalle und latentes Merkmal



kreuzt, wird durch die eigene Position auf der Skala für Körpergrößen bestimmt. Diese Position liegt in einem der Intervalle von kleiner als 1.52 m bis größer als 1.83. Eine Person, die größer als 1.83 m ist, wird dann die Alternativen 1(a), 2(a) und 3(c) ankreuzen. Eine Person, deren Körpergröße in das Intervall 1.78 m bis 1.83 m fällt, wird die Alternativen 1(b), 2(a) und 3(c) ankreuzen, etc. Insgesamt hat man die folgenden Muster:

Tabelle 1: Antwortmuster

Intervall	Antwortmuster			Typ
< 1.52	1d	2b	3a	H
$1.52 \leq x < 1.57$	1c	2b	3a	G
$1.57 \leq x < 1.63$	1c	2b	3a	F
$1.63 \leq x < 1.67$	1c	2b	3b	E
$1.67 \leq x < 1.73$	1b	2c	3b	D
$1.73 \leq x < 1.78$	1b	2a	3b	C
$1.78 \leq x < 1.83$	1b	2a	3c	B
$1.83 \leq x$	1a	2a	3c	A

Es gibt also insgesamt 8 Antwortmuster. Nun werden im Allgemeinen die Kategoriengrenzen (1.52 m, 1.57 m , 1.63 m etc) nciht explizit vorgegeben, man sucht eine zumindest ordinale Ordnung zwischen ihnen. Diese läßt sich aus den Antwortmustern, wie sie in Tablle 1 angeben werden, herleiten. Ein Antwortmuster besteht aus der Folge der Alternativen, die jeweils in den Items angekreuzt werden, also für den Antworttyp A (a) bei Item 1, (a) bei Item 2 und (c) bei Item 3. Durch die Typen A bis H werden die Körpergrößen von "groß" bis "klein" geordnet. Personen vom Typ A sind größer als die vom Typ B, etc. Man kann die Typen erneut in einer Tabelle, dem *Skalogramm* anschreiben, wobei aber die Reihenfolge der Itembeantwortung so geändert wird, dass ein Parallelogramm von Einsen (1 steht für "angekreuzt", 0 für "nicht angekreuzt") entsteht. Es entsteht die Tabelle 2. Wie man der Tabelle entnimmt, wird der Typ A nun durch die Folge (1a, 3c, 2a) statt, wie in Tabelle 1, durch (1a, 2a, 3c) repräsentiert. Der

Tabelle 2: Skalogramm für drei Items mit mehreren Antwortalternativen

	Itemkategorien									
Typ	1a	3c	2a	1b	2c	3b	1c	2b	3a	1d
A	1	1	1	0	0	0	0	0	0	0
B	0	1	1	1	0	0	0	0	0	0
C	0	0	1	1	0	1	0	0	0	0
D	0	0	0	1	1	1	0	0	0	0
E	0	0	0	0	1	1	1	0	0	0
F	0	0	0	0	0	1	1	1	0	0
G	0	0	0	0	0	0	1	1	1	0
H	0	0	0	0	0	0	0	1	1	1

Typ B wird durch die Folge (3c, 2a, 1b) repräsentiert, in Tabelle 1, durch die Folge (1b, 2a, 3c). Die Folge der Itemkategorien, wie sie in der Tabelle 2 angegeben wird, also (1a, 3c, 2a, 1b, etc) entspricht in Abbildung 10 der Folge der Antwortkategorien, wenn man von rechts nach links geht. Man hat eine ordinale Ordnung der Intervalle erhalten.

Ein interessanter Aspekt des Skalogramms ist, dass nicht nur Information über die Ordnung der Items geliefert wird, sondern der Ort einer urteilenden Person angedeutet wird. Denn eine Person gehört notwendig zu einem der verschiedenen Typen, und ein Typ korrespondiert zu einem Bereich auf der Skala. Kennt man den Typ, zu dem eine Person gehört, kann man die Antworten der Person vorhersagen.

Dichotome Items: Es werden dichotome Items formuliert, die sich hinsichtlich der Ausprägung des Merkmales (der Einstellung), die sie ausdrücken, rangordnen lassen. In Abb. 2 (A) entsprechen den Itemfunktionen drei Items, von denen (a) das Merkmal am geringsten ausdrückt, und (c) drückt das Merkmal am meisten aus. Allgemein gelte für n Items

$$I_1 \prec I_2 \prec \cdots \prec I_n, \tag{2.6.25}$$

Wobei I_1 die geringste Merkmalsausprägung, I_n die – für die gegebene Menge von n Items – höchste Merkmalsausprägung repräsentiere. Eine Person kann nun jedem Item zustimmen oder auch nicht. Als Beispiel werden die folgenden vier Items betrachtet:

1. Sind Sie größer als 1.83 m?
2. Sind Sie größer als 1.73 m?
3. Sind Sie größer als 1.63 m?
4. Sind Sie größer als 1.52 m?

Hier ist es sicherlich so, dass eine Person, die dem Item 1. zustimmt, sicher auch

allen anderen zustimmt; diese Personen entsprechen dem Antworttyp A mit dem Muster (1, 1, 1, 1), wobei eine 1 für angekreuzt, eine Null für nicht angekreuzt steht. Der Antworttyp B ist größer als 1.73 m, aber kleiner als 1.83 m, man wird das Antwortmuster (0, 1, 1, 1) erhalten. Für den Typ C gilt, dass er größer als 1.63 m, aber kleiner als 1.73 m ist, etc. Man erhält die folgende Darstellung: Die Tabelle 3 ist redundant, weil sie sowohl die angekreuzten Antworten wie die

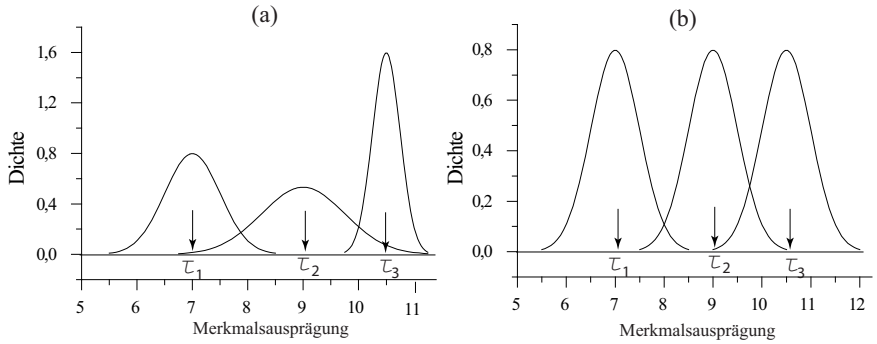
Tabelle 3: Skalogramm für dichotome Items

	Item								
	Zustimmung				Ablehnung				
Typ	1	2	3	4	1	2	3	4	Anz. Personen
A	1	1	1	1	0	0	0	0	n_1
B	0	1	1	1	1	0	0	0	n_2
C	0	0	1	1	1	1	0	0	n_3
D	0	0	0	1	1	1	1	0	n_4
E	0	0	0	0	1	1	1	1	n_5

nichtangekreuzten Antworten enthält; der Sinn dieser Darstellung ist, zu zeigen, dass wieder ein Parallelogramm von Einsen entsteht. Alle Information in den Daten ist natürlich im ersten. linken Teil der Tabelle (Zustimmung) enthalten. Ein solches Dreieck (bzw. Parallelogramm) von Einsen entsteht natürlich nur, wenn das Merkmal eindimensional ist und die Personen fehlerfrei antworten. Die Eindimensionalität ist bei den gewählten Items natürlich trivial, aber diese Items dienen nur der Illustration. In der Realität werden die Kategoriengrenzen (1.52, 1.63, etc) nicht bekannt sein.

Die Guttman-Skala beruht auf einem determinstischen Ansatz, dh es sind werden keinerlei probabilistischen Mechanismen postuliert, die Abweichungen von korrekten Wahlen der Alternativen, die der Position der urteilenden Person entsprechen, erklären. Da die Grenzen der Intervalle, die durch die Alternativen implizit definiert werden, i. A. nicht explizit bekannt sind, kann es zu Fehlurteilen kommen, zumal wenn Kategoriengrenzen dicht beieinander liegen. Man muß dann entscheiden, ob Antwortmuster, die nicht einer 1-dimensionalen Skala entsprechen, einfach nur Fehler sind oder aber eine Mehrdimensionalität des Merkmals anzeigen. Dafür hat Guttman einen Koeffizienten vorgeschlagen, den *Reproduzierbarkeitskoeffizienten*. Der Ausdruck beruht darauf, dass "fehlerhafte" Antwortmuster in Beziehung gesetzt werden können zu dem Muster, das dem beobachteten am ähnlichsten ist. Die Person wird dann dem entsprechenden Typ zugeordnet, und der Fehler ist ein Fehler, der von Personen dieses Typs gemacht werden kann. Personen dieses Typs ist es eigentlich nicht möglich – wenn die zufälligen Effekte bei der Urteilsbildung nicht wären – das beobachtete Muster zu reproduzieren, deswegen ist die Rede von einem *Reproduzierbarkeitsfehler*. Eine Person, die das Antwortmuster (1, 1, 0, 1) erzeugt hat, ist höchstwahrscheinlich

Abbildung 10: Discriminal Dispersions (Thurstone, 1928); τ_1, τ_2 und τ_3 Erwartungs- bzw. Skalenwerte für die Items I_1, I_2 und I_3.



vom Typ (1, 1, 1, 1); die Person hat einen Fehler gemacht. Muster wie (1, 0, 1, 0) oder (1, 1, 0, 0) enthalten zwei Fehler. Der Reproduzierbarkeitskoeffizient ist dann durch

$$\text{Rep} = 1 - \frac{\text{Gesamtzahl Fehler}}{\text{Gesamtzahl Antworten}}. \tag{2.6.26}$$

Effiziente Berechnungsmodi diskutiert Green (1956).

2.6.7 Thurstone-Skalen und die Positionierung von Probanden

Eine Möglichkeit, Skalen für die Wahrnehmung von Objekten zu erhalten, ist der Thurstonesche Paarvergleich (Thurstone (1928)). Bei den Objekten kann es sich um Personen, oder Waren (im Rahmen einer Marktforschungsuntersuchung), Meinungsaussagen etc handeln. Für je zwei Items I_g und I_h (Objekten, Aussagen, etc) soll die befragte Person angeben, ob I_g mehr (oder weniger) von einem bestimmten Merkmal aufweist als I_h oder nicht. Es kann sich auch um Präferenzurteile handeln: man gibt an, ob man das Objekt I_g dem Objekt I_h vorzieht oder nicht. Läßt man $\succ$ für "hat mehr von" oder "präferiere" stehen, so bedeutet also

$$I_g \succ I_j,$$

dass I_g mehr von einem Merkmal aufweist als I_j, oder dass I_g dem I_h präferiert wird, etc.

Sind nun die Objekte I_g, I_h, $g, h = 1, \ldots, n$ hinreichend nahe beieinander, so zeigt sich, dass Urteile der Form $I_g \succ I_h$ nur mit einer bestimmten Wahrscheinlichkeit gefällt werden. Es ist möglich, dass einige Bpn (Befragungspersonen) das Urteil $I_g \succ I_h$, andere das Urteil $I_g \prec I_h$ fällen. Es ist auch möglich, dass ein und dieselbe Person bei wiederholten Befragungen einmal $I_g \succ I_h$, das andere Mal $I_g \prec I_h$ urteilt. Anders als Guttman geht Thurstone davon aus, dass die

wahrgenommene Ausprägung des Merkmals, die durch ein Item angezeigt wird, zufällig variiert. Ist also η_g die mit dem Item I_g assoziierte wahrgenommene Ausprägung des in Frage stehenden Merkmals, so muß η_g als zufällige Veränderliche konzipiert werden. Thurstone (1928) nahm an, dass η_g normalverteilt mit einem Erwartungswert $\mathbb{E}(\eta_g) = \tau_g$ und der Varianz σ_g^2 ist, d.h. $\eta_g \sim N(\tau_g, \sigma_g^2)$. Thurstone nahm weiter an, dass $\tau_g = \mathbb{E}(\eta_g)$ ein Erwartungswert über alle Bpn ist. Für eine Person $a \in \mathcal{P}$ soll also gelten

$$\eta_{ag} = \tau_g + \varepsilon_{ag}, \quad \varepsilon_{ag} \sim N(0, \sigma_g^2). \tag{2.6.27}$$

η_{ag} heißt *Discriminal Process*[19] (Thurstone). Die Nebenbedingung für ε_{ag} bedeutet, dass der Erwartungswert der "Fehler" gleich Null ist, d.h. $\mathbb{E}(\eta_{ag}) = \tau_g$. τ_g wird als Skalenwert für das Item I_g interpretiert.

Die Person $a \in \mathcal{P}$ vergleiche nun zwei Items, I_g und I_h. Gilt $\eta_{ag} > \eta_{ah}$, so sagt sie, dass das Item I_g mehr als I_h von dem in Frage stehenden Merkmal habe, andernfalls sagt sie, I_h habe mehr von dem Merkmal als I_g. Man kann demnach schreiben

$$P_{agh} = P(I_g \succ I_h | a) = P(\eta_{ag} > \eta_{ah}), \tag{2.6.28}$$

und offenbar $P_{ahg} = P(I_g < I_h) = 1 - P(\eta_{ag} > \eta_{ah})$. Dies wiederum bedeutet, dass das Urteil von der Verteilung der Differenz $D_{agh} = \eta_{ag} - \eta_{ah}$ abhängt. Da $\eta_{ag} > \eta_{ah}$ genau dann, wenn $D_{agh} > 0$, also

$$P_{agh} = P(D_{agh} > 0). \tag{2.6.29}$$

Andererseits ist

$$D_{agh} = \eta_{ag} - \eta_{ah} = \tau_g - \tau_h + \varepsilon_{ag} - \varepsilon_{ah}. \tag{2.6.30}$$

Die Differenz zweier normalverteilter Variablen ist ebenfalls normalverteilt, in diesem Fall mit dem Erwartungswert und der Varianz

$$\mathbb{E}(D_{gh}) = \tau_g - \tau_h \tag{2.6.31}$$

$$\mathbb{V}(D_{gh}) = \sigma_{gh}^2 = \sigma_g^2 + \sigma_h^2 - 2\rho_{gh}\sigma_g\sigma_h, \tag{2.6.32}$$

wobei ρ_{gh} die Korrelation zwischen η_{ag} und η_{ah}^2 ist. Gelingt es nun, die Größen $\mathbb{E}(D_{gh}) = \tau_g - \tau_h$ für verschiedene I_g und I_h abzuschätzen, so kann man aus diesen Differenzen eine Skala für die Items konstruieren. Die folgenden Überlegungen zeigen, dass dies im Prinzip möglich ist.

Thurstones Law of Comparative Judgment Man kann nun wieder eine Indikatorvariable X_{agh} einführen: $X_{agh} = \{0, 1\}$, und zwar $X_{agh} = 1$, wenn $D_{agh} > 0$, und $X_{agh} = 0$ sonst. Gesucht ist die Wahrscheinlichkeit von $X_{agh} = 1$. Mit $U = D_{agh}$ gilt

$$P(X_{agh} = 1 | \tau_g, \tau_h, \sigma_{gh}^2) = P(U > 0), \tag{2.6.33}$$

[19] Eine Eindeutschung soll erst gar nicht versucht werden.

d.h.

$$P(X_{agh} = 1|\tau_g, \tau_h, \sigma_{gh}^2) = \frac{1}{\sigma_{gh}\sqrt{2\pi}} \int_0^\infty \exp\left[-\frac{(u - (\tau_g - \tau_h))^2}{2\sigma_{gh}^2}\right] du, \quad (2.6.34)$$

Es sei $Z = (U - (\tau_g - \tau_h))/\sigma_{gh}$. Dann ist

$$P(Z \leq z) = P\left(\frac{U - (\tau_g - \tau_h)}{\sigma_{gh}} \leq z\right) = P(U \leq \sigma_{gh}z + (\tau_g - \tau_h)).$$

Nun ist, wegen der Symmetrie der Normalverteilung, $P(U > 0) = P(U \leq 0)$, so dass (2.6.33) sein muß für denjenigen z-Wert, für den

$$\sigma_{gh}z + (\tau_g - \tau_h) = 0,$$

also

$$\tau_h - \tau_g = z\sigma_{gh} \quad (2.6.35)$$

gilt. Diese Gleichung ist Thurstones *Law of Comparative Judgment* (LCJ). Der z-Wert korrespondiert zu der Wahrscheinlichkeit, mit der $X_{agh} = 1$ beobachtet wird, d.h. mit der $\eta_{ag} > \eta_{ah}$ wahrgenommen wird. Hat man eine empirische Abschätzung $\hat{\pi}_{gh}$ dieser Wahrscheinlichkeit, so kann man den zugehörigen z-Wert bestimmen und damit über die Gleichung (2.6.35) die Differenz der Skalenwerte $\tau_h - \tau_g$ abschätzen. So kann man mit verschiedenen Item-Paaren verfahren, erhält jedesmal die Differenz der entsprechenden Skalenwerte und kann auf diese Weise eine Skala für alle Items I_g gewinnen.

Dieses Vorgehen setzt voraus, dass man eine Schätzung von σ_{gh} erhalten kann. im Allgemeinen Fall hat man nach (2.6.32) die Gleichung

$$\sigma_{gh}^2 = \sigma_g^2 + \sigma_h^2 - 2\rho_{gh}\sigma_g\sigma_h.$$

Eine Abschätzung der σ_{gh} erfordert also Schätzungen der n Varianzen σ_g, $g = 1, \ldots, n$, sowie der $n(n-1)/2$ Korrelationen ρ_{gh}, d.h. es müssen

$$n + n(n-1)/2 = \frac{n(n+1)}{2}$$

freie Parameter geschätzt werden. Denen stehen aber nur $n(n-1)/2$ Schätzungen $\hat{\pi}_{gh}$ gegenüber, woraus folgt, dass die σ_{gh} nicht ohne einschränkende Annahmen aus den Daten bestimmt werden können. Thurstone diskutierte eine Reihe von Fällen, die durch zunehmend restriktive Forderungen bezüglich der σ_g und ρ_{gh} definiert sind. Der einfachste, schätzbare Fall ist der 5-te Fall (Case V):

Spezialfall: Case V Bei diesem Fall wird angenommen, dass (i) $\sigma_g = \sigma$ für alle I_g, und (ii) $\rho_{gh} = 0$ für alle g, h mit $g \neq h$. Das LCJ wird dann zu

$$\tau_h - \tau_g = \alpha z_{gh}, \quad \alpha = 2\sigma \quad (2.6.36)$$

d.h. die Differenzen $\tau_h - \tau_g$ sind proportional zu z_{gh}, und z_{gh} kann unmittelbar – bis auf eine beliebige Proportionalitätskonstante α – aus den Daten geschätzt werden. Natürlich müssen die Annahmen (i) und (ii) überprüft werden, worauf hier aber nicht weiter eingegangen werden soll.

Die Wahrscheinlichkeit für $X_{agh} = 1$ vereinfacht sich nun zu

$$P(X_{agh} = 1|\tau_g, \tau_h) = \frac{1}{\sqrt{2\pi}} \int_{-(\tau_g - \tau_h)/\sigma}^{\infty} \exp(-z^2/2)dz. \tag{2.6.37}$$

Die Gleichungen (2.6.36) und (2.6.37) zeigen an, dass man die Differenz $\tau_h - \tau_g$ direkt aus den relativen Konfusionshäufigkeiten gewinnen kann, denn offenbar ist

$$z_{gh} = \Phi^{-1}[P(X_{agh} = 1|\tau_g, \tau_h)] = \tau_h - \tau_g, \quad \alpha = 1. \tag{2.6.38}$$

Eigentlich ist $\alpha = 1/\sigma$, aber σ ist unbekannt, und die Setzung $\alpha = 1$ bedeutet nur die Festlegung einer Einheit. Man kann die z_{12}, z_{23} etc nun aneinanderlegen und hat damit eine Intervallskala für die Items I_g, – vorausgesetzt, die Daten sind mit der Annahme der Eindimensionalität kompatibel.

Personenparameter Im LCJ (2.6.35) taucht der Index a, der eine Person indiziert, nicht mehr auf, denn die Erwartungswerte und damit die Varianzen und Kovarianzen werden durch Mittelung über die $a \in \mathcal{P}$ gebildet. Es kann nun angenommen werden, dass nicht nur die Items eine Position τ_g auf der Skala haben, sondern eine Person a, die die Items beurteilt, hat ebenfalls eine Position, etwa θ_a. Man kann dies so deuten, dass θ_a eine für die Person a ideale Merkmalsausprägung repräsentiert. Man kann dann annehmen, dass die Person a dem Item I_g zustimmt, wenn

$$D_{ag} = \tau_g - \theta_a + \varepsilon^*_{ag} > 0, \tag{2.6.39}$$

wobei ε^*_{ag} eine normalverteilte "Fehler"variable ist. Für den Erwartungswert von D_{ag} gilt

$$\mathbb{E}(D_{ag}) = \tau_g - \theta_a, \tag{2.6.40}$$

also $\mathbb{E}(\varepsilon^*_{ag}) = 0$. Für die Varianz ergibt sich

$$\mathbb{V}(D_{ag}) = \mathbb{V}(\theta_a + \varepsilon^*_{ag}) = \sigma^2_\theta + \sigma^2_g, \tag{2.6.41}$$

denn τ_g ist eine für I_g charakteristische Konstante, $\sigma^2_{g*} = \mathbb{V}(\varepsilon_{ag})$ und $\sigma^2_\theta = \mathbb{V}(\theta_a)$ (Erwartungswertbildung über alle $a \in \mathcal{P}$).

Man kann nun die Differenz der Wahrnehmungen von I_g und I_h für die Person a betrachten. Es ist

$$D_{agh} = \tau_g - \theta_a + \varepsilon^*_{ag} - (\tau_h - \theta_a + \varepsilon^*_{ah}),$$

d.h.

$$D_{agh} = \tau_g - \tau_h + \varepsilon^*_{ag} - \varepsilon^*_{ah}. \tag{2.6.42}$$

Diese Gleichung entspricht der Gleichung (2.6.30), bis auf den Fehlerterm. In (2.6.27) wurde der Discriminal Process $\eta_{ag} = \tau_g + \varepsilon_{ag}$ eingeführt. η_{ag} ist in der Tat identisch mit D_{ag}, wie in (2.6.39) eingeführt, wenn man

$$\varepsilon_{ag} = -\theta_a + \varepsilon^*_{ag} \tag{2.6.43}$$

setzt. Wie Andrich (1978) anmerkt hat Thurstone den Personparamter θ_a mit in den Fehlerterm hineingenommen, und in (2.6.43) wird dieser Term nur weiter aufgeschlüsselt. Andererseits ist

$$\varepsilon^*_{ag} - \varepsilon^*_{ah} = \varepsilon_{ag} - \varepsilon_{ah} + \theta_a - \theta_a = \varepsilon_{ag} - \varepsilon_{ah},$$

so dass die Definition von D_{agh} am Ende gleich bleibt, wie es auch sein soll.

Logistische statt Gauß-Verteilung: Das BTL-Modell Für den Case V zeigt (2.6.37), dass die Wahrscheinlichkeit für $X_{agh} = 1$ bis auf den Parameter σ nur von der Differenz $\tau_h - \tau_h$ abhängt. Die Gauß-Verteilung ist insofern unhandlich, als das Integral nicht in geschlossener Form darstellbar ist. Andererseits ist die logistische Verteilung von der Gauß-Verteilung für alle praktischen Zwecke so gut wie nicht zu unterscheiden. Setzt man also die logistische Verteilung an, so erhält man

$$P(X_{agh} = 1|\tau_g, \tau_h, \sigma) = \frac{\exp((\tau_g - \tau_h)/\sigma)}{1 + \exp((\tau_g - \tau_h)/\sigma)}. \tag{2.6.44}$$

Da σ eine Konstante ist, kann sie in die Parameter τ_g und τ_h absorbiert werden. Damit hat man das von Bradley & Terry (1952) sowie Luce (1959) vorgeschlagene Modell für den Paarvergleich, in dem *Discriminal processes* nicht mehr vorkommen (vergl Beispiel B.2.3, Seite 295 im Anhang). Andererseits muß man sehen, dass man hier nicht durch messtheoretische Überlegungen zu (2.6.44) gekommen ist, sondern über Thurstones Betrachtungen über den Urteilsprozess, in dem die logistische Verteilung zunächst als Approximation für die Normalverteilung eingeführt wurde. Man hätte allerdings auch gleich die logistische Verteilung für die Discriminal Processes annehmen können. Wichtig ist, dass dem BTL-Modell[20] (2.6.44) die Case V-Annahme zugrunde liegt.

Latent-Trait-Modell In (2.6.39) ist die Variable $D_{ag} = \tau_g - \theta_a + \varepsilon^*_{ag}$ eingeführt worden. Zur Veranschaulichung werde angenommen, dass das Item I_g eine Meinungsäußerung sei, der man zustimmen kann oder nicht. Man stimmt I_g zu, wenn die eigene Position auf der Skala, die das Meinungsspektrum abbildet, höher ist als die des Items, wenn also

$$\theta_a - \varepsilon_{ag} > \tau_g$$

ist. Das ist aber der Fall, wenn

$$0 > \tau_g - \theta_a + \varepsilon_{ag} = D_{ag}$$

[20] Bradley-Terry-Luce

ist. Gilt $\varepsilon_{ag} \sim N(0, \sigma_g^2)$, so hat man mit $P(X_{ag} = 1|\tau_g, \theta_a) = (D_{ag} < 0|\tau_g, \theta_a)$

$$P(X_{ag} = 1|\tau_g, \theta_a) = P(\varepsilon_{ag} < -(\tau_g - \theta_a)) = \int_{-(\tau_g - \theta_a)/\sigma_g}^{\infty} f(z)dz, \qquad (2.6.45)$$

f die Standardnormalverteilung. Die Approximation durch die logistische Verteilung liefert dann

$$P(X_{ag} = 1|\tau_g, \theta_a) = \frac{\exp((\theta_a - \tau_g)/\sigma_g)}{1 + \exp((\theta_a - \tau_g)/\sigma_g)}. \qquad (2.6.46)$$

Dies ist das Birnbaum-Modell für die Beantwortung von Items. Während bei der ursprünglichen Thurstone-Skala nur die Items skaliert werden, erlaubt (2.6.46) die simultane Skalierung von Items und Personen.

2.6.8 Latente Merkmale und Klassifikation

Die latente Variable θ kann kategorial sein; die Werte von θ sind dann Klassennamen. Im obigen Beispiel der Abhängigkeit der Antworten bei zwei Items I_g und $I_{g'}$ konnten die Probanden zwei Klassen zugeordnet werden, nämlich der Klasse der weiblichen und der Klasse der männlichen Probanden. Man kann $\theta = m$ für 'männlich' und $\theta = w$ für 'weiblich' setzen, oder $\theta = 0$ für 'männlich' und $\theta = 1$ für 'weiblich'. Innerhalb jeder Klasse waren die Antworten auf die beiden Items stochastisch unabhängig. Das Geschlecht einer Person ist im Allgemeinen wahrnehmbar, aber bei Merkmalen wie Berufseignung, Depression (situativ versus endogen) ist die Ausprägung nicht unmittelbar wahrnehmbar, sondern muß über einen Test erfasst werden. Die Frage ist nun, ob die lokale stochastische Unabhängigkeit bei konstanter Merkmalsausprägung (also bei Personen mit identischer Merkmalsausprägung) eine Möglichkeit liefert, latente Klassen anhand der Antworten auf Items zu finden und Probanden diesen Klassen zuzuordnen. Dies ist in der Tat möglich, wie im Folgenden gezeigt wird. Es wird zunächst der einfachste Fall von nur zwei latenten Klassen $\mathcal{C}$ und $\bar{\mathcal{C}}$ behandelt.

θ ist in diesem Fall einfach eine Indikatorvariable, also

$$\theta = \begin{cases} 1, & a \in \mathcal{C} \\ 0, & a \notin \mathcal{C}, \text{ d.h. } a \in \bar{\mathcal{C}} \end{cases} \qquad (2.6.47)$$

Die Wahrscheinlichkeit, dass eine beliebig ausgewählte Person der Klasse $\mathcal{C}$ angehört, sei $P(\theta = 1) = p_c$, und dementsprechend ist die Wahrscheinlichkeit, dass sie der Komplementärklasse $\bar{\mathcal{C}}$ angehört, $P(\theta = 0) = 1 - p_c$. Für alle Probanden mit $\theta = 1$ – die also zur Klasse $\mathcal{C}$ gehören – gilt lokale Unabhängigkeit, ebenso für alle Probanden mit $\theta = 0$, das sind die in der Komplementärklasse $\bar{\mathcal{C}}$ (p_c definiert also die Grundquote bezüglich $\mathcal{C}$). Für alle Probanden mit $\theta = 1$ – die also zur Klasse $\mathcal{C}$ gehören – gilt lokale Unabhängigkeit, ebenso für alle Probanden mit $\theta = 0$, das sind die in der Komplementärklasse $\bar{\mathcal{C}}$.

Die Wahrscheinlichkeit, dass eine Person aus $\mathcal{C}$ das Item I_g "löst", sei nun π_{g1} (der zweite Index 1 bezieht sich darauf, dass für alle Personen aus $\mathcal{C}$ die Zuordnung $\theta = 1$ gilt). Man bemerke, dass π_{g1} den gleichen Wert für alle Personen aus $\mathcal{C}$ hat, die Annahme ist ja, dass der Wert von θ die "Lösungs"wahrscheinlichkeit für ein Item bestimmt. Die Wahrscheinlichkeit, dass das Item nicht beantwortet oder gelöst wird, ist dann $1 - \pi_{g1}$. Wie üblich soll $X_g = 1$ signalisieren, dass das Item gelöst wurde. Das Antwortverhalten für das Item I_g läßt sich dann kompakt in der Form

$$h_g(X_g|\theta = 1) = \pi_{g1}^{X_g}(1 - \pi_{g1})^{1-X_g} \tag{2.6.48}$$

schreiben: h_g ist die Wahrscheinlichkeit, dass $X_g = 1$, gegeben ein Wert von θ. Für $Xg = 1$ ergibt sich die Wahrscheinlichkeit

$$h_g(X_g = 1|\theta = 1) = \pi_{g1}^{X_g}(1 - \pi_{g1})^{1-X_g} = \pi_{g1},$$

denn dann ist $1 - X_g = 0$ und $(1 - \pi_{g1})^0 = 1$. Für $X_g = 0$ ergibt sich in analoger Weise die Wahrscheinlichkeit

$$h_g(X_g = 0|\theta = 1) = 1 - \pi_{g1}.$$

Betrachtet wird nun die Wahrscheinlichkeit f, dass die X_g die speziellen Werte $x_g = 1$ oder $x_g = 0$ annehmen: wegen der lokalen Unabhängigkeit gilt dann

$$f(x_1, \ldots, x_n|\theta = 1) = \prod_{g=1}^{n} h_g(x_g|\theta = 1) = \prod_{g=1}^{n} \pi_{g1}^{x_g}(1 - \pi_{g1})^{1-x_g} \tag{2.6.49}$$

Für Probanden aus $\bar{\mathcal{C}}$ sei die Wahrscheinlichkeit, das Item I_g zu beantworten, gleich π_{go}. Alle vorangegangenen Betrachtungen übertragen sich auf diesen Fall, so dass analog zu (2.6.49) die Beziehung

$$f(x_1, \ldots, x_n|\theta = 0) = \prod_{g=1}^{n} h_g(x_g|\theta = 0) = \prod_{g=1}^{n} \pi_{g0}^{x_g}(1 - \pi_{g0})^{1-x_g} \tag{2.6.50}$$

Wird nun eine beliebige Person gewählt, so gehört sie mit der Wahrwscheinlichkeit p_c zu $\mathcal{C}$ und mit der Wahrscheinlichkeit $1 - p_c$ zu $\bar{\mathcal{C}}$. Nach dem Satz der totalen Wahrscheinlichkeit gilt dann

$$f(x_1, \ldots, x_n) = p_c \prod_{g=1}^{n} h_g(x_g|\theta = 1) + (1 - p_c) \prod_{g=1}^{n} h_g(x_g|\theta = 0). \tag{2.6.51}$$

Klassifikation: Sind nun die Wahrscheinlichkeiten p_c und π_{g1} sowie π_{g0} für $g = 1, \ldots, n$ bekannt (sie müssen im Allgemeinen geschätzt werden!), so kann eine Person auf der Basis ihrer Werte $x_1, \ldots, x_n$ einer der beiden Klassen zugeordnet werden. Während $p_c = P(\theta = 1)$ eine *a-priori*-Wahrscheinlichkeit für

die Zugehörigkeit zur Klasse $\mathcal{C}$ ist, muß für eine Klassifikation eine *a-posteriori*-Wahrscheinlichkeit

$$P(\theta = 1|x_1, \ldots, x_n)$$

für die Zugehörigkeit zu $\mathcal{C}$ bestimmt werden. Natürlich gilt

$$P(\theta = 0|x_1, \ldots, x_n) = 1 - P(\theta = 1|x_1, \ldots, x_n).$$

Nach dem Satz von Bayes gilt

$$P(\theta = 1|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n \cap \theta = 1)}{f(x_1, \ldots, x_n)} = \frac{f(x_1, \ldots, x_N|\theta = 1)p_c}{f(x_1, \ldots, x_N)}, \quad (2.6.52)$$

und

$$P(\theta = 0|x_1, \ldots, x_n) = 1 - P(\theta = 1|x_1, \ldots, x_n). \quad (2.6.53)$$

Gegeben die $x_1, \ldots, x_n$ kann man eine Person nun der Klasse $\mathcal{C}$ zuordnen, wenn

$$P(\theta = 1|x_1, \ldots, x_n) > P(\theta = 0|x_1, \ldots, x_n), \quad (2.6.54)$$

und der Klasse $\bar{\mathcal{C}}$, wenn umgekehrt $P(\theta = 1|x_1, \ldots, x_n) < P(\theta = 0|x_1, \ldots, x_n)$.

Die Gleichung (2.6.52) kann noch umgeschrieben werden, so dass bestimmte Eigenschaften dieser Wahrscheinlichkeit deutlicher werden. Berücksichtigt man (2.6.49) und (2.6.51), so erhält man mit $\vec{x} = (x_1, \ldots, x_n)$ den zunächst etwas unübersichtlicheren Ausdruck

$$P(\theta = 1|\vec{x}) = \frac{p_c \prod_{g=1}^{n} \pi_{g1}^{x_g}(1 - \pi_{g1})^{1-x_g}}{p_c \prod_{g=1}^{n} \pi_{g1}^{x_g}(1 - \pi_{g1})^{1-x_g} + (1 - p_c) \prod_{g=1}^{n} \pi_{g0}^{x_g}(1 - \pi_{g0})^{1-x_g}}. \quad (2.6.55)$$

Teilt man nun den Zähler und den Nenner durch $p_c \prod_{g=1}^{n} \pi_{g1}^{x_g}(1 - \pi_{g1})^{1-x_g}$ und berücksichtig man, dass

$$p_c \prod_{g=1}^{n} \pi_{g1}^{x_g}(1 - \pi_{g1})^{1-x_g} = p_c \exp\left[\sum_{g=1}^{n}(x_g \log \pi_{g1} + (1 - x_g) \log(1 - \pi_{g1}))\right]$$

und analog

$$(1 - p_c) \prod_{g=1}^{n} \pi_{g0}^{x_g}(1 - \pi_{g0})^{1-x_g} = p_c \exp\left[\sum_{g=1}^{n}(x_g \log \pi_{g0} + (1 - x_g) \log(1 - \pi_{g0}))\right],$$

so erhält man für $P(\theta = 1|x_1, \ldots, x_n)$ den Ausdruck

$$P(\theta = 1|\vec{x}) = \frac{1}{1 + \left(\frac{1-p_c}{p_c}\right) \exp\left[\sum_{g=1}^{n}\left(x_g \log \frac{\pi_{g0}}{\pi_{g1}} + (1 - x_g) \log \frac{1-\pi_{g0}}{1-\pi_{g1}}\right)\right]}. \quad (2.6.56)$$

Es ist aber $P(\theta = 0|\vec{x}) = 1 - P(\theta = 1|\vec{x})$, so dass[21]

$$P(\theta = 0|\vec{x}) = \frac{\left(\frac{1-p_c}{p_c}\right) \exp\left[\sum_{g=1}^{n}\left(x_g \log\frac{\pi_{g0}}{\pi_{g1}} + (1-x_g)\log\frac{1-\pi_{g0}}{1-\pi_{g1}}\right)\right]}{1 + \left(\frac{1-p_c}{p_c}\right) \exp\left[\sum_{g=1}^{n}\left(x_g \log\frac{\pi_{g0}}{\pi_{g1}} + (1-x_g)\log\frac{1-\pi_{g0}}{1-\pi_{g1}}\right)\right]}. \tag{2.6.57}$$

Die Nenner in den Ausdrücken für $P(\theta = 1|\vec{x})$ und $P(\theta = 0|\vec{x})$ sind identisch und spielen deshalb beim Vergleich dieser beiden Wahrscheinlichkeiten keine Rolle; betrachtet man etwa den Quotienten von $P(\theta = 1|\vec{x})$ und $P(\theta = 0|\vec{x})$, so kürzt sich der Nenner heraus und man hat

$$\frac{P(\theta = 1|\vec{x})}{P(\theta = 0|\vec{x})} = \frac{1}{\left(\frac{1-p_c}{p_c}\right) \exp\left[\sum_{g=1}^{n}\left(x_g \log\frac{\pi_{g0}}{\pi_{g1}} + (1-x_g)\log\frac{1-\pi_{g0}}{1-\pi_{g1}}\right)\right]}. \tag{2.6.58}$$

Entscheidet man sich für die Zuordnung $a \to \mathcal{C}$, wenn $P(\theta = 1|\vec{x}) > P(\theta = 0|\vec{x})$, so muß für diese Entscheidung also

$$1 > \left(\frac{1-p_c}{p_c}\right) \exp\left[\sum_{g=1}^{n}\left(x_g \log\frac{\pi_{g0}}{\pi_{g1}} + (1-x_g)\log\frac{1-\pi_{g0}}{1-\pi_{g1}}\right)\right]$$

oder

$$\frac{p_c}{1-p_c} > \exp\left[\sum_{g=1}^{n}\left(x_g \log\frac{\pi_{g0}}{\pi_{g1}} + (1-x_g)\log\frac{1-\pi_{g0}}{1-\pi_{g1}}\right)\right]$$

gelten. Da der Logarithmus eine monoton steigende Funktion ist, kann man auch beide Seite logarithmieren und die Logarithmen miteinander vergleichen. Mit

$$\operatorname{Logit} p = \log\frac{p}{1-p} \tag{2.6.59}$$

erhält man dann

$$\operatorname{Logit} p_c > \sum_{g=1}^{n}\left(x_g \log\frac{\pi_{g0}}{\pi_{g1}} + (1-x_g)\log\frac{1-\pi_{g0}}{1-\pi_{g1}}\right) \tag{2.6.60}$$

Dieser Ausdruck läßt sich weiter vereinfachen zu

$$0 > \sum_{g=1}^{n} x_g(\operatorname{Logit}\pi_{g0} - \operatorname{Logit}\pi_{g1}) + \sum_{g=1}^{n}\log\frac{1-\pi_{g0}}{1-\pi_{g1}} - \operatorname{Logit} p_c,$$

oder, gleichbedeutend damit, zu

$$X = \sum_{g=1}^{n} x_g(\operatorname{Logit}\pi_{g0} - \operatorname{Logit}\pi_{g1}) > \sum_{g=1}^{n}\log\frac{1-\pi_{g1}}{1-\pi_{g0}} + \operatorname{Logit} p_c. \tag{2.6.61}$$

Die *Diskriminanzfunktion* X hängt linear von x_g ab! Lineare Diskriminanzfunktionen spielen in der Klassifikation von Personen oder Objekten generell eine zentrale Rolle.

[21] $P(\theta = 1|\vec{x})$ hat die Form $1/(1+A)$; dann ist $P(\theta = 0|x_1, \ldots, x_n)$ von der Form $1-1/(1+A) = (1 + A - 1/(1 + A) = A/(1 + A)$.

2.6.9 Der theoretische Status latenter Variablen

Modelle und ihre Semantik: Latente Variable sind Variable, die auf der Basis von Messungen bestimmt werden, also abgeleitete Größen. Im Rahmen psychometrischer Tests sind die Messungen die Antworten der Probanden auf Items, und die Beziehung zwischen der Antwort und der eigentlich interessierenden latenten Variablen wird entweder durch eine Itemfunktion oder auf der Basis eines faktoranalytischen Modells hergestellt: werden einer Stichprobe von m Probanden n Items vorgelegt und ist x_{ij} die numerische Repräsentation der Antwort der i-ten Person auf das j-te Item, so kann x_{ij} in der Form[22]

$$x_{ij} = \theta_{i1}a_{1j} + \cdots + \theta_{in}a_{nj}, \;\; i = 1, \ldots, m; \; j = 1, \ldots, n > m. \tag{2.6.62}$$

wobei θ_{ik} der Faktorscore der i-ten Person bezüglich der k-ten latenten Variable ist, also die Ausprägung der k-ten latenten Dimension bei der i-ten Person ist , und a_{kj} ist die Ladung des j-ten Tests oder Items auf der k-ten latenten Variablen. Die Repräsentation der x_{ij} gilt exakt, ist aber "trivial", weil so viele latente Variablen postuliert werden, wie es Items gibt. Um eine Approximation an die Faktorenanalyse zu erhalten, versucht man, die x_{ij} mit nur $r < n$ latenten Werten zu approximieren, also

$$x_{ij} \approx \theta_{i1}a_{1j} + \cdots + \theta_{ir}a_{rj}, \quad r < n \tag{2.6.63}$$

zu setzen; auf die hiermit verbundenen Fragen muß aber im hier gegebenen Zusamenhang nicht eingegangen werden. Für die θ_{ik} und a_{jk} gelten jedenfalls die Gleichungen[23]

$$\theta_{ik} = a_{1k}x_{i1}/\lambda_k + \cdots + a_{nk}x_{in}/\lambda_k. \tag{2.6.64}$$

$$a_{jk} = \theta_{1j}x_{1j} + \cdots + \theta_{mj}x_{mj}. \tag{2.6.65}$$

Formal kann man die θ_{ik} als eine Art gewogenes Mittel der Scores der i-ten Person betrachten, und die Ladungen a_{jk} als ein gewogenes Mittel der Scores für das j-te Item (gemittelt über die m Probanden). Die Tabelle 4 zeigt noch einmal die Messwerte sowie die Skalenwerte θ_{ik} für die Personen und a_{jk} für die Items, $k = 1, \ldots, n$. Die Tabelle 4 gibt das Schema an, nach dem die a_{jk} und θ_{ik} gewonnen werden. Bei einer Anwendung im Rahmen der Klassischen Testtheorie, d.h. bei der Berechung des Summenscores einer getesteten Person, wird von Gewichten a_{ik} Gebrauch gemacht, die anhand einer großen, repräsentativen Stichprobe

[22] Der Einfachheit halber wird nur der Ansatz der Hauptachsentransformation dargestellt.

[23] Die Gleichungen entsprechen der *Singularwertzerlegung* der $m \times n$-Datenmatrix X, derzufolge stets $X = Q\Lambda^{1/2}P'$ gilt, wobei P die normierten Eigenvektoren von $X'X$ und Q die normierten Eigenvektoren von XX' – jeweils als Spaltenvektoren – enthält. $\Lambda^{1/2}$ ist die Diagonalmatrix der Wurzeln der von Null verschiedenen Eigenwerte von $X'X$ und XX'. Die Matrizen P und Q sind orthonormal. $A = P\Lambda^{1/2}$, d.h. für das Element a_{jk} (Ladung des j-ten Items auf der k-ten latenten Variablen) gilt $a_{jk} = p_{jk}\sqrt{\lambda_k}$. Die Gleichungen (2.6.64) und (2.6.65) ergeben sich aus $X = QA'$, denn wegen der Orthonormalität folgt einerseits $Q'X = A'$, d.h. $A = X'Q$, andererseits $XA = QA'A = Q\Lambda$, d.h. $Q = XA\Lambda^{-1}$, und $\Lambda^{-1} = \mathrm{diag}(1/\lambda_1, \ldots, \lambda_n)$.

Tabelle 4: Daten, Item-(a_{jk}) und Personenparameter (θ_{ik}); m Personen, n Items. Maximalzahl latenter Variablen ist n, interpretiert werden i.A. $r < n$

	Items $I_1, \ldots, I_n$				Faktorwerte		
	x_{11}	x_{12}	$\cdots$	x_{1n}	θ_{11}	$\cdots$	θ_{1n}
Personen	x_{21}	x_{22}	$\cdots$	x_{2n}	θ_{21}	$\cdots$	θ_{2n}
			$\vdots$			$\vdots$	
$(i = 1, \ldots, m)$	x_{i1}	x_{i2}	$\cdots$	x_{in}	θ_{i1}	$\cdots$	θ_{in}
			$\vdots$			$\vdots$	
	x_{m1}	x_{m2}	$\cdots$	x_{mn}	θ_{m1}	$\cdots$	θ_{mn}
	a_{11}	a_{12}	$\cdots$	a_{1n}			
Faktorladungen	a_{21}	a_{22}	$\cdots$	a_{2n}			
			$\vdots$				
(Items $j = 1, \ldots, n$)	a_{j1}	a_{j2}	$\cdots$	a_{jn}			
			$\vdots$				
	a_{n1}	a_{n2}	$\cdots$	a_{nn}			

gewonnen wurden und zu der der Proband nicht gehört, – wohl aber sollte er (oder sie) zu der Population gehören, aus der die Stichprobe ausgewählt wurde. Die a_{ik} repräsentieren dann Merkmale der Population.

Die Existenz der θ_{ik} und der a_{jk} ist durch ein Resultat der Algebra gesichert, ganz unabhängig davon, was die x_{ij} bedeuten. Das durch die Gleichungen (2.6.62), (2.6.64) und (2.6.65) definierte Modell ist also eigentlich nichts anderes als eine Interpretation der Anwendung eines algebraischen Theorems zur Beschreibung von Daten, das selbst nichts mit der Bedeutung der Messungen zu tun hat. Dieser Sachverhalt reflektiert ein Problem. Denn es ist ja denkbar, dass die Antworten auf die Items $I_1, \ldots, I_n$ in der Tat durch Ausprägungen von latenten Variablen bestimmt werden, – die aber nicht linear, also additiv, wirken. So seien etwa η_1 und η_2 die durch die $I_1, \ldots, I_n$ erfassten Variablen, und es gelte

$$x_{ij} = b_{i1}\eta_{j1} + b_{i2}\eta_{j2} + b_{i3}\eta_{j1}\eta_{j2} + e_{ij}. \tag{2.6.66}$$

b_{i1} und b_{i2} seien die Ausprägungen von η_1 und η_2 bei der i-ten Person, η_{j1} und η_{j2} seien die Maße, mit denen η_1 bzw.η_2 durch das j-te Item erfasst werden. b_{i3} sei die Ausprägung der Wechselwirkung zwischen den beiden latenten Variablen bei der i-ten Person, wobei die Wechselwirkung durch das Produkt $\eta_{j1}\eta_{j2}$ beschrieben werde. Dies bedeutet, dass es außer einer "einfachen" Wirkung der Variablen η_{j1} und η_{j2} noch eine weitere Wirkung von η_{1j} gibt, die proportional zu η_{2j} (und umgekehrt) wirkt. Wechselwirkungen, die durch Produkte beschrieben werden, kennt man z.B. aus der Chemie: die Intensität der Reaktion zweier Substanzen

σ_1 und σ_2 hängt von der Anzahl möglicher Molekülkombinationen (s_1, s_2) ab, wobei s_1 ein Molekül der Substanz σ_1 und s_2 ein Molekül der Substanz σ_2 ist, und insgesamt gibt es $n_1 \times n_2$ mögliche Kombinationen, wenn n_j die Anzahl der Moleküle der Substanz σ_j ist (Massewirkungsgesetz). Ob man eine derartige Interpretation auf psychische oder kognitive Reaktionen übertragbar ist, hängt von der Theorie ab, in deren Rahmen die Reaktionen betrachtet werden. Die faktorenanalytischen Methoden (hier: die Hauptachsentransformation), die auf die Repräsentation (2.6.62) zielen, werden stets die θ_{ik} und a_{jk} liefern, die den Bedingungen (2.6.64) und (2.6.65) genügen, weil eben aus rein algebraischen Gründen eine lineare Dekomposition stets möglich ist; das wahre Modell (2.6.66) bleibt unidentifiziert. Man kann den Fall haben, dass die Dimensionen, die durch den linearen Ansatz bestimmt werden, u. U. (d.h. durch entsprechende Rotation der Achsen) durch bestimmte Items repräsentiert werden können, sofern die Antworten auf die Items unkorreliert sind; die Antworten auf die übrigen Items lassen sich dann durch Linearkombination der Dimensionen gewissermaßen "vorhersagen". Es bleibt aber dabei, dass ein objektiv falsches Modell akzeptiert wird, denn das wahre Modell ist eben (2.6.66) und nicht (2.6.62). Wird die Faktorenanalyse zur Bestimmung latenter Dimensionen gewählt und dabei ein lineares Modell angenommen, so kann die nichtlineare Beziehung (2.6.66) nicht identifiziert werden.

Dieses Gedankenexperiment führt auf einen *konstruktivistischen* bzw. *operationalistischen* Ansatz zur Interpretation latenter Variablen. Der konstruktivistischen Auffassung zufolge sind Naturgesetze und dementsprechend eben auch latente Variable Konstruktionen des menschlichen Geistes, die in der Natur gar nicht existieren müssen. Zur Veranschaulichung denke man an den Kraftbegriff der newtonschen Physik, demzufolge die Kraft F proportional zu Beschleunigung b eines Körpers ist, wobei der Proportionalitätsfaktor die Masse m des Körpers ist, also $F = mb$. Die Frage ist, ob die Kraft F eine tatsächliche Größe in der Natur oder ob F nur ein Konstrukt unserer Denkweise ist, mit dem sich gewisse Vorgänge in der Natur exzellent beschreiben lassen. Nicht identisch, aber verwandt mit diesem Ansatz ist der operationalistische Ansatz, bei dem Begriffe durch die Art und Weise – eben durch die Operation – ihrer Erfassung, also in einem allgemeinen Sinne ihrer Messung, definiert werden. Indem man die Masse m und die Beschleunigung b eines Körpers – die wiederum operational definiert werden – mißt und das Produkt mb bildet, hat man operational die einwirkende Kraft definiert bzw. bestimmt; was hier zum Ausdruck kommt, ist die eigenartig zwischen Definition und Gesetz schwebende Charakterisierung des Kraftbegriffs. In der Psychologie ist der Merkspruch "Intelligenz ist, was der Intelligenztest mißt" bekannt, der dem operationalen und in gewisser Weise auch dem konstruktivistischen Vorgehen entspricht. Andererseits wird man von den Fähigkeiten, die in einem Intelligenztest gemessen werden, kaum sagen können, dass sie *nichts* mit Intelligenz zu tun haben. Hier werden für den wissenschaftstheoretisch Interessierten Fragen bezüglich des verwendeten Wahrheitsbegriffes aufgeworfen. Der intuitiv naheliegende Korrespondenzbegriff der Wahrheit – eine Aussage ist wahr

genau dann, wenn die Dinge in der Welt so sind, wie die Aussage sie beschreibt – scheint auf den ersten Blick nicht ganz angemessen zu sein. Andererseits werden mit einem Intelligenztest betimmte Fähigkeiten erfaßt, und die Charakterisierung dieser Fähigkeiten als 'Intelligenz' kann als nominalistische Definition dieses Merkmalkomplexes aufgefaßt werden, die von der Wahrheit der Messungen unabhängig ist. In Borsboom, Mellenbergh & van Heerden (2003) werden einige dieser Fragen in Bezug auf die Testtheorie angesprochen.

Diese Betrachtungen übertragen sich auf den Begriff der latenten Variablen, wie er in Item-Response-Theorien verwendet wird. Während aber bei faktorenanalytischen Ansätzen die θ_{ik} direkt aus den x_{ij} errechnet werden, wird bei den IR-Theorien ein probabilistischer Zusammenhang zwischen der latenten Variablen und dem beobachtbaren Antwortverhalten postuliert. Die Frage ist hier, wie dieser Zusammenhang zu deuten ist.

Stochastische Probanden: In Abschnitt 2.6.3 wurde bereits der Ansatz der 'unterliegenden Variablen' dargestellt. Diesem Ansatz entsprechend fluktuiert das gemessene Merkmal innerhalb der Person und das Antwortverhalten hängt vom Wert der latenten Variablen zur Zeit der Messung ab. Lazarsfeld (1950) (zitiert nach Lord & Novick (1968, p. 29) dikutiert bereits einen Mr. Brown, der zu wiederholt zu seiner Einstellung zu den Vereinten Nationen (UN) befragt wird. Damit die Reaktionen von Mr. Brown unabhängig voneinander sind, sei es notwendig, ihn nach jeder Befragung einer Hirnwäsche ("to wash his brain") zu unterziehen. Nach einer hinreichend großen Zahl von Befragungen könne man dann die Wahrscheinlichkeit einer bezüglich den UN positiven Antwort als relative Häufigkeit bestimmen. Dabei könne das Antwortverhalten vom Zustand Mr. Browns abhängen: hat er keinen Alkohol getrunken, äußert er sich eher positiv, weil er sich der sozialen Erwünschheit einer positiven Einstellung bewußt ist, steht er dagegen unter Alkoholeinfluß, läßt er seiner Abneigung gegen internationale Institutionen freien Lauf. Die Antwort von Herrn Brown wird, so Lazarsfeld, durch Herrn Browns jeweilige *propensity* [24] bestimmt. Lord & Novick (1968), p. 30, betrachten Studierende, die ihre Examensleistungen u. a. auf ihren Zustand während des Examens zurückführen, der in nicht genau kontrollierbarer Weise fluktuiere. Die entsprechende Verteilung heißt dann, der Lazarsfeldschen Anregung zufolge, *propensity distribution*. In Abschnitt 2.12, diskutieren Lord & Novick dementsprechend dann *replicate measurements*; bei einer gegebenen Person $a \in \mathcal{P}$ variiert der Score X_{ag} für das Item I_g von Messung zu Messung, d.h. X_{ag} ist, für gegebenes Item I_g, eine zufällige Veränderliche. Ursache für die zufällige Variation von X_{ag} sind Einflüsse auf die Messungen, die nicht weiter kontrolliert werden können, wobei es sich auch im "innere" Einflüsse wie Aufmerksamkeitsschwankungen etc.

[24] Den Begriff der propensity (Propensität, etwa: Neigung) geht, im Zusammenhang mit Wahrscheinlichkeiten, auf Popper zurück, der damit einen objektiven Wahrscheinlichkeitsbegriff begründen wollte, vergl. Mortensen, U.: Wissenschaftstheorie III, Abschnitt 2.2 (http://www.uwe-mortensen.de/WisstheorieIII.pdf)

handeln kann.

Ellis & van Wollenberg (1993) argumentieren, dass *Latent trait* Modelle auf drei Basisannahmen beruhen: (i) die Eindimensionalität des gemessenen Merkmals, (ii) dem Prinzip der lokalen Unabhängigkeit, und (iii) der Monotonität der Itemfunktion, d.h. je größer die Fähigkeit oder Ausgeprägtheit des Merkmals, desto größer soll die Wahrscheinlichkeit einer Antwort sein. Die Möglichkeit, dass es stochastische Aspekte im Antwortverhalten einer Person gibt derart, dass die bedingten Wahrscheinlichkeiten $F_g(\theta)$ und $P(\eta > \eta_0|\theta)$ identisch sind, definieren die Autoren als *lokale Homogenität*. Die üblichen *Latent trait* Modelle postulieren diese lokale Homogenität nicht, und Ellis et al. führen aus, dass das Postulat eine vierte, von den ersten drei Annahmen unabhängige Annahme ist. Sie bedeutet, wie bereits Lord & Novick (1968), p. 539, ausgeführt haben, dass Personen mit dem gleichen Wert von θ auch die gleichen Antwortwahrscheinlichkeiten haben. In Abschnitt 2.6.3 wurde bereits für die Spezialfälle des Ogiven- sowie des logistischen Modells ausgeführt, dass die Verteilungen der unterliegenden Variablen für die verschiedenen Personen die gleichen Parameterwerte etwa für die Varianzen haben müssen; die Argumente von Ellis et al. beziehen sich allgemeiner auf alle möglichen *Latent trait* Modelle. Gilt die Annahme der lokalen Homogenität, sprechen Ellis et al. von einem *homogenen monotonen IRT Modell*. In der Diskussion argumentieren sie, dass nur solche Modelle für die Psychologie von Interesse seien, – wenn man die Psychologie als die Wissenschaft vom Verhalten individueller Personen betreiben wolle: "So the probabilities described by an inhomogeneous model seem not particularly important for psychological theory." (Ellis et al., p. 427) Die Autoren diskutieren Möglichkeiten, die Homogentitätsannahme zu testen, worauf an dieser Stelle aber nicht eingegangen werden kann.

Holland (1990) nimmt einen anderen Standpunkt ein. Er spricht im Zusammenhang mit der Möglichkeit stochastischen Antwortverhaltens von einem *stochastic subject*. Holland (1990) zitiert eine Reihe von Autoren, die eine stochastische Zuständlichkeit von Personen annehmen, äußert aber die Ansicht: "I believe that no completely satisfactory justification of the "stochastic subject" is possible," und hält die Erklärungen, die bezüglich der Mechanismen, die das Antwortverhalten einer Person stochastisch erscheinen lassen, für tautologisch: "subjects are stochastic, because human behavior is uncertain". Aber eine solche Aussage ist wohl eher als Polemik zu verstehen, mit der er seine eigene Auffassung rethorisch stützen will. Denn die Annahme des stochastischen Subjekts ist eine Folge der beobachteten Unsicherheit (im Sinne von mangelnder Vorhersagbarkeit) im menschlichen Verhalten, und dieses Verhalten ist keine Folge der Annahme eines stochastischen Subjekts. Hollands Erklärung für die Zufälligkeit der Antworten beim Testen von Personen basiert auf dem Begriff der zufälligen Auswahl von Probanden (*random sampling of examinees*). Tatsächlich kann ein Proband, der von einem Psychologen getestet werden soll, aus der Sicht des Psychologen wie die zufällige Wahl eines Probanden aus einer Population möglicher Probanden gesehen werden. Holland postuliert:

> "There are no stochastic mechanisms operating except the sampling process by which examiness are selected from C [der Population $\mathcal{P}$] and tested with T [dem Test]. I shall make no probabilistic assumptions about *how* an examinee arrives at his or her particular response vector. I only assume that an examinee, tested with T, will produce some response vector. This assumption, along with the random sampling from C, will be the sole basis for statistical inferences. This is the random sampling perspective and should be distinguished from the stochastic subject perspective ..." (Holland (1990), p. 579).

Holland räumt ein, dass abgesehen vom Test-Item das Verhalten in einem Test von vielen Faktoren abhängen kann, die nicht alle kontrollierbar sind. Dementsprechend führt er weiter aus:

> "It should be emphasized that I do *not* make the deterministic assumption that if an examinee where *retested* with T then the *same* response pattern would be produced by that examinee. Even if retesting were done under identical conditions, the examinee is no longer the same person in the sense that relevant experiences (i.e., prior exposure to T) have occurred that had not occured prior to the first testing. In short, the tested sample of examinees has changed from a sample from C to a sample from a new population, C'. Thus, we explicitly do not include individual variability in test performance over time as a source of probability in defining $p(x)$. Rather, $p(x)$ is merely the proportion of C who would produce response vector x if tested with T. Individual variability over time is absorbed into the definition of the population C. As the population C "ages" in some important ways, it changes and so do the values of $p(x)$."(Holland (1990), p. 579).

Holland nimmt damit an, dass das Verhalten einer Person *unter gegebenen Umständen* vollständig determiniert ist. Diese Annahme ist ebenso kühn wie umstritten, und im gegebenen Zusammenhang ist sie überflüssig.

Beim *random sampling* Modell kommt es nur auf die Unterschiede zwischen den getesteten Personen an; die Rede ist dementsprechen auch von *between-subject* Modellen, im Gegensatz zu den *within-subject* Modellen, die die *stochastic subject*-Annahme machen (Borsboom, Mellenbergh & van Heerden, 2003). Die Hollandsche Art, die Antwortvariabilität eines Probanden zu deuten, mag einem einigermaßen verquer erscheinen. Sie leitet aber über zu einer Interpretation von Testmodellen, die in der Tat die Möglichkeit der Stochastizität des Probandenverhaltens vernachlässigen. Borsboom, Mellenbergh und van Heerden (2003) haben hierfür einige Argumente vorgelegt. Die vierte Annahme der lokalen Homogenität, wie sie von Ellis et al. (1993) eingeführt wurde und derzufolge ein Modell, dass für eine Population gilt, auch für jedes Mitglied der Population gilt, *muß*

ja keineswegs gelten. Molenaar, Huizenga & Nesselroade (2003) haben Simulationen durchgeführt, bei denen Testantworten aufgrund verschiedener Modelle, korrespondierend zu verschiedenen Personen, erzeugt wurden: bei einigen Personen wurde ein Einfaktormodell angenommen, bei anderen ein Zweifaktormodell, und so fort. Wurde dann an diese Daten ein sogenanntes *between-subject* Modell angepasst, so lieferte dieses bei der Annahme von nur einem, höchstens zwei Faktoren einen sehr guten Fit. Andere Autoren kamen zu ähnlichen Ergebnissen (Borkenau & Ostendorf (1998), Mischel & Shoda (1998), Feldman (1995), Cervone (1977)). Diese Befunde zeigen, dass *between-* und *within-subject* Modelle nicht viel miteinander zu tun haben müssen (Borsboom et al.).

Verläßt man den *within-subject*-Standpunkt und gibt damit den Begriff des *stochastic subject*s auf, so muß die Existenz zufälliger Variabilität in den Antwortmustern durch das zufällige Ziehen der Probanden aus der Population $\mathcal{P}$ begründet werden. Der Merkmalsparameter θ einer Person reflektiert ein konstantes Merkmal und wird als ein die Person charakterisierender Skalenwert interpretiert. Der Antwortprozess wird nicht mehr, wie beim *stochastic-subject*-Modell, mit dem Testmodell gleichgesetzt. Wie Borsboom et al. ausführen, bedeutet dann das Big-Five-Modell *nicht* mehr, dass dieses Modell auch für jede Person gilt. Diese Betrachtungsweise wird plausibel, wenn man sich noch einmal das am Anfang dieses Abschnitts eingeführte Faktorenmodell ansieht. Nach (2.6.64) wird θ_{ik}, also der Wert der i-ten Person auf der k-ten latenten Variablen, als gewogenes Mittel (über die Items)

$$\theta_{ik} = \sum_{j=1}^{n} a_{jk} x_{ij}$$

der Messwerte berechnet. Die Meßwerte x_{ij} sind zwar spezifisch für die i-te Person, die "Gewichte" a_{jk} aber nicht, – diese ergeben sich als gewogenes Mittel (über die Personen)

$$a_{jk} = \sum_{i=1}^{m} \theta_{jk} x_{jk}$$

der Messwerte der Personen für das j-te Item, und die "Gewichte" sind hier die θ_{ik}. Die a_{jk} als Mittelwerte über die Personen bestimmen, ob die Antwort auf ein Item I_j mit einem großen, mittleren oder kleinen Anteil in die θ_{ik} eingehen, und in diesem Sinne sind die θ_{ik} Messwerte, die die Relationen zwischen den Personen definieren. Die a_{jk} und θ_{ik} werden nur durch die einmalig bestimmten x_{ij} bestimmt und bilden gewissermaßen nur die Ausprägungen auf den latenten Variablen zum Messzeitpunkt ab, nicht aber interne zeitliche Fluktuationen. Ob die durch die θ_{ik} und a_{jk} definierte Struktur auch der Struktur *innerhalb* jeder Person entspricht, bleibt prinzipiell eine offene Frage, wobei aber eine negative Antwort wahrscheinlich ist. Die Interpretation der Personparameter als Messwerte, die die Relationen zwischen Personen abbilden, ist also keine neue Vorstellung, sondern bereits immanent in den faktorenanalytischen Modellen enthalten.

2.6.10 Die Schätzung von Parametern

Wichtige Parameter in jeder Testtheorie sind die Schwierigkeit der Items, die "Fähigkeit" (Ausprägung des jeweils gemessenen Merkmals), und natürlich andere Größen wie die Gültigkeit und Reliabilität des Tests und der einzelen Items. Die Werte dieser Parameter sind unbekannt und müssen aus den Daten geschätzt werden. Es soll hier ein kurzer Überblick über die wichtigsten Methoden zur Schätzung von Parametern und die statistischen Eigenschaften dieser Schätzungen gegeben werden.

Überblick über die gängigen Methoden: Die am meisten verwendeten Methoden sind

1. **Die Momentenmethode:** Will man z.B. den Erwartungswert und die Varianz einer zufälligen Veränderlichen X Schätzen, so berechnet man einfach die zu den Definitionen dieser Größen korrespondierenden Stichprobenmomente. So ist der Erwartungswert durch

$$\mathbb{E}(X) = \sum_{alle\ i} p_i x_i \quad X \text{ ist diskret, bzw. } \mathbb{E}(X) = \int_I x f(x) dx \quad X \text{ ist stetig} \tag{2.6.67}$$

   definiert (I ist der Definitionsbereich der Variablen X), und die Varianz ist durch

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) \tag{2.6.68}$$

   definiert. Die Schätzungen von $\mu = \mathbb{E}(X)$ und $\sigma^2 = \mathbb{V}(X)$ nach der Momentenmethode sind dann der Stichprobenmittelwert und die Stichprobenvarianz

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{2.6.69}$$

   Man beachte, dass bei der Schätzung von σ^2 nicht durch $n-1$, sondern durch n geteilt wurde, denn s^2 ist eben eine Schätzung von σ^2, die der Definition der Varianz als *durchschnittliche* quadrierte Abweichung vom Erwartungswert entspricht; die Varianz ist demnach ein arithmetisches Mittel *aller* möglichen Abweichungen von X vom Erwartungswert $\mathbb{E}(X)$.

2. **Die Methode der Kleinsten Quadrate:** Gegeben sind Messwerte y_i, $i = 1, \ldots, n$, und ein Modell, mit dem man die y_i "erklären" oder "vorhersagen" will. Das Modell enthält freie – d.h. noch zu bestimmende – Parameter a, b, ... Die Vorhersagen der y_i-Werte haben die Form $\hat{y}_i(a, b, \ldots)$. Für gegebene Parameterwerte unterscheiden sich die y_i- und $\hat{y}_i$-Werte um "Fehler" e_i, d.h. $y_i = \hat{y}_i + e_i$. Die Parameter werden geschätzt, indem man die Funktion

$$Q(a, b, \ldots) = \sum_{i=1}^{n}(y_i - \hat{y}_i(a, b, \ldots))^2 = \sum_{i=1}^{n} e_i^2 \tag{2.6.70}$$

als Funktion der freien Parameter $a,$, $b,\ldots$ minimalisiert. Das einfachste Modell ist $y_i = \mu + e_i$, $\mu = \mathbb{E}(Y)$. Mit $\mu = a$ kann man nun die Funktion $Q(\mu)$ minimalisieren. Man betrachtet demnach

$$Q(\mu) = \sum_{i=1}^{n} (y_i - \mu)^2,$$

und findet

$$\frac{dQ(\mu)}{d\mu} = -2 \sum_{i=1}^{n} (y_i - \mu).$$

Für $\mu = \hat{\mu}$ wird $dQ/d\mu = 0$, d.h.

$$\left.\frac{dQ(\mu)}{d\mu}\right|_{\mu=\hat{\mu}} = -2 \sum_{i=1}^{n} (y_i - \hat{\mu}) = 0,$$

woraus $\sum_i y_i = n\hat{\mu}$ oder

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

folgt. Die Kleinste-Quadrate-Schätzung für μ ist das arithmetische Mittel, und

$$Q(\hat{\mu}) = \sum_{iu=1}^{n} (y_i - \bar{y})^2$$

ist der kleinstmögliche Wert für Q. Offenbar ist $Q(\hat{\mu}) = ns_y^2$, so dass man mit dieser Schätzung die Varianz der y-Werte um die Schätzung $\bar{y}$ minimalisiert.

Für das Modell $y_i = bx_i + a + e_i$ erhält man bekanntlich die KQ-Schätzungen

$$\hat{b} = \frac{Kov(x, y)}{s_x^2}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x}, \tag{2.6.71}$$

und die Korrelation zwischen x und y ist $r_{xy} = \hat{b}s_x/s_y$; r_{xy} kann als KQ-Schätzung des Regressionsparameters b für den Fall, dass x und y standardisiert wurden, betrachtet werden. Da viele Größen der KTT als Korrelationen definiert sind, kann man sagen, dass die Methode der Kleinsten Quadrate zentral für die KTT ist. Es muß noch hinzugefügt werden, dass die spezielle Annahme einer bestimmen Verteilungsform für die y-Werte zunächst nicht erforderlich ist; eine solche Annahme wird erst nötig, wenn inferenzstatistische Aussagen über die Schätzungen gewünscht werden.

Die Anwendung der Methode der Kleinsten Quadrate setzt nicht die Annahme der Normalverteilung voraus, – diese Annahme kann sich aber, wenn sie berechtigt ist, als nützlich erweisen, wenn man etwas über die Verteilung der Schätzungen sagen möchte.

3. **Die Maximum-Likelihood-Methode (ML-Methode):** *Likelihood* ist zunächst nur ein anderer Ausdruck für *probability*, bedeutet also im Allgemeinen Sinn soviel wie Wahrscheinlichkeit. Schätzt man die freien Parameter nach dieser Methode, so bestimmt man die Parameter so, dass die Wahrscheinlichkeit der Daten, gegeben diese speziellen Parameterwerte, maximal wird. Die Idee dabei ist, dass im Allgemeinen eben das geschieht, was am wahrscheinlichsten ist. In R. A. Fishers (1922, p. 310) Worten: "The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed."

   Diese Methode kann auf Modelle der Form $y = \hat{y} + e$ angewendet werden, erfordert dann aber die Annahme einer Wahrschleinlichkeitsverteilung für y bzw. für die Fehler e. Generell folgt schon aus dem Ansatz der ML-Methode, dass die Annahme einer Wahrscheinlichkeitsverteilung Voraussetzung für die Anwendung der Methode ist, da man ja sonst keinen Ausdruck für die Wahrscheinlichkeit der Daten anschreiben könnte. Es muß sich nicht um die Normalverteilung handeln: es kann sich um die Binomial-, die Exponential- oder irgendeine andere Verteilung handeln. Das Interessante an der ML-Methode ist allerdings, dass sich zeigen läßt (vergl. Kendall & Stuart (1973), p. 240), dass die Schätzungen für die Parameter *asymptotisch normalverteilt* sind. Diese Eigenschaft von ML-Schätzungen sind natürlich von Bedeutung für inferenzstatistische Betrachtungen.

   Beispiel 2.6.1 Für eine Stichprobe mit n Messungen y_i gelte $y_i = \mu + e_i$, wobei die Fehler e_i normalverteilt mit dem Erwartungswert 0 und der Varianz σ^2 seien. Die y_i sind dann normalverteilt mit dem Erwartungswert μ und der Varianz σ^2. Gesucht sind die ML-Schätzungen für μ und σ^2. Es wird die Unabhängigkeit der y_i vorausgesetzt. Die Likelihood der Stichprobe ist dann

$$\begin{aligned} \mathcal{L}(y_1, \ldots, y_n) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma}\right)^2\right), \quad (2.6.72) \end{aligned}$$

   und die *log-Likelihood*

$$L = \lg \mathcal{L} = -n \log \sigma - \frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma}\right)^2. \quad (2.6.73)$$

Die partiellen Ableitungen nach μ und σ sind

$$\frac{\partial L}{\partial \mu} = \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)\frac{1}{\sigma} \tag{2.6.74}$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\mu)^2 \tag{2.6.75}$$

Setzt man hier $\partial L/\partial\mu|_{\mu=\hat{\mu}} = 0$, so ergibt sich $\sum_i(y_i-\hat{\mu})/n = 0$ und damit $\hat{\mu} = \sum_i y_i/n$. Setzt man $\partial L/\partial\sigma|_{\sigma=\hat{\sigma}} = 0$, so ergibt sich $\hat{\sigma}^2 = \sum_i(y_i-\bar{y})^2/n$, wobei für μ bereits die Schätzung $\bar{y}$ eingesetzt wurde; zusammengefasst hat man mithin die ML-Schätzungen

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{2.6.76}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{2.6.77}$$

□

Die Maximum-Likelihood-Methode (ML-Methode) ist insbesondere dann eine mögliche Methode, wenn ein Modell mit additivem Fehler der Form $y = \hat{y} + e$ nicht postuliert wird oder nicht postuliert werden kann, dafür aber die Wahrscheinlichkeitsverteilung bekannt ist.

Beispiel 2.6.2 Die zufällige Veränderliche X sei binomialverteilt mit den Parametern p und n, so dass

$$P(X=k) = \binom{n}{k}p^k(1-p)^{n-k}. \tag{2.6.78}$$

Die Daten liegen als Folge von Werten einer Indikatorvariable $X_i = \{0,1\}$ vor: beim i-ten Versuch ist entweder $X_i = 1$ oder $X_i = 0$. Man hat etwa

$$X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1, \ldots$$

Die Likelihood dieser Daten ist dann das Produkt

$$\mathcal{L}(X_1,\ldots,X_n) = p(1-p)(1-p)p\cdots$$

Da es auf die Reihenfolge beim Multiplizieren nicht ankommt, kann man, wenn es insgesamt k "Erfolge" gegeben hat,

$$\mathcal{L}(X_1,\ldots,X_n) = p^k(1-p)^{n-k} \tag{2.6.79}$$

schreiben, und die Log-Likelihood ist

$$L = \log\mathcal{L} = k\log p + (n-k)\log(1-p). \tag{2.6.80}$$

L – und damit $\mathcal{L}$ – wird maximal, wenn $dL/dp|_{p=\hat{p}} = 0$. Es ist

$$\frac{dL}{dp} = \frac{k}{p} - \frac{n-k}{1-p},$$

und

$$\frac{k}{\hat{p}} - \frac{n-k}{1-\hat{p}} = 0.$$

Nach $\hat{p}$ aufgelöst ergibt sich daraus die ML-Schätzung

$$\hat{p} = \frac{k}{n}; \tag{2.6.81}$$

die ML-Schätzung für p ist also gerade durch die relative Häufigkeit der "Erfolge" gegeben. Formal entspricht diese Schätzung wieder dem arithmetischen Mittel der X_i-Werte, und damit entspricht die Schätzung dem Resultat, das man bei Anwendung der Momentenmethode bekommt.

$\hat{p}$ ist nicht notwendig der wahre Wert von p, – es ist eben nur der Schätzwert mit maximaler Wahrscheinlichkeit. Man kann sich einen Überblick über die Verteilung möglicher p-Werte verschaffen indem man Bayes Theorem anwendet. Demnach kann man

$$P(p|x) = P(x|p)\frac{P(p)}{P(x)} \tag{2.6.82}$$

schreiben: $P(p|x)$ ist die a posteriori-Wahrscheinlichkeit für den Wert p, $P(p)$ ist die a priori-Wahrscheinlichkeit für den Wert p, $P(x|p)$ ist die Wahrsfheinlichkeit der Daten x unter der Bedingung, dass der Binomialparameter den Wert p hat, und $P(x)$ ist die Wahrscheinlichkeit, dass die Daten x beobachtet werden. $(x|p)$ ist die Likelihood der Daten x, gegeben p. Hat man n Verscuhe mit k "Erfolgen" gemacht, so ist

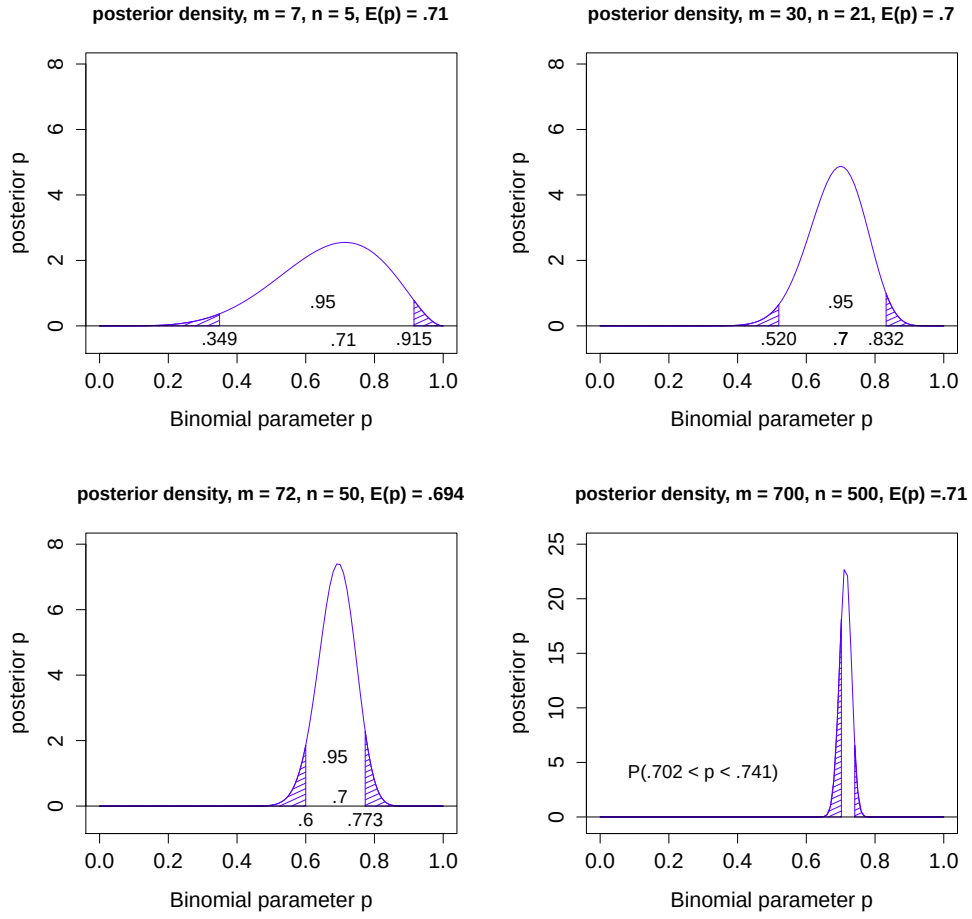
$$P(x|p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.6.83}$$

$P(p)$ muß durch eine Annahme festgelegt werden. Eine Möglichkeit ist, vorangegangene Schätzungen für p einzusetzen. Eine andere Möglichkeit ist, davon auszugehen, dass man nichts über p weiß, so dass man für $P(p)$ eine Gleichverteilung annimmt. Für $P(X)$ erhält man nach dem Satz über die Totale Wahrscheinlichkeit

$$P(x) = \int_0^1 P(x|p)P(p)dp = \binom{n}{k} \int_0^1 p^k (1-p)^{n-k} dp, \tag{2.6.84}$$

wobei der Gleichverteilungsannahme entsprechend $P(p) = 1$ gesetzt wurde. Man sieht, wie der Bereich, in dem der wahre Wert aller Wahrscheinlichkeit nach liegt, um so enger wird, je größer der Wert von n wird. □

Abbildung 11: A posteriori-Verteilungen für Binomialparameter $k/n \approx .7$; schraffierte Flächen: jeweils $p = .25$



Die ML-Methode führt nicht immer auf einfache Ausdrücke für die Schätzungen, wie anhand der logistischen Verteilung deutlich wird:

Beispiel 2.6.3 Die Stichprobe $(y_1, \ldots, y_n)$ komme aus einer logistisch verteilten Population, d.h.

$$f(y|\mu, \sigma) = \frac{1}{1 + \exp\left(-\frac{y-\mu}{\sigma}\sqrt{\frac{\pi}{3}}\right)} \tag{2.6.85}$$

(vergl. (2.6.23), Seite 45.) Kann die Unabhängigkeit der y_i-Werte vorausgesetzt werden, so ist die Likelihood der Daten durch

$$\mathcal{L}(y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{1 + \exp\left(-\frac{y_i-\mu}{\sigma}\alpha\right)}, \quad \alpha = \sqrt{\pi/3} \tag{2.6.86}$$

gegeben, und die Log-Likelihood durch

$$L = \log \mathcal{L} = -\sum_{i=1}^{n} \log\left(1 + \exp\left(\frac{y_i - \mu}{\sigma}\alpha\right)\right). \tag{2.6.87}$$

Setzt man zur Vereinfachung $A_i = \alpha(y_i - \mu)/\sigma$, so lassen sich die Ableitungen nach einem Parameter θ, wobei $\theta = \mu$ oder $\theta = \sigma$ sein kann, in der Form

$$\frac{\partial L}{\partial \theta} = -\sum_{i=1}^{n} \frac{1}{1 + e^{A_i}} \frac{\partial(1 + e^{A_i})}{\partial \theta} = -\sum_{i=1}^{n} \frac{e^{A_i}}{1 + e^{A_i}} \frac{\partial A_i}{\partial \theta} \tag{2.6.88}$$

schreiben. Insbesondere ist

$$\frac{\partial A_i}{\partial \mu} = -\frac{\alpha}{\sigma}, \quad \frac{\partial A_i}{\partial \sigma} = -\frac{\alpha(y_i - \mu)}{\sigma^2}.$$

L nimmt ein Maximum an für $\mu = \hat{\mu}$ und $\sigma = \hat{\sigma}$ und

$$\begin{aligned} \left.\frac{\partial L}{\partial \mu}\right|_{\mu=\hat{\mu}} &= \frac{\alpha}{\hat{\sigma}} \sum_{i=1}^{n} \frac{e^{\hat{A}_i}}{1 + e^{\hat{A}_i}} = 0 \\ \left.\frac{\partial L}{\partial \mu}\right|_{\mu=\hat{\sigma}} &= \frac{\alpha}{\hat{\sigma}^2} \sum_{i=1}^{n} \frac{e^{\hat{A}_i}(y_i - \hat{\mu})}{1 + e^{\hat{A}_i}} = 0 \end{aligned}$$

wobei $\hat{A}_i = \alpha(y_i - \hat{\mu})/\hat{\sigma}$. Da $\alpha/\hat{\sigma} \neq 0$ vorausgesetzt werden können, müssen die jeweiligen Summen gleich Null sein. Demnach müssen Schätzungen $\hat{\mu}$ und $\hat{\sigma}$ gefunden werden, für die die beiden Gleichungen

$$\sum_{i=1}^{n} \frac{e^{\hat{A}_i}}{1 + e^{\hat{A}_i}} = 0 \tag{2.6.89}$$

$$\sum_{i=1}^{n} \frac{e^{\hat{A}_i}(y_i - \hat{\mu})}{1 + e^{\hat{A}_i}} = 0 \tag{2.6.90}$$

gelten. Explizite Gleichungen für $\hat{\mu}$ und $\hat{\sigma}$ lassen sich nicht finden, d.h. die Gleichungen müssen numerisch gelöst werden.

Man kann auch von einer reparametrisierten Formel für die Dichte ausgehen. Denn es kann ja

$$\frac{\alpha(y_i - \mu)}{\sigma} = Ay_i + B$$

geschrieben werden, mit $A = \alpha/\sigma$, $B = -\alpha\mu/\sigma$. Die Log-Likelihood-Funktion hat dann die Form

$$L(A, B) = -\sum_{i=1}^{n} \log(1 + e^{Ay_i + B}). \tag{2.6.91}$$

Aus den partiellen Ableitungen nach A und B findet für die Schätzungen $\hat{A}$ und $\hat{B}$ die Gleichungen

$$\sum_{i=1}^{n} \frac{y_i e^{\hat{A}y_i+\hat{B}}}{1+e^{\hat{A}y_i+\hat{B}}} = 0 \tag{2.6.92}$$

$$\sum_{i=1}^{n} \frac{e^{\hat{A}y_i+\hat{B}}}{1+e^{\hat{A}y_i+\hat{B}}} = 0. \tag{2.6.93}$$

Hieraus lassen sich dann die Schätzungen $\hat{\sigma} = \alpha/\hat{A}$ für σ und $\hat{\mu} = -\hat{B}\hat{\sigma}/\alpha$ für μ ausrechnen. □

Eigenschaften von Schätzungen: Die Schätzungenvon Parametern – die sogenannten *Schätzer* – werden anhand der in der Stichprobe beobachteten Werte berechnet. In dem Maße, in dem diese Werte zufällig sind – jede Stichprobe ist ja eine zufällige Auswahl von Beobachtungen oder "Messungen" aus einer Population möglicher Werte – sind auch die Schätzungen zufällig, d.h. die Schätzungen $\hat{a}$, $\hat{b}$, ... sind zufällige Variablen, die ihrerseits wieder Erwartungswerte, Varianzen haben, und die Schätzungen verschiedener Parameter a, b etc können miteinander kovariieren. Die statistischen Eigenschaften der Schätzungen drücken sich in diesen Erwartungswerten, Varianzen und Kovarianzen aus und können von der Methode der Schätzung abhängen. Generell wird ein Parameter als eine *Statistik* über die gegebenen Messwerte berechenet, d.h. $\hat{\theta} = T(y_1, \ldots, y_n)$, wobei T die Funktion ist, die zur Berechnung von $\hat{\theta}$ angewendet wird. Soll also $\bar{y}$ als Schätzung von $\mathbb{E}(Y) = \mu$ berechnet weren, so ist

$$\hat{\mu} = T(y_1, \ldots, y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Die wichtigen Eigenschaften von Schätzungen sollen hier kurz aufgelistet werden, da sie bei der Diskussion einer Testtheorie von Bedeutung sind.

1. **Erwartungstreue:** Es sei θ ein zu schätzender Parameter und $\hat{\theta}$ eine Schätzung. Die Schätzung ist *erwartungstreu*, wenn

   $$\mathbb{E}(\hat{\theta}) = \theta. \tag{2.6.94}$$

   Dies heißt, dass die Schätzung $\hat{\theta}$ zwar vom wahren Wert θ abweichen kann, aber nicht systematisch um einen bestimmten Wert $\Delta\theta$ von θ abweicht. $\Delta\theta$ heißt *systematischer Fehler* oder auch kurz *Bias*, von engl. bias = systematischer Fehler.

   Das arithmetische Mittel $\bar{x}$ ist ein erwartungstreuer Schätzer des Erwartungswertes $\mathbb{E}(X)$ einer zufälligen Veränderlichen (sofern dieser überhaupt existiert), denn

   $$\mathbb{E}(\bar{x}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) = \frac{1}{n}n\mathbb{E}(X) = \mathbb{E}(X).$$

Nicht jeder Schätzer ist erwartungstreu:

Beispiel 2.6.4 die ML-Schätzung (2.6.77) der Varianz einer normalverteilten zufälligen Veränderlichen ist nicht erwartungstreu. Um dies zu sehen, kann man den Ausdruck für $\hat{\sigma}^2$ wie folgt schreiben:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{E}(Y) + \mathbb{E}(Y) - \bar{y})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{E}(Y))^2 - (\bar{y} - \mathbb{E}(Y))^2 - 2\sum_{i=1}^{n}(y_i - \mathbb{E}(Y)(\bar{y} - \mathbb{E}(Y)\end{aligned}$$

schreiben, wobei sich leicht zeigen läßt, dass das Kreuzprodukt auf der rechten Seite stets gleich Null sein muß. Dann ist

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(y_i - \mathbb{E}(Y))^2 + \mathbb{E}[(\bar{y} - \mathbb{E}(Y))^2].$$

Es ist aber

$$\mathbb{E}(y_i - \mathbb{E}(Y))^2 = \sigma^2, \quad \mathbb{E}[(\bar{y} - \mathbb{E}(Y))^2] = \frac{\sigma^2}{n},$$

$\sigma^2 = \mathbb{V}(X)$, und der zweite Term ist ja gerade die Varianz von $\bar{y}$, die bekanntlich gleich σ^2/n ist. Also hat man

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{n}n\sigma^2 - \frac{\sigma^2}{n} = \left(1 - \frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2.$$

Da $1 - 1/n < 1$ für $n \geq 1$ folgt, dass $\hat{\sigma}^2$ *im Durchschnitt* — also gemittelt über verschiedene Stichproben – kleiner als der wahre Wert σ^2 ist. Es folgt

$$\frac{n}{n-1}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right) = \sigma^2,$$

d.h. der Schätzer

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{2.6.95}$$

hat, im Unterschied zu s^2, keinen systematischen Fehler, ist also *biasfrei*.□

2. **Konsistenz:** Es ist intuitiv naheliegend, zu vermuten, dass eine Schätzung $\hat{\theta}$, die auf einem kleinen Stichprobenumfang n beruht, eine größere Varianz hat als eine Schätzung, die auf einem größeren Stichprobenumfang beruht. Man kann vermuten, dass die Schätzung $\hat{\theta}_n$, basierend auf einem Stichprobenumfang n, gegen θ strebt, wenn n größer wird. Insbesondere heißt ein Schätzer $\hat{\theta}$ *konsistent*, wenn

$$\lim_{n\to\infty}\mathbb{E}[(\hat{\theta}_n - \theta)^2] = 0, \tag{2.6.96}$$

wenn $\hat{\theta}_n$ also *quadratischen Mittel* gegen θ konvergiert. Dies bedeutet, dass (i) die Varianz von $\hat{\theta}_n$ und auch ein Bias b_n für wachsenden Wert von n gegen Null gehen. Dass man das quadratische Mittel und nicht einfach die Differenz $\hat{\theta}_n - \theta$ betrachtet beruht darauf, dass die Konvergenz im quadratischen Mittel gleichbedeutend damit ist, dass $\hat{\theta}_n$ in Wahrscheinlichkeit gegen θ konvergiert, d.h. wenn (2.6.96) gilt, so auch

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0, \text{ für beliebiges } \varepsilon. \tag{2.6.97}$$

Es läßt sich zeigen, dass die üblichen Schätzungen der Momente einer Verteilung, also $\bar{x}$ für $\mathbb{E}(X)$, s^2 bzw. $\hat{s}^2$ für σ^2, etc., konsistente Schätzungen für die Momente sind.

3. **Effizienz:** Es werde angenommen, dass zwei verschiedene Schätzmethoden zwei verschiedene Schätzungen $\hat{\theta}_1$ und $\hat{\theta}_2$ liefern. Die Schätzung $\hat{\theta}_1$ ist effizienter als die Schätzung $\hat{\theta}_2$, wenn

$$\mathbb{E}[(\hat{\theta}_1 - \theta)^2] \leq \mathbb{E}[(\hat{\theta}_2 - \theta)^2]. \tag{2.6.98}$$

Die Effizienz eines Schätzers kann, muß aber nicht vom Wert des Parameters θ abhängen. Man kann auch den Begriff der *relativen Effizienz* einführen: dies ist der Quotient

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{\mathbb{E}[(\hat{\theta}_1 - \theta)^2]}{\mathbb{E}[(\hat{\theta}_2 - \theta)^2]}. \tag{2.6.99}$$

Es läßt sich zeigen, dass im Falle erwartungstreuer Schätzer dieser Quotient gerade gleich dem Quotienten der Varianzen der beiden Schätzer ist; dann ist derjenige Schätzer effizienter als der andere, wenn er die kleinere Varianz $\mathbb{V}(\hat{\theta})$ hat; der effizienteste Schätzer ist dann der mit der minimalen Varianz.

4. **Suffizienz:** Wie oben angedeutet, wird eine Schätzung $\hat{\theta}$ als eine Funktion $T(y_1, \ldots, y_n)$ der Messungen $y_1, \ldots, y_n$ berechnet. Bildlich gesprochen, zieht T die Information über θ aus den Daten. Eine Statistik T zur Schätzung von θ heißt nun *suffizient* oder *erschöpfend*, wenn es keine andere Statistik T' gibt, die mehr Information bezüglich θ aus der Stichprobe zieht als T ("... the statistic chosen should summarise the whole of the relevant information supplied by the sample." Fisher, 1922, p. 316). Der Begriff der Suffizienz spielt eine zentrale Rolle in der Argumentation für das Rasch-Modell. Es ist deshalb sinnvoll, kurz auf die Bedingungen, unter denen Statistiken suffizient sein können, einzugehen.

   Zunächst muß eine formale Definition des Begriffes gegeben werden, die es ermöglicht, für eine gegebene Wahrscheinlichkeitsverteilung zu entscheiden, ob die Schätzung eines ihrer Parameter (oder aller Parameter) suffizient ist oder nicht. Die folgende Definition ist die übliche:

Definition 2.6.3 *Es sei $Y = (y_1, \ldots, y_n)$ eine gegebene Stichprobe von Messungen, und $T(y)$ sei eine Statistik für den Parameter θ. $T(Y)$ ist eine* suffiziente Statistik für θ, *wenn*

$$P(Y = y|T(Y) = t, \theta) = P(Y = y|T(Y) = t) \tag{2.6.100}$$

gilt.

Erläuterung: Ein gegebener Wert t von $T(Y)$ kann sich für verschiedene Stichproben ergeben; man denke an einen Mittelwert $\bar{y}$, der den gleichen Wert für ganz verschiedene Stichproben haben kann. Soll also $\bar{y}$ eine suffiziente Statistik für den Erwartungswert $\mathbb{E}(Y) = \mu$ sein, so soll die bedingte Wahrscheinlichkeit einer Stichprobe Y für gegebenen $\bar{y}$-Wert unabhängig von μ sein: die ganze Information über μ ist ja beeits in $\bar{y}$ enthalten, wenn $\bar{y}$ eine suffiziente Statistik für μ ist.

Beispiel 2.6.5 (Lindgren (1962), p. 191). Es sei $y_i = 1$ mit der Wahrscheinlichkeit p oder $y_i = 0$ mit der Wahrscheinlichkeit $1-p$, für $i = 1, \ldots, n$, und es sei $Y = y_1 + \cdots + y_n$. Y ist binomialverteilt. Die Wahrscheinlichkeit für $Y = y_i$ kann man in der Form

$$P(Y = y_i) = p^{y_i}(1-p)^{1-y_i}, \quad y_i = \{0, 1\}$$

schreiben. Eine Statistik bezüglich des unbekannten Parameters p ist

$$T(y_1, \ldots, y_n) = \sum_{i=1}^{n} y_i,$$

denn dann ist $T(y_1, \ldots, y_n) = k$, k die Anzahl der "Erfolge", und k/n als relative Häufigkeit dieser Erfolge kann ja als Schätzung für p gewählt werden. Natürlich gilt

$$P(T(y_1, \ldots, y_N) = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Man kann nun die bedingte Wahrscheinlichkeit, dass man die Stichprobe $(y_1, \ldots, y_n)$ erhält, gegeben dass $T = k$ ist, betrachten:

$$P(Y|T = k) = \frac{P(Y = (y_1, \ldots, y_n) \cap T = k)}{P(T = k)}.$$

Ist nun für die *gefundene* Stichprobe $\sum_i y_i \neq k$, so ist $Y = (y_1, \ldots, y_n) \cap T = k = \emptyset$ und damit ist $P(Y = (y_1, \ldots, y_n) \cap T = k) = 0$. Ist aber $\sum_i y_i = k$, so ist

$$P(Y = (y_1, \ldots, y_n) \cap T = k) = p^k (1-p)^{n-k},$$

denn auf die Reihenfolge der Nullen und Einsen kommt es ja nicht an, und man erhält

$$P(Y|T=k) = \frac{p^k(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} = \frac{1}{\binom{n}{k}}. \tag{2.6.101}$$

Die bedingte Wahrscheinlichkeit $P(Y|T=k)$ ist offenbar unabhängig vom Parameter p.

Dies bedeutet, dass die Information über den unbekannten Parameter p, die in der Stichprobe Y enthalten ist, bereits in der Statistik $T = \sum_i y_i$ enthalten ist. Man kann demnach auch schreiben

$$P(Y|T=k,p) = P(Y|T=k), \tag{2.6.102}$$

d.h. kennt man T, so ist keine weitere Information über p nötig, um die Wahrscheinlichkeit von Y, gegeben den Wert von T, zu berechnen. In diesem Sinne ist die Statistik T *hinreichend* – d.h. eben suffizient – für die Bestimmung der Wahrscheinlichkeit der Stichprobe Y. □

Ob eine Statistik suffizient ist, hängt von der Wahrscheinlichkeitsverteilung der Messwerte ab. Der folgende Satz chrakterisiert die Verteilungen, die suffiziente Statistiken zulassen.

Satz 2.6.1 *(Fishers Faktorisierungstheorem) Es sei $T(Y)$ eine Statistik für einen Parameter θ einer Familie $f(Y;\theta)$ von Verteilungen (f kann diskret oder stetig sein). Notwendig und hinreichend für die Suffizienz von T ist die Bedingung*

$$f(Y;\theta) = g(T(Y),\theta)h(Y), \tag{2.6.103}$$

d.h. die Verteilung kann in einen von T und θ abhängigen Term und in einen nur von Y abhängigen Term faktorisiert werden.

Der Satz soll hier nicht in aller Allgemeinheit bewiesen werden, es genügt, ihn zu illustrieren. Dazu werde angenommen, dass Y diskret ist[25]. Es sei $P_\theta(Y=y)$ die Wahrscheinlichkeit, dass die spezielle Stichprobe y gezogen wird, gegeben der Parameter θ. P_θ sei faktorisierbar im Sinne von (2.6.103), so dass

$$P_\theta(Y=y) = g[T(y),\theta]h(y). \tag{2.6.104}$$

Für einen speziellen Wert T_0 von T; also $T(y)=T_0$, gilt dann

$$\begin{aligned} P(T=T_0) &= \sum_{\forall y, T(y)=T_0} P_\theta(Y=y) = \sum_{\forall y, T(y)=T_0} g(T(y),\theta)h(y) \\ &= g(T_0,\theta) \sum_{\forall y, T(y)=T_0} h(y), \end{aligned} \tag{2.6.105}$$

[25]Damit wird vermieden, auf Fragen der Integrierbarkeit eingehen zu müssen.

wobei $\forall$ für "für alle" steht, d.h. hier, dass über alle y summiert werden soll, für die $T(y) = T_0$ ist. Man betrachtet die bedingte Verteilung von Y, gegeben $T(y) = T_0$:

$$P(Y|T_0) = \frac{P(Y = y \cap T = T_0)}{P(T = T_0)} = \begin{cases} 0, & T \neq T_0 \\ P(Y = y)/P(T = T_0), & T = T_0 \end{cases} \tag{2.6.106}$$

Für den Quotienten $P(Y = y)/P(T = T_0)$ findet man

$$\frac{P(Y = y)}{P(T = T_0)} = \frac{g(T_0, \theta)}{g(T_0, \theta) \sum_{\forall y, T(y)=T_0} h(y)} = \frac{1}{\sum_{\forall y, T(y)=T_0} h(y)},$$

d.h. der Quotient ist unabhängig von θ. Die Faktorisierung (2.6.104) ist also eine *hinreichende* Bedingung für die Suffizienz.

Nun sei umgekehrt $P(Y = y|T(y) = T_0)$ unabhängig von θ. Dann kann man

$$P(Y = y|T(y) = T_0) = \frac{P(Y = y \cap T = T_0)}{P(T = T_0)} = a(y, T_0)$$

schreiben, wobei a eine Funktion nur von y und T_0, nicht aber von θ ist. Es ist $P(Y = y) = P(Y = y \cap T(y) = T_0)$, d.h. aber

$$P(Y = y) = P(Y = y \cap T(y) = T_0) = a(y, T_0)P(T = T_0).$$

Der erste Faktor auf der rechten Seite – $A(y, T_0)$ – ist unabhängig von θ, und der zweite Faktor hängt ab von y nur über die Statistik $T(y)$. Also bedeutet die Unabhängigkeit der bedingten Wahrscheinlichkeit $P(Y = y|T(y) = T_0)$ von θ, dass sie als Produkt im Sinne von (2.6.103) angeschrieben werden kann; diese Faktorisierung ist, da sie von der Unabhängigkeit von θ impliziert wird, auch eine *notwendige* Bedingung. □

Die folgenden allgemeinen Betrachtungen erweisen sich als nützlich, wenn man überprüfen will, ob ein Testmodell suffiziente Schätzungen für die Parameter des Modells liefert.

Dazu $X_g = \{0, 1\}$ sei wieder eine Indikatorvariable: wenn $X_g = 1$, ist das g-te Item gelöst oder beantwortet worden, und wenn $X_g = 0$, dann nicht. Es ist $P(X_g = 1|\theta) = F_g(\theta)$, $P(X_g = 0|\theta) = 1 - 1 - F_g(\theta)$. Die Wahrscheinlichkeitsverteilung für X_g kann dann in kompakter Form wie

$$P(X_g|\theta) = F_g^{X_g}(1 - F_G)^{1-X_g}, \tag{2.6.107}$$

also $P(X_g|\theta) = F_g(\theta)$ für $X_g = 1$, und $P(X_g|\theta) = 1 - F_g(\theta)$ für $X_g = 0$. Es sei $\vec{x} = (X_1, \ldots, X_n)'$ ein Antwortvektor. Die Wahrscheinlichkeit, d.h. die Likelihood, für $\vec{x}$ ist dann für einen gegebenen θ-Wert, logale Unabhängigkeit vorausgesetzt,

$$P(\vec{x}|\theta) = \prod_{i=1}^{n} P(X_g(\theta)). \tag{2.6.108}$$

Nun ist

$$P(X_g|\theta) = F_g^{X_g}(1-F_G)^{1-X_g} = F_g^{X_g}\frac{1-F_g}{(1-F_g)^{X_g}} = (1-F_g)\frac{F_g^{X_g}}{(1-F_g)^{X_g}},$$

also

$$P(\vec{x}|\theta) = \prod_{g=1}^{n} P_g(X_g|\theta) = \prod_{g=1}^{n}\left((1-F_g)\frac{F_g^{X_g}}{(1-F_g)^{X_g}}\right) \tag{2.6.109}$$

Beispiel 2.6.6 (Logistisches Modell) Es werde speziell das logistische Modell (2.6.6), also

$$F_g(\theta) = \frac{e^{\alpha_g(\theta-b_g)}}{1+e^{\alpha_g(\theta-b_g)}}, \quad \alpha_g > 0.$$

angenommen. Es ist

$$1-F_g(\theta) = 1 - \frac{e^{\alpha_g(\theta-b_g)}}{1+e^{\alpha_g(\theta-b_g)}} = \frac{1+e^{\alpha_g(\theta-b_g)}-e^{\alpha_g(\theta-b_g)}}{1+e^{\alpha_g(\theta-b_g)}} = \frac{1}{1+e^{\alpha_g(\theta-b_g)}}.$$

Die Gleichung (2.6.109) kann in der Form

$$P(\vec{x}|\theta) = \prod_{g=1}^{n}(1-F_g)\prod_{g=1}^{n}\frac{F_g^{X_g}}{(1-F_g)^{X_g}} = \prod_{g=1}^{n}(1-F_g)\prod_{g=1}^{n}\left(\frac{F_g}{1-F_g}\right)^{X_g}$$

geschrieben werden. Für das logistische Modell hat man aber

$$\frac{F_g}{1-F_g} = \frac{e^{\alpha_g(\theta-b_g)}}{1+e^{\alpha_g(\theta-b_g)}}(1+e^{\alpha_g(\theta-b_g)}) = e^{\alpha_g(\theta-b_g)},$$

so dass

$$\begin{aligned} P(\vec{x}|\theta) &= \prod_{g=1}^{n}(1-F_g)\prod_{g=1}^{n}\exp\left(e^{\alpha_g(\theta-b_g)}\right)^{X_g} \\ &= \left(\prod_{g=1}^{n}(1-F_g)\right)\exp\left(\theta\sum_{g=1}^{n}\alpha_g X_g\right)\exp\left(\sum_{g=1}^{n}\alpha_g b_g X_g\right) \end{aligned} \tag{2.6.110}$$

$P(\vec{x}|\theta)$ genügt offenbar Fishers Faktorisierungstheorem, mit

$$g[T(Y),\theta] = \left(\prod_{g=1}^{n}(1-F_g)\right)\exp\left(\theta\sum_{g=1}^{n}\alpha_g X_g\right), \quad h(Y) = \exp\left(\sum_{g=1}^{n}\alpha_g b_g X_g\right),$$

wenn man $T(Y) = \sum_{g=1}^{n}\alpha_g X_g$ setzt. □

2.6.11 Sensitivität, Spezifität, Signalentdeckung

Es werden dichotome Merkmale betrachtet. Eine Diagnose ist eine Form der Signalentdeckung: Es soll festgestellt werden, ob bei einer Person ein Merkmal $\mathcal{M}$ (das "Signal") vorhanden ist ($\mathcal{M}_+$) oder nicht ($\mathcal{M}_-$), und gegebenenfalls soll noch die Ausprägung, in der es vorhanden ist, bestimmt werden. Ein häufiges Problem dabei ist, dass das "Signal" in ein "Rauschen" eingebettet ist, d.h. in Komponenten, die zufällig von einer Situation zur nächsten variieren – dies ist das *detection of signals in noise*-Problem, demzufolge das Entdecken des "Signals" ein probabilistischer Prozess ist, der den deterministischen Fall als Spezialfall enthält. Die Möglichkeit, dass zu beurteilende Merkmale fehlerhaft in Bezug auf Entscheidungskriterien beurteilt werden, kann ebenfalls als ein Effekt von "Rauschen" im kognitiven System des Beurteilers gesehen werden. Zunächst wird nur der einfache, dichotome Fall betrachtet (vergl. Tabelle 5). Es sei T der Test, mit dem das Vorhandensein von $\mathcal{M}$ geprüft wird, T_+ sei das Ereignis, dass der Test das Vorhandensein von $\mathcal{M}$ signalisiert, T_- sei das Ereignis, dass der Test das Fehlen des Merkmals $\mathcal{M}$ anzeigt. $\mathcal{M}_+$ bedeutet, dass in einem vorliegenden Fall das Merkmal $\mathcal{M}$ vorhanden ist, $\mathcal{M}_-$ bedeutet, dass es nicht vorhanden ist. Allgemein ist die (bedingte) Wahrscheinlichkeit eines Testausgangs T unter der Bedingung $\mathcal{M}$ wieder durch

$$P(T|\mathcal{M}) = \frac{P(T \cap \mathcal{M})}{P(\mathcal{M})} \tag{2.6.111}$$

gegeben, also hat man zB für $T = T_+$ und $\mathcal{M} = \mathcal{M}_+$ bzw. $T = T_-$, $\mathcal{M} = \mathcal{M}_-$

$$P(T_+|\mathcal{M}_+) = \frac{P(T_+ \cap \mathcal{M}_+)}{P(\mathcal{M}_+)}, \quad P(T_-|\mathcal{M}_-) = \frac{P(T_- \cap \mathcal{M}_-)}{P(\mathcal{M}_-},$$

etc. Weiter ist

$$P(T_+) = P((T_+ \cap \mathcal{M}_+) \cup (T_+ \cap \mathcal{M}_-)) = P(T_+ \cap \mathcal{M}_+) + P(T_+ \cap \mathcal{M}_-),$$

da sich die Ereignisse $T_+ \cap \mathcal{M}_-$ und $T_+ \cap \mathcal{M}_-$ ja ausschließen; $\cup$ steht dabei für "oder". Aus (2.6.111) folgt $P(T|\mathcal{M})P(\mathcal{M}) = P(T \cap \mathcal{M})$ d.h. man erhält die "totale Wahrscheinlichkeiten" $P(T_+)$ und $P(T_-)$ gemäß

$$P(T_+) = P(T_+|\mathcal{M}_+)P(\mathcal{M}_+) + P(T_+|\mathcal{M}_-)P(\mathcal{M}_-) \tag{2.6.112}$$
$$P(T_-) = P(T_-|\mathcal{M}_+)P(\mathcal{M}_+) + P(T_-|\mathcal{M}_-)P(\mathcal{M}_-) \tag{2.6.113}$$

In Tabelle 5 werden die verschiedenen Fälle zusammengefasst. Die hier auftretenden bedingten und absoluten Wahrscheinlichkeiten (d.h. die $P(T_+)$ und $P(T_-)$) können aus den Werten der Randverteilungen der Tabelle 5, analog zu denen der Tabelle (2.7.1), geschätzt werden. Die auf der rechten Seite von (2.6.112) erscheinende bedingte Wahrscheinlichkeit $P(T_+|\mathcal{M}_+)$ und die auf der rechten Seite von (2.6.113) erscheinende bedingte Wahrschinlichkeit $P(T_-|\mathcal{M}_-)$ haben in der Theorie der Diagnostik besondere Namen erhalten, die zunächst eingeführt werden sollen.

Tabelle 5: Sensitivität und Spezifität im dichotomen Fall

	$\mathcal{M}$		
T	$\mathcal{M}_+$	$\mathcal{M}_-$	Σ
T_+	a	b	$a+b$
T_-	c	d	$a+d$
Σ	$a+c$	$b+d$	N

1. **Sensitivität:** sie ist die Wahrscheinlichkeit, dass $\mathcal{M}$ entdeckt wird, wenn bei einer Person $\mathcal{M}$ auch tatsächlich vorliegt. Demnach ist die Sensitivität die bedingte Wahrscheinlichkeit $P(T_+|\mathcal{M}_+)$. Zur Abkürzung wird dafür der Buchstabe η verwendet:
$$\eta := P(T_+|\mathcal{M}_+) \tag{2.6.114}$$
definiert. Besteht der Test in der Beurteilung durch eine Person A, so wird für T_+ einfach A_+ geschrieben. In Bezug auf die Tabelle 5 hat man die Schätzung
$$\hat{\eta} = \frac{a}{a+c}. \tag{2.6.115}$$

2. **Spezifität:** Sie ist die Wahrscheinlichkeit, dass $\mathcal{M}$ als *nicht* vorhanden diagnostiziert wird, wenn $\mathcal{M}$ tatsächlich nicht vorhanden ist. Repräsentiert T_- das Ereignis, dass $\mathcal{M}$ als nicht vorhanden diagnostiziert wird, so ist also die Spezifität durch die bedingte Wahrscheinlichkeit
$$\vartheta := P(T_-|\mathcal{M}_-) \tag{2.6.116}$$
definiert. Besteht der Test in der Beurteilung durch eine Person A, so wird für T_- einfach A_- geschrieben. Wieder in Bezug auf die Tabelle 5 hat man die Schätzung
$$\hat{\vartheta} = \frac{d}{b+d}. \tag{2.6.117}$$

3. **Prävalenz:** Die Prävalenz ist der Anteil, mit dem das Merkmal $\mathcal{M}$ überhaupt in der betrachteten Population vorkommt, sie entspricht also der Wahrscheinlichkeit $P(\mathcal{M}_+)$; im folgenden wird zur Abkürzung θ für die Prävalenz geschrieben, also $\theta = P(\mathcal{M}_+)$. In Bezug auf die Tabelle 5 erhält man eine Schätzung
$$\hat{\theta} = \frac{a+c}{N}. \tag{2.6.118}$$

In der neuen Notation hat man also für $P(T_+)$ und $P(T_-)$ wegen $P(T_-|\mathcal{M}_+) = 1-\eta$, $P(T_+|\mathcal{M}_-) = 1-\vartheta$

$$P(T_+) = \eta\theta + (1-\vartheta)(1-\theta), \tag{2.6.119}$$
$$P(T_-) = (1-\eta)\theta + \vartheta(1-\theta). \tag{2.6.120}$$

Predictive Values: Bisher sind nur die Wahrscheinlichkeiten verschiedener Urteile betrachtet worden. Es sind aber noch zwei weitere Wahrscheinlichkeiten von Interesse: (a) die Wahrscheinlichkeit, dass eine zufällig gewählte Person oder ein zufällig gewähltes Objekt das Merkmal, das bei ihr oder ihm als vorhanden getestet wurde, auch tatsächlich hat; man spricht vom *positive predictive value* (PPV), $P(\mathcal{M}_+|T_+)$, und (b) die Wahrscheinlichkeit, dass eine zufällig gewählte Person oder ein zufällige gewähltes Objekt das Merkmal *nicht* hat, wenn es negativ getestet wurde; die der *negative predictive value* (NPV), $P(\mathcal{M}_-|T_-)$. Man hat

$$\begin{aligned} PPV = P(\mathcal{M}_+|T_+) &= P(T_+|\mathcal{M}_+)\frac{P(\mathcal{M}_+)}{P(T_+)} \\ &= \frac{\eta\theta}{\eta\theta + (1-\vartheta)(1-\theta)} \end{aligned} \tag{2.6.121}$$

$$\begin{aligned} NPV = P(\mathcal{M}_-|T_-) &= P(T_-|\mathcal{M}_-)\frac{P(\mathcal{M}_-)}{P(T_-)} \\ &= \frac{\vartheta(1-\theta)}{(1-\eta)\theta + \vartheta(1-\theta)} \end{aligned} \tag{2.6.122}$$

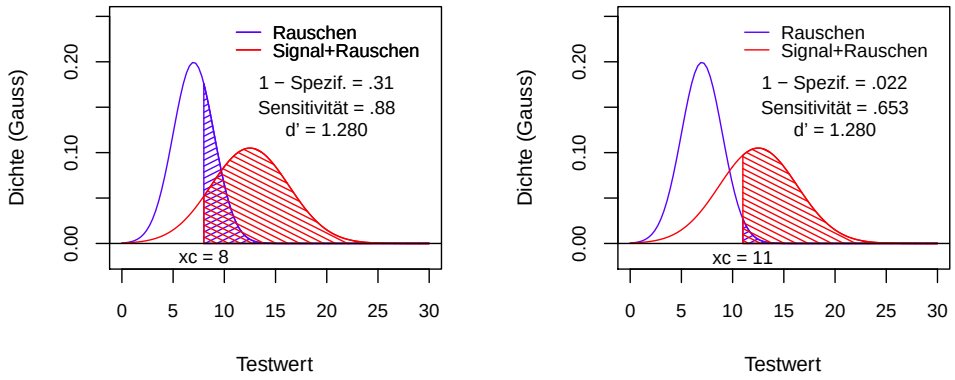
Diese Größen hängen jeweils von den drei Parametern Sensitivität η, Spezifität ϑ und Prävalenz θ ab, d.h. man hat zwei Gleichungen, aber drei Unbekannte. Dieser Sachverhalt verweist auf Probleme, die Parameter zu schätzen. Lösungen können gefunden werden, wenn mehr als nur zwei Urteiler betrachtet werden; eine Übersicht findet man bei Walter & Irvig (1988).

Theorie der Signalentdeckung und die ROC-Analyse: T und $\mathcal{M}$ sind bisher als dichotome Variable bzw. als Indikatorvariable betrachtet worden. Das mag in speziellen Fällen auch in dieser Direktheit so gelten: ein Tisch ist entweder rund ($\mathcal{M}_+$) oder eckig ($\mathcal{M}_-$), und ein Blick auf den Tisch stellt die Rundheit T_+ fest oder nicht (T_-). In diagnostischen Situationen, seinen sie nunr psychologischer oder medizinischer Art, geht es aber im Allgemeinen weniger direkt zu. Das zu diagnostizierende Merkmal kann bei einem Individuum mehr oder weniger ausgeprägt sein, und ebenso können die Symptome mehr oder weniger ausgeprägt sein. So sei X eine zufällige Variable, die die Ausprägung eines Symptoms repräsentiere. Es werde angenommen, dass das zu diagnostiziernde Merkmal in einem bestimmten Minimalausmaß vorhanden sein muß, damit es als existent betrachtet wird; diese Ausprägung sei $\mathcal{M}_+$; in diesem Fall existiert ein Signal, im anderen Fall nicht. Ist dieses Ausmaß nicht gegeben, so werde $\mathcal{M}_-$ kodiert. T sei eine Indikatorvariable:

$$T(x_0) = \begin{cases} T_-, & X \leq x_0 \\ T_+, & X > x_0 \end{cases} \tag{2.6.123}$$

X kann eine diskrete oder stetige Variable sein; hier wird der Kürze halber nur der stetige Fall betrachtet. Für X sei eine Dichtefunktion f erklärt, deren Parameter von der Ausprägung von $\mathcal{M}$ abhängen. In Anlehnung an die in der signalent-

Abbildung 12: Gauss-Verteilungen für X unter der Bedingung des reinen Rauschens ($\mu_1 = 7.0, \sigma_1^2 = 2.0$) und des Signals plus Rauschen ($\mu_2 = 12.5, \sigma_2^2 = 3.8$), für die Werte von $x_0 = 8$ und $x_0 = 11$. 1 - Spezifität = p(fA), Sensitivität = p(hit).



deckungstheorie üblichen Bezeichnungen werde wegen der größeren Übersichtlichkeit wegen die folgende Bezeichnung eingeführt:

$$n \text{ wenn } \mathcal{M} = \mathcal{M}_-, \quad sn(= s + n) \text{ wenn } \mathcal{M} = \mathcal{M}_+. \tag{2.6.124}$$

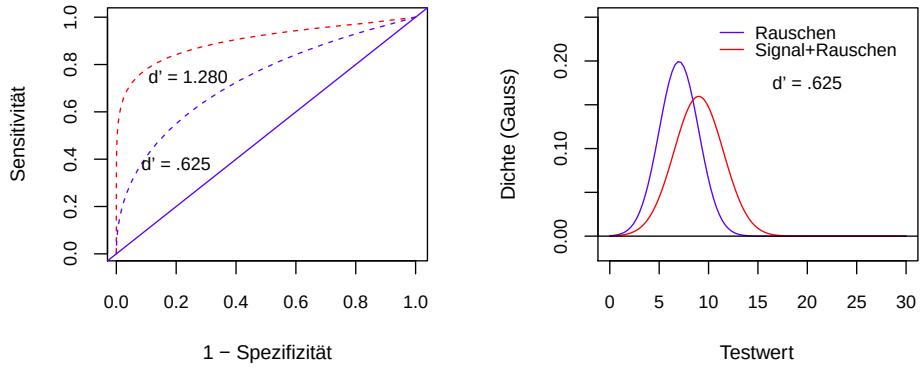
n steht hier für *noise*, als Rauschen, und sn für *signal plus noise*, also Rauschen plus Signal. Die Schreibweise sn für $\mathcal{M}_+$ und n für $\mathcal{M}_-$ erweist sich für die folgenden Betrachtungen als übersichtlicher als die bisherige. Für X werden nur Verteilungen betrachtet, die von zwei Parametern abhängen: dem Erwártungswert $\mu = \mathbb{E}(X)$, und der Varianz $\sigma^2 = Var(X)$, bzw. der Standardabweichung σ; insbesondere wird die Gauß-Verteilung und die Log-Normalverteilung betrachtet. X ist log-normalverteilt, wenn $\log X$ Gauß-verteilt ist. Es seien μ_0 und σ_0 die Parameter von f, wenn kein Signal vorhanden ist. Dann ist also $f(X) = f_n(X) = f(X; \mu_0, \sigma_0)$. Ist ein Signal vorhanden, so ist $f(X) = f_{sn}(X) = f(X; \mu_s, \sigma_s)$, wobei μ_s und σ_s der Erwartungswert und die Streuung von X sind, wenn das Signal vorhanden ist.

Da T gemäß (2.6.123) von x_0 abhängt, folgt, dass die Sensitivität $P(T_+|\mathcal{M}_+)$ und die Spezifität $P(T_-|\mathcal{M}_-)$ keine absoluten Größen sind. In Abhängigkeit von x_0 können diese beiden Größen alle Werte zwischen 0 und 1 annehmen. Der Fall, dass $T = T_+$, wenn $\mathcal{M} = \mathcal{M}_- =$ n, wird auch als "falscher Alarm" (fA) bezeichnet. Allgemein gilt

$$P(T_+|\text{sn}) + P(T_-|\text{sn}) = 1, \tag{2.6.125}$$

$$P(T_+|\text{n}) + P(T_-|\text{n}) = 1. \tag{2.6.126}$$

Abbildung 13: ROC-Kurven (links) und alternative n- und $s+n$-Verteilungen



$P(T_+|\text{sn})$ bezeichnet jetzt also die Sensitivität, und $P(T_-|\text{n})$ die Spezifität. Dann folgt

$$P(T_+|\text{n}) = 1 - P(T_-|\text{n}) = P(\text{fA}). \tag{2.6.127}$$

Wie das Beispiel in Abbildung 12 illustriert, bedeutet ein "kleiner" x_0-Wert zwar einen "großen" Sensitivitätswert, aber auch eine "große" Wahrscheinlichkeit für einen falschen Alarm. Umgekehrt impliziert ein "großer x_0-Wert eine kleine Wahrscheinlichkeit für einen falschen Alarm, aber auch eine reduzierte Sensitivität.

Man kann nun die Sensitivitätswerte gegen die zugehörigen Wahrscheinlichkeiten eines falschen Alarms, also gegen 1 - Spezifität für Werte von x_0 aus dem Bereich, auf dem x definiert ist, auftragen; es entsteht die *Receiver-Operating-Characteristic* oder kurz *ROC*-Kurve[26]. Die ROC-Kurve fällt mit der Diagonalen zusammen, wenn die Dichten von X für den Fall n des reinen Rauschens und des Falles sn zusammmenfallen; dann ist das Signal gleich Null (sn = n). Das Signal wird um so besser entdeckt werden, je verschiedener die Dichten von X für den Fall n einerseits und andererseits für den Fall sn sind. Dementsprechend ist "d-Strich"

$$d' = \frac{\mu_{sn} - \mu_n}{\sqrt{\sigma_{sn}^2 + \sigma_n^2}} \tag{2.6.128}$$

ein Maß für die Detektierbarkeit des Signals[27]. Je größer $\sqrt{\sigma_{sn}^2 + \sigma_n^2}$ relativ zur Differenz $\mu_{\text{sn}} - \mu_{\text{n}}$, desto mehr werden sich die Dichten f_{n} und f_{sn} überlappen

[26] Die Signal-Entdeckungstheorie wurde in der Nachrichtentechnik entwickelt; der Receiver ist der Empfänger des Signals, und es sind die Eigenschaften (operating characteristics) des Empfängers, die Wahrscheinlichkeit der Signalentdeckung bestimmen.

[27] d' wird auch Sensitivität genannt, aber diese Bezeichung führt zu Verwechslungen mit der durch $P(T_+|\mathcal{M}_+)$ definierten Sensitivität.

und umso kleiner wird die Detektierbarkeit d' sein, und umgekehrt wird die Detektierbarkeit um so größer sein, desto kleiner $\sqrt{\sigma_{sn}^2 + \sigma_n^2}$ relativ zur Differenz der Erwartungswerte ist.

In den Abbildungen (12) und (13) werden Gauß-Verteilungen gezeigt. Gauß-Verteilungen werden oft als "Normalverteilungen" bezeichnet,

Der Ausdruck "Normalverteilung"[28] ist oft eher irreführend, weshalb hier auch der Ausdruck "Gauß-Verteilung" verwendet wird; repräsentiert X insbesondere biologische Größen, so ist die log-Normalverteilung viel eher der Normalfall. Ist zum Beisliel klar, dass X keine negativen Werte annehmen kann, so kann die Annahme der Gauß-Verteilung zu Problemen führen, wenn der Erwartungswert klein relativ zur Varianz ist. Die sist zum Beispiel der Fall, wenn X PSA-Werte[29] gemessen werden. Die Erwartungswerte bei gesunden Personen liegen alterabhängig in der Nachbarschaft von 1 oder 2, während die Varianz groß genug sein muß, um auch Werte größer als 4 oder 5 zuzulassen. Die Gauß-Verteilung kann diesen Fall nicht adäquat abbilden, man müßte schon die *trunkierte Gauß-Verteilung* betrachten, bei der (in diesem Fall) Werte kleiner als Null die Wahrscheinlichkeit Null haben. Allerdings wird dann die Verteilung der Messwerte immer noch inadäquat beschrieben.

In der Signalentdeckungstheorie werden üblicherweise Gauß-Verteilungen angenommen, aber diese Annahme ist nicht notwendigerweise auch korrekt. In Egan (1975) wird eine Reihe alternativer Verteilungen diskutiert. Hier soll insbesondere die Lognormal-Verteilung illustriert werden (die icht in Egan (1975) behandelt wird). Die zufällige Veränderliche X heißt *log-normalverteilt*, wenn der (natürliche) Logarithmus von X normalverteilt ist[30]; für die Dichtefunktion von X hat man dann

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right), & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{2.6.129}$$

[28] Der Ausdruck 'Normalverteilung' geht wohl auf den belgischen Astronomen und Statistiker Lambert Adolphe Jacques Quételet (1796 bis 1874) zurück, der u.a. viele biometrische Messungen durchführte und dabei feststellte, dass die Messungen in guter Näherung symmetrisch und glockenförmig verteilt sind; dieser Befund führte zum Ausdruck 'Normalverteilung'. Diese Verteilung wurde zuerst 1730 von dem französischen Mathematiker Abraham de Moivre (1667 – 1754) als Grenzverteilung der Binomialverteilung für den Fall $p = 1/2$ hergeleitet, und Piere Simon de Laplace (1749 – 1827) zeigte 1812 die Gültigkeit dieser Herleitung für den allgemeinen Fall. Carl Friedrich Gauß (1777 – 1855) beschäftigte sich um 1795 mit der Methode der Kleinsten Quadrate (die unabhängig von Gauß durch den französischen Mathematiker Adrien-Marie Legendre (1752 – 1833) entwickelt worden war) und in diesem Zusammenhang ebenfalls mit der später nach ihm benannnten Gauß-Verteilung.

[29] Prostate-Specific Antigen; PSA-Werte werden u.A. bei Vorsorgeuntersuchungen zur frühzeitigen Entdeckung von Prostatakarzinomen erhoben.

[30] Als Rechtfertigung für die Annahhme der Gauß-Verteilung wird oft der Zentrale Grenzwertsatz angeführt, demzufolge die *Summme* unabhängiger Größen mit steigender Zahl der Summanden gegen eine Gauß-Verteilung strebt. Wird die gemessene Größe dagegen eher durch ein Produkt repräsentiert, so ist der Logarithmus der Größe durch die Summe der Logarithmen der Faktoren gegeben, auf die dann der Zentrale Grenzwertsatz angewendet werden kann, d.h. der Logarithmus der gemessenen Größe ist asymptotisch Gauß-verteilt.

μ ist der Erwartungswert, σ^2 ist die Varianz der $\log x$-Werte. Die Beziehung von μ und σ^2 zum Erwartungswert $\mathbb{E}(X)$ und zur Varianz $Var(X)$ von X ist durch die Gleichungen

$$\mathbb{E}(X) = e^{\mu+\sigma^2/2} \tag{2.6.130}$$

$$Var(X) = e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right) \tag{2.6.131}$$

gegeben. Der Maximalwert (Modus) von f wird für

$$x_{\max} = e^{\mu-\sigma^2} \tag{2.6.132}$$

angenommen. Die Frage ist, welcher cut-off-Punkt der gemessenen Variablen gewählt werden soll, um die Sensitivität und die Spezifität zu bestimmen. Dazu geht man davon aus, dass $P(T_+|\mathrm{n})$, $P(T_+|\mathrm{sn})$, $P(T_-|\mathrm{n})$ und $P(T_-|\mathrm{sn})$ bedingte Wahrscheinlichkeiten sind. $P(\mathrm{n})$ und $P(\mathrm{sn})$ sind die *a priori*-Wahrscheinlichkeiten, wobei natürlich $P(\mathrm{n}) = 1 - P(\mathrm{sn})$, desgleichen $P(\mathrm{sn}|x)$, $P(\mathrm{n}|x)$, etc. Allgemein gilt für $X = x$

$$P(\mathrm{sn}|x) = \frac{P(\mathrm{sn}\cap x)}{P(x)} = \frac{P(x|\mathrm{sn})P(\mathrm{sn})}{P(x)} \tag{2.6.133}$$

$$P(\mathrm{n}|x) = \frac{P(\mathrm{n}\cap x)}{P(x)} = \frac{P(x|\mathrm{n})P(\mathrm{n})}{P(x)} \tag{2.6.134}$$

wobei

$$P(\mathrm{sn}|x) + P(\mathrm{n}|x) = 1 \tag{2.6.135}$$

Division der Gleichung (2.6.133) durch die Gleichung (2.6.134) liefert den *Odds Ratio* (OR)

$$\frac{P(\mathrm{sn}|x)}{P(\mathrm{n}|x)} = \frac{P(\mathrm{sn}|x)}{1-P(\mathrm{sn}|x)} = \frac{P(x|\mathrm{sn})}{P(x|\mathrm{n})}\cdot\frac{P(\mathrm{sn})}{P(\mathrm{n})} \tag{2.6.136}$$

$P(x|\mathrm{sn})/P(x|\mathrm{n})$ ist der *Likelihood-Quotient*. Da X als stetige zufällige Veränderliche betrachtet wird, kann für $P(x|\mathrm{sn})$ auch $f(x|\mathrm{sn}) = f_{\mathrm{sn}}(x)$ geschrieben werden; analog dazu $P(x|\mathrm{n}) = f(x|\mathrm{n}) = f_{\mathrm{n}}(x)$, etc.

Es werden die bedingten Wahrscheinlichkeiten $P(T_+|\mathrm{sn})$ und $P(T_+|\mathrm{n})$ betrachtet:

$$P(T_+|\mathrm{sn}, x_0) = \int_{x_0}^{\infty} f(x|\mathrm{sn})dx \tag{2.6.137}$$

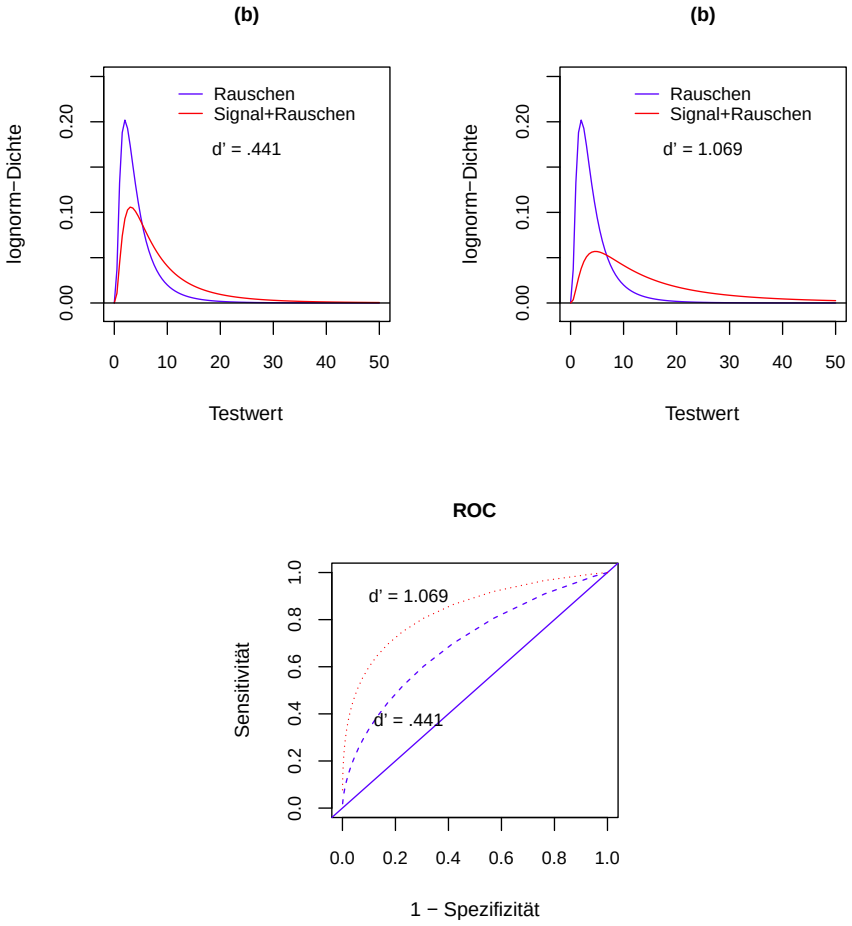
$$P(T_+|\mathrm{n}, x_0) = \int_{x_0}^{\infty} f(x|\mathrm{n})dx \tag{2.6.138}$$

Dann ist

$$\frac{dP(T_+|\mathrm{sn}, x_0)}{dx_0} = -f(x_0|\mathrm{sn}) \tag{2.6.139}$$

$$\frac{dP(T_+|\mathrm{n}, x_0)}{dx_0} = -f(x_0|\mathrm{n}) \tag{2.6.140}$$

Abbildung 14: Lognormalverteilungen; die "Rausch-Verteilung" (kein Signal) ist in (a) und (b) dieselbe: $\mathbb{E}(X) = 4.624$, $\sigma_1 = 4.018$. Die Rauschen+Signal-Verteilung in (a) hat den Erwartungswert und die Streuung $\mathbb{E}_1(X) = 9.127$, $\mathbb{E}_s(X) = 9.172$, $\sigma_2 = 9.395$ und in (b) $\mathbb{E}_3(X) = 28.574$, $\sigma_3 = 22.037$. Die Maxima der Dichten liegen bei $x_{max,1} = 1.988$, $x_{max,2} = 3.087$ und $x_{max,3} = 4.699$.



so dass

$$\frac{dP(T_+|\mathrm{sn}, x_0)}{dP(T_+|\mathrm{n}, x_0)} = \frac{-f(x_0|\mathrm{sn})}{-f(x_0|\mathrm{n})} = \frac{f(x_0|\mathrm{sn})}{f(x_0|\mathrm{n})} \tag{2.6.141}$$

$dP(T_+|\mathrm{sn}, x_0)/dP(T_+|\mathrm{n}, x_0)$ ist die Steigung der Tangente an der ROC-Kurve für $x = x_0$, und wegen (2.6.136) ist diese Steigung gerade gleich dem Likelihood-Quotienten $L(x_0)$ für den *cut off*-Wert x_0.

Man kann nach allgemeinen Entscheidungskriterien fragen; diese legen den Wert von x_0 fest. Das erste Kriterium ist einfach, die Warhscheinlichkeit einer

korrekten Entscheidung zu maximieren. C repräsentiere eine korrekte Entscheidung. Dann gilt

$$\begin{aligned} P(C) &= P(T_+|\mathrm{sn})P(\mathrm{sn}) + P(T_-|\mathrm{n})P(\mathrm{n}) \\ &= P(T_+|\mathrm{sn})P(\mathrm{sn}) + (1 - P(T_+|\mathrm{n})P(\mathrm{n}) \end{aligned} \tag{2.6.142}$$

$P(C)$ kann insbesondere als Funktion der Rate (Wahrscheinlichkeit) des falschen Alarms $(T_+|\mathrm{n})$ aufgefasst werden. Die Ableitung von $P(C)$ nach $P(T_+|\mathrm{n})$ ist dann

$$\frac{dP(C)}{dP(T_+|\mathrm{n})} = \frac{dP(T_+|\mathrm{sn})}{dP(T_+|\mathrm{n})}P(\mathrm{sn}) + \frac{d(1 - P(T_+|\mathrm{n}))}{dP(T_+|\mathrm{n})}P(\mathrm{n}) = \frac{dP(T_+|\mathrm{sn})}{dP(T_+|\mathrm{n})}P(\mathrm{sn}) + P(\mathrm{n}).$$

Setzt man diese Ableitung gleich Null, so erhält man unter Berücksichtigung von (2.6.141)

$$P(\mathrm{sn})L(x_0) = P(\mathrm{n}) \Rightarrow L(x_0) = \frac{P(\mathrm{n})}{P(\mathrm{sn})}, \tag{2.6.143}$$

d.h. man entscheidet nach Maßgabe des Quotienten der a-priori-Wahrscheinlichkeiten $P(\mathrm{n})$ und $P(\mathrm{sn})$. Ist $L(x) > P(\mathrm{n})/P(\mathrm{sn})$, so nimmt man an, dass ein Signal empfangen wurde, anderfalls folgert man, dass kein Signal empfangen wurde. Sind die a-priori-Wahrscheinlichkeiten bekannt, so maximiert diese Entscheidungsregel die Wahrscheinlichkeit, eine korrekte Entscheidung zu treffen.

Eine andere Entscheidungsregel kann angewendet werden, wenn der Nutzen bzw. die Kosten V – allgemein die *pay-offs* – der Entscheidung bekannt sind. Dafür betrachte man die folgende Tabelle: Der Erwartungswert des pay-offs V ist

Tabelle 6: Pay-offs für Entscheidungen

	Natur	
Entscheid.	sn	n
T_+	$V_{\mathrm{sn},+}$	$V_{\mathrm{n},+}$
T_-	$V_{\mathrm{sn},-}$	$V_{\mathrm{n},-}$

dann

$$\begin{aligned} \mathbb{E}(V) &= V_{\mathrm{sn},+}P(T_+|\mathrm{sn})P(\mathrm{sn}) + V_{\mathrm{n},-}P(T_+|\mathrm{n})P(\mathrm{n}) \\ &\quad + V_{\mathrm{n},-}P(T_-|\mathrm{n})P(\mathrm{n}) + V_{\mathrm{sn},-}P(T_-|\mathrm{sn})P(\mathrm{sn}) \end{aligned} \tag{2.6.144}$$

Gegeben sei der Wert $X = x$ der Testvariablen. Der Erwartungswert von V sei $\mathbb{E}(V|x, T_+)$, wenn angenommen wird, dass ein Signal vorliegt, und $\mathbb{E}(V|x, T_-)$, wenn angenommen wird, dass kein Signal vorliegt. Die Entscheidung soll nach Maßgabe des größeren Erwartungswertes, $\mathbb{E}(V|x, T_-)$ oder $\mathbb{E}(V|x, T_+)$, gefällt werden. Dann ist

$$\mathbb{E}(V|x, T_+) = V_{\mathrm{sn},+}P(\mathrm{sn}|x) + V_{\mathrm{n},T_+}P(\mathrm{n}|x) \tag{2.6.145}$$

$$\mathbb{E}(V|x, T_-) = V_{\mathrm{n},-}P(\mathrm{n}|x) + V_{\mathrm{sn},T_-}P(\mathrm{sn}|x) \tag{2.6.146}$$

Angenommen, man wählt T_+, wenn $\mathbb{E}(V|x,T_+) > \mathbb{E}(V|x,T_-)$, d.h. wenn $\mathbb{E}(V|x,T_+)/\mathbb{E}(V|x,T_-) > 1$. Division der ersten Gleichung durch die zweite liefert

$$\frac{\mathbb{E}(V|x,T_+)}{\mathbb{E}(V|x,T_-)} = \frac{V_{\mathrm{sn},+}P(\mathrm{sn}|x) + V_{\mathrm{n},+}P(\mathrm{n}|x)}{V_{\mathrm{n},-}P(\mathrm{n}|x) + V_{\mathrm{sn},-}P(\mathrm{sn}|x)} > 1,$$

und ein Rearrangement der Terme liefert die Entscheidungsregel

$$\text{Wenn } \frac{P(\mathrm{sn}|x)}{P(\mathrm{n}|x)} > \frac{V_{\mathrm{n},-} - V_{\mathrm{n},+}}{V_{\mathrm{sn},+} - V_{\mathrm{sn},-}}, \text{ dann } T_+, \text{ sonst } T_- \tag{2.6.147}$$

Wegen (2.6.136) hat man, mit $L(x) = P(x|\mathrm{sn})/P(x|\mathrm{n})$ der Likelihood-Quotient, die äquivalente Regel

$$L(x) > \frac{P(\mathrm{n})}{P(\mathrm{sn})} \cdot \frac{V_{\mathrm{n},-} - V_{\mathrm{n},+}}{V_{\mathrm{sn},+} - V_{\mathrm{sn},-}}, \tag{2.6.148}$$

ist also für gegebenen Testwert der Likelihood-Quotient $L(x)$ größer als Ausdruck auf der rechten Seite von (2.6.148), so entscheidet man sich für die Annahme, dass ein Signal vorliegt, andernfalls entscheidet man sich dagegen. Diese Entscheidung ist optimal, wenn die a priori-Wahrscheinlichkeiten bekannt sind. In der Medizin kann für $P(\mathrm{sn})$ die Inzidenzrate der Erkrankung ("Signal") angenommen werden, allerdings müssen dann immer noch die pay-offs V explizit bekannt sein. Bei bestimmten Erkrankungen kann man aber auf die Differenz $V_{\mathrm{sn},+} - V_{\mathrm{sn},-}$ fokussieren: die Kosten $V_{\mathrm{sn},-}$ können immens relativ zu den Kosten $V_{\mathrm{sn},+}$ sein, so dass auch kleine Werte von $L(x)$ eine Entscheidung für "Signal" (etwa: Tumor) nahelegen, – auch wenn die Kosten $V_{\mathrm{n},+}$ für einen falschen Alarm nicht gering sind.

Beziehung zur Theorie der statistischen Entscheidungen: Aus der Theorie der Signifikanztest sind zwei Arten von Fehlern bekannt: der Fehler erster Art oder α-Fehler – man entscheidet sich für die Alternativhypothese H_1, obwohl H_0 korrekt ist, und der Fehler zweiter Art oder β-Fehler – man entscheidet sich für H_0, obwohl H_1 korrekt ist. Üblicherweise wird nach Maßgabe des Wertes einer zufälligen Veränderlichen X entschieden: Ist etwa $X > x_0$, so entscheidet man für H_1, sonst für H_0. Bezeichnet man mit T_+ das zufällige Ereignis $X > x_0$, und mit T_- das Komplementärereignis $X \leq x_0$, so ist also

$$\alpha = P(T_+|H_0) \tag{2.6.149}$$
$$\beta = P(T_-|H_1) \tag{2.6.150}$$

$1 - \beta = P(T_+|H_1)$ ist die *Power* des Tests. Bezeichnet man mit T_+ das zufällige Ereignis, dass ein Signal registriert wird, und mit T_- das Komplementärereignis, das kein Signal registriert wurde, und bezeichnet s die Bedingung, dass tatsächlich ein Signal gesendet wurde, während $\neg s$ den Fall bezeichnet, dass kein Signal

gesendet wurde, so hat man die den Gleichungen (2.6.149) und (2.6.150) entsprechenden Gleichungen

$$\begin{array}{rcll} \alpha & = & P(T_+|\neg s), & (\text{"falsch positiv"}) \\ \beta & = & P(T_-|s), & (\text{"falsch negativ"}) \end{array} \tag{2.6.151}$$

Setzt man H_0 bzw $\neg s$ gleich $\mathcal{M}_-$ und H_1 bzw s gleich M_+, so ist klar, dass

$$\eta = 1-\beta, \ (\text{Sensitivität = Power}) \tag{2.6.152}$$

$$\vartheta = 1-\alpha, \ (\text{Spezifität}) \tag{2.6.153}$$

die Sensitivität entspricht also gerade der Power eines Tests und ist damit komplementär zur Wahrscheinlichkeit eines Fehlers zweiter Art bzw komplementär zur Wahrscheinlichkeit eines 'falsch negativen' Resultates ist, während die Spezifität komplementär zur Wahrscheinlichkeit eines Fehlers erster Art bzw komplementär zur Wahrscheinlichkeit eines Falschen Alarms ist.

2.7 Objektivität und Urteilerübereinstimmung

2.7.1 Urteilerübereinstimmnung

Im Allgemeinen Fall gebe es K Beurteilungskategorien $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$. Für einen gegebenen "Fall" (Objekt, Person, etc) wählt der Beurteiler A etwa die Kategorie $\mathcal{C}_i$, und Beurteiler B die Kategorie $\mathcal{C}_j$. Die Gesamthäufigkeit, mit der diese Kombination von Kategorisierungen vorkomme, sei n_{ij}. Diese Häufigkeiten können in einer Kontingenztabelle, die in diesem Zusammenhang auch *Übereinstimmungsmatrix* genannt wird, zusammenfassen:

	B				
A	$\mathcal{C}_1$	$\mathcal{C}_2$	$\cdots$	$\mathcal{C}_K$	$\sum$
$\mathcal{C}1$	n_{11}	n_{12}	$\cdots$	n_{2K}	n_{1+}
$\mathcal{C}_2$	n_{21}	n_{22}	$\cdots$	n_{2K}	n_{2+}
$\vdots$			$\vdots$		$\vdots$
$\mathcal{C}_K$	n_{K1}	n_{K2}	$\cdots$	n_{KK}	n_{K+}
$\sum$	n_{+1}	n_{+2}	$\cdots$	n_{+K}	$n_{++} = N$

(2.7.1)

Die Häufigkeiten $(n_{1+}, n_{2+}, \ldots, n_{K+})$ sind die Randverteilung für den Beurteiler A, und die Häufigkeiten $(n_{+1}, n_{+2}, \ldots, n_{+K})$ sind die Randverteilung für den Beurteiler B. Die Randverteilungen sind (im strengen Sinne) *homogen*, wenn $n_{k+} = n_{+k}$ für alle $k = 1, \ldots, K$.

Man kann die bedingte Wahrscheinlichkeit, dass der Beurteiler A die Kategorie $\mathcal{C}_j$ wählt $(A(\mathcal{C}_j))$ unter der Bedingung, dass der Beurteiler B die Kategorie $\mathcal{C}_k$ wählt $(B(\mathcal{C}_k))$, allgemein durch den Satz für bedingte Wahrscheinlichkeiten ausdrücken:

$$P(A(\mathcal{C}_j)|B(\mathcal{C}_k)) = \frac{P(A(\mathcal{C}_j) \cap B(\mathcal{C}_k))}{P(B(\mathcal{C}_k))}, \quad j,k = 1, \ldots K \tag{2.7.2}$$

wobei $\cap$ für "und" steht. Analog dazu gilt natürlich

$$P(B(\mathcal{C}_k)|A(\mathcal{C}_j)) = \frac{P(A(\mathcal{C}_j) \cap B(\mathcal{C}_k))}{P(A(\mathcal{C}_j))}. \tag{2.7.3}$$

Sofern die Urteiler unabhängig voneinander urteilen gilt natürlich

$$P(A(\mathcal{C}_j) \cap B(\mathcal{C}_k)) = P(A(\mathcal{C}_j))P(B(\mathcal{C}_k)) \tag{2.7.4}$$

und man erhält

$$P(A(\mathcal{C}_j)|B(\mathcal{C}_k)) = P(A(\mathcal{C}_j)), \quad P(B(\mathcal{C}_k)|A(\mathcal{C}_j)) = P(B(\mathcal{C}_k)). \tag{2.7.5}$$

Schätzungen für diese Wahrscheinlichkeiten lassen sich aus den Randhäufigkeiten der Kontingenztabelle gewinnen:

$$\hat{P}(A(\mathcal{C}_j)) = \frac{n_{j+}}{n_{++}}, \quad \hat{P}(B(\mathcal{C}_k)) = \frac{n_{+k}}{n_{++}}, \quad \hat{P}(A(\mathcal{C}_j) \cap B(\mathcal{C}_k)) = \frac{n_{jk}}{n_{++}} \tag{2.7.6}$$

sowie

$$\hat{P}(A(\mathcal{C}_j)|B(\mathcal{C}_k)) = \frac{n_{jk}}{n_{k+}}, \quad \hat{P}(B(\mathcal{C}_k)|A(\mathcal{C}_j)) = \frac{n_{jk}}{n_{+j}} \tag{2.7.7}$$

Unter der Voraussetzung, dass alle Kategorien bei den Beurteilungen gewählt werden, folgt $P(A(\mathcal{C}_j)) \neq 0$ und $P(B(\mathcal{C}_k)) \neq 0$ für alle j, k. Unabhängigkeit der Urteiler bedeutet dann aber insbesondere, dass

$$\hat{P}(A(\mathcal{C}_j \cap B(\mathcal{C}_k)) = \hat{P}(A(\mathcal{C}_j))\hat{P}(B(\mathcal{C}_k)) = \frac{n_{j+}n_{+k}}{n_{++}^2} \neq 0$$

für alle j, k. Es werde nun angenommen, dass die beiden Beurteiler unabhängig voneinander urteilen in dem Sinne, dass keiner der beiden Urteiler Informationen über die Urteile des anderen hat. Es ist aber denkbar, dass sie stets exakt nach den gleichen Kriterien urteilen und die Bewertung der Merkmale hinsichtlich dieser Kriterien stets identisch ausfällt. Dann folgt

$$P(A(\mathcal{C}_j) \cap B(\mathcal{C}_k)) = \begin{cases} 0, & j \neq k \\ \\ 1, & j = k \end{cases}$$

Andererseits sollte die Unabhängigkeit der Beurteiler nach (2.7.4)

$$P(A(\mathcal{C}_j) \cap B(\mathcal{C}_k)) \neq 0 \text{ für } j \neq k \tag{2.7.8}$$

implizieren. Der Widerspruch, der sich hier zu ergeben scheint, ist allerdings nur scheinbar. Denn da die Urteiler nach Voraussetzung exakt dieselben oder doch perfekt äquivalente Kriterien verwenden kann man das Urteil es einen Beurteilers auf der Bassis des anderen Urteilers voraussagen und *insofern* sind die Urteile abhängig voneinander, auch, wenn die Beurteiler nicht miteinander kommunizieren.

Übereinstimmung der Urteiler bedeutet also Abhängigkeit wegen des Gebrauchs derselben oder äquivalenter Kriterien, wobei eine perfekte Übereinstimmung im Sinne von (2.7.8) eher selten zu beobachten sein wird, weil eine perfekte, fehlerfreie Bewertung der Merkmale der Objekte in Bezug auf die Kriterien eher eine Ausnahme sein wird. Ein gute Urteilerübereinstimmung zeigt sich also, wenn die n_{jj} groß sind, wenn also die Masse der Häufigkeiten in der Diagonale der Tabelle 2.7.1 zu finden sind und für die meisten der $n_{jk} = 0$, $j \neq k$, gilt.

Urteilerübereinstimmung und Unabhängigkeit der Urteiler: Die meisten Maße für die Urteilerübereinstimmung hängen in der einen oder anderen Weise von diesen Parametern ab, d.h. sind Funktionen dieser Parameter. Im Wesentlichen wird stets der Anteil übereinstimmender Urteile zu dem Anteil, der bei rein zufälliger Übereinstimmung zu erwarten ist, in Beziehung gesetzt. In einem Großteil der Literatur zu diesen Maßen wird die Rolle inhomogener Randverteilungen sowie der Prävalenz diskutiert. Eine andere Frage ist die nach der Bedeutung des Ausdrucks 'zufällige Übereinstimmung'. Gemeinhin wird angenommen, dass die Wahrscheinlichkeit für das Ereignis $A_+ \cap B_+$ bei 'zufälliger Übereinstimmung' durch $P(A_+ \cap B_+) = P(A_+)P(B_+)$ gegeben ist und damit durch $n_{+k}n_{k+}/N$ abgeschätzt werden kann, denn die bekannte Formel für die bedingte Wahrscheinlichkeit impliziert

$$P(A_+ \cap B_+) = P(A_+|B_+)P(B_+) = P(A_+)P(B_+),$$

wenn $P(A_+|B_+) = P(A_+)$, das Urteil A_+ des ersten Urteilers also unabhängig vom Urteil B_+ des zweiten ist (dann ist auch das Urteil des zweiten unabhängig vom Urteil des ersten). Andererseits wird man bei einer Untersuchung zur Urteilerübereinstimmung geradezu fordern, dass A und B unabhängig voneinander diagnostizieren. Würden sie es nicht tun, wüßte man nicht, ob ihre Übereinstimmung ein Resultat dieser Abhängigkeit ist, die etwa in einer Diskussion eines Falles vor der jeweiligen Diagnose bestehen könnte, oder ob sie ein Resultat ähnlicher Interpretation der Symptome ist, die auch ohne jede Diskussion vor der Diagnose stattfindet. Hier wird argumentiert werden, dass die Unabhängigkeit der Urteiler *bedingte Unabhängigkeit* der Urteiler bedeutet:

$$P(A_+ \cap B_+|\mathcal{C}_+) = P(A_+|\mathcal{M}_+)P(B_+|\mathcal{M}_+) \tag{2.7.9}$$

(und analog dazu für alle anderen Konjunktionen) definiert ist. □

Prävalenz und Bias: Die Prävalenz $P(\mathcal{M}_+)$ ist zunächst eine objektive Größe. Allerdings muß sie den Urteilern nicht bekannt sein, und zwei verschiedene Beurteiler A und B können implizit verschiedene Annahmen über den Wert von $P(M_+)$ machen. Byrt, Bishop & Carlin (1993) haben dazu den Begriff des *Bias* eingeführt: der bestünde im Falle verschiedener Annahmen über die Prävalenz; denn er impliziere inhomogene Randverteilungen (P_{A+}, P_{A-}) und (P_{B+}, P_{B-}).

Der Bias könne durch einen *Bias-Index* ausgedrückt werden:

$$\mathrm{BI} = \frac{a+b}{N} - \frac{a+c}{N} = \frac{a-c}{N}, \tag{2.7.10}$$

N die Gesamtzahl der Fälle. Byrt et al.s Argument basiert allerdings nicht auf einer ausführlichen Diskussion der Urteilsmechanismen sondern scheint auf der Annahme zu basieren, dass die als subjektiv angenommenen θ-Werte der beiden Diagnostiker direkt aus der Tabelle abgelesen werden können. Es wird im Folgenden argumentiert werden, dass die möglicherweise existierenden subjektiven Annahmen über die Prävalenz alle in den Sensitivitäts- und Spezifitätsparametern absorbiert sind, so dass der Bias-Begriff von Byrt et al. nicht weiter diskutiert zu werden braucht; diese Diskussion klärt gleichzeitig die Begriffe der Sensitivität und Spezifität.

Zum Byrt et al.'schen Bias-Begriff Maße, die das Ausmaß an Urteilerübereinstimmung reflektieren sollen, gehen im Allgemeinen vom Anteil gemeinsamer positiver und negativer Beurteilungen aus und unterscheiden sich im Wesentlichen nur in der Art, in der der Anteil zufälliger Übereinstimmung herausgerechnet wird. Wie bei einem Diagnostiker ein positives oder negatives Urteil zustande kommt wird in den Gleichungen (2.6.119) und (2.6.120) angegeben. Formal sind T_+ und T_- zufällige Ereignisse – was manche Diagnostiker oder Diagnostikerinnen nicht gerne hören werden –, und die Gleichungen lassen auch den Spezialfall $\eta = \vartheta = 1$ zu, also perfekte Sensitivität und Spezifität. Dieser Fall ist der wichtige Spezialfall des sogenannten *Goldstandards*. Dies sind Diagnosen, deren Korrektheit nicht angezweifelt wird – in der Medizin sind dies häufig Diagnosen anhand eines histologischen Befundes. Auch histologische Befunde können fehlerbehaftet sein, aber man vernachlässigt die Möglichkeit von Fehlern, weil man keine bessere Diagnose hat. In diesem Fall ist $P(T_+) = \theta$, $P(T_-) = 1 - \theta$, d.h. die Wahrscheinlichkeit eines "positiven" Urteils T_+ ist durch die Wahrscheinlichkeit, überhaupt auf eine Person zu treffen, die das Merkmal $\mathcal{M}$ hat gegeben.

Die Gleichung (2.6.119) liefert die Basis für eine Diskussion des Effekts unterschiedlicher Randverteilungen und anderer Aspekte der Tabelle auf den Wert der Übereinstimmungsmaße. In Bezug auf Byrt et al.s Bias-Begriff stellt sich die Frage, welche der Größen in (2.6.119) subjektiv sind. Die Sensitivität η und die Spezifität ϑ sind Parameter, die den oder die jeweilige(n) Diagnostiker(in) charakterisieren; sie sind Maße, die die Genauigkeit der Wahrnehmung und/oder die Kenntnis über die Möglichkeiten bzw Symptome, mit denen $\mathcal{M}$ signalisiert wird abbilden, – insofern repräsentieren sie subjektive Aspekte des Urteilsprozesses. Die nächste Frage ist, ob die Prävalenz $\theta = P(\mathcal{M}_+)$ als objektive oder, wie von Byrt et al. postuliert, als subjektive Größe gedeutet werden muß.

Die Häufigkeiten in der Tabelle (2.7.1) enthalten typischerweise die Reaktionen der Diagnostiker D auf zu beurteilenden Personen, wie sie eben dem Beurteilungsprozess zugeführt werden: Die Auswahl dieser Personen gleicht formal dem

zufälligen Ziehen von Kugeln aus einer Urne, wobei die Kugeln entweder $\mathcal{M}_+$- oder $\mathcal{M}_-$-Kugeln sind. Die Wahrscheinlichkeit, eine Kugel mit $\mathcal{M}_+$ zu ziehen, entspricht (bei guter Durchmischung der Kugeln) dem Anteil der Kugeln in der Urne, die mit $\mathcal{M}_+$ gekennzeichnet sind. Dies bedeutet, dass $\theta = P(\mathcal{M}_+)$ eine objektive Größe ist. Die Frage ist nur, ob die Situation, in der diagnostiziert wird, dem Ziehen aus einer Urne entspricht, deren Zusammensetzung aus $\mathcal{M}_+$- und $\mathcal{M}_-$-Fällen dem der Population entspricht. Soll in einer schulpsychologischen Praxis geprüft werden, ob ein Schüler unter ADHS[31] leidet, so kommt der Schüler im Allgemeinen aus einer Teilpopulation irgendwie auffälliger Schüler, in der der Anteil ADHS-Kranker höher ist als in der Gesamtpopulation, und möglicherweise ist deshalb der θ-Wert, den die Schulpsychologin ihrer Diagnose zugrundelegt, höher als der der Gesamtpopulation. Analog ist die Situation in einer Arztpraxis, wo der Anteil der Patienten der Praxis, die etwa an Krebs leiden, höher ist als in der Gesamtpopulation. Der Arzt muß sich entscheiden, welchen Anteil er zugrunde legen soll, den Anteil der an Krebs Erkrankten in der Gesamtpopulation oder in der Teilpopulation derjenigen, die überhaupt an einer gesundheitlichen Störung leiden. Möglicherweise kommt auf diese Weise θ als subjektive Größe in die Diagnose.

Subjektive Einschätzungen von θ können allerdings unabhängig von der Frage der Teilpopulationen in den Sensitivitäten und Spezifitäten versteckt sein. Es werde die Sensitivität $\eta = P(T_+|\mathcal{M}_+)$ betrachtet. Dieser Parameter soll ausdrücken, wie ein Test – also D – reagiert, wenn eine Person mit dem Merkmal $\mathcal{M}_+$ bezüglich $\mathcal{M}$ zu diagnostizieren ist, ohne dass der Diagnostiker weiß, dass der Fall $\mathcal{M}_+$ vorliegt. Alles, was D für einen gegebenen Fall zur Verfügung hat, sind Symptome S. Die Wahrscheinlichkeit, dass D für diesen Fall $\mathcal{M}_+$ diagnostiziert, ist $P(\mathcal{M}_+|S)$. Sind die Symptome abzählbar, so kann insbesondere S_1,oder S_2 oder S_3 etc vorliegen und man hat, $S_j \cap S_k = \emptyset$ vorausgesetzt,

$$p(T_+|\mathcal{M}_+) = \sum_j P(\mathcal{M}_+|S_j)P(S_j|\mathcal{M}_+).$$

Kann man annehmen, dass ein Komplex von Symptomen durch einen Vektor $\mathbf{x} = (x_1, \ldots, x_n)'$ beschrieben werden kann, erhält man einen allgemeinen Ausdruck für die Sensitivität:

$$\eta = P(T_+|\mathcal{M}_+) = \int_X P(\mathcal{M}_+|\mathbf{x})P(\mathbf{x}|\mathcal{M}_+)d\mathbf{x} \tag{2.7.11}$$

Hier ist $P(\mathbf{x}|\mathcal{M}_+)$ eine objektive Größe insofern, als die Verteilung der Symptome nur von $\mathcal{M}_+$ und nicht von einem Diagnostiker abhängt.

Beispiel 2.7.1 $P(\mathcal{M}_+|\mathbf{x})$ kann subjektive Größen enthalten. Denn formal ist

$$P(\mathcal{M}_+|\mathbf{x}) = P_s(\mathbf{x}|\mathcal{M}_+)\frac{P_s(\mathcal{M}_+)}{P_s(\mathbf{x})} = P_s(\mathcal{M}_+|\mathbf{x}). \tag{2.7.12}$$

[31] ADHS = Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung

Die Wahrscheinlichkeiten auf der rechten Seite sind mit s für 'subjektiv' indiziert, denn sie reflektieren die Annahmen von D bezüglich des Auftretens von $\mathcal{M}_+$ und der Symptome $\mathbf{x}$, und damit ist $P(\mathcal{M}_+|\mathbf{x})$ selbst eine subjektive Größe. □

Ausdrücke wie (2.7.12) legen nahe, dass D sich bei ihrem Urteilsprozess nach Bayes' Theorem richtet, was nicht sein muß. Jedenfalls ist $P_s(\mathbf{x})$ eine subjektive Schätzung von $P(\mathcal{M}_+)$, die zur Abkürzung mit θ^s bezeichnet werden soll. Es zeigt sich aber, dass die implizite Annahme, dass D bayesianisch urteilt, nicht wesentlich ist. Dazu wende man den Satz der totalen Wahrscheinlichkeit auf $P_s(\mathbf{x})$ an:

$$P_s(\mathbf{x}) = P_s(\mathbf{x}|\mathcal{M}_+)\theta^s + P_s(\mathbf{x}|\mathcal{M}_-)(1-\theta^s).$$

Substituiert man diesen Ausdruck in (2.7.12) und dividiert man Zähler und Nenner der rechten Seite von (2.7.12) durch $P_s(\mathbf{x}|\mathcal{M}_+)P_s(\mathcal{M}_+)$, so erhält man

$$P(\mathcal{M}_+|\mathbf{x}) = \frac{1}{1+\lambda^{-1}}, \tag{2.7.13}$$

mit

$$\lambda = \lambda(\mathbf{x}) = \frac{P_s(\mathbf{x}|\mathcal{M}_+)\theta^s}{P_s(\mathbf{x}|\mathcal{M}_-)(1-\theta^s)} \tag{2.7.14}$$

Der natürliche Logarithmus (d.h. der zur Basis e) von λ^{-1} ist $y = \log_e \lambda^{-1} = -\log_e \lambda = -\log \lambda$ – der Index e kann zur Vereinfachung weggelassen werden. Durch Exponierung kann die Logarithmierung wieder rückgängig gemacht werden: $e^y = e^{-\log\lambda} = \lambda^{-1}$. Auf diese Weise erhält man erhält man für (2.7.13)

$$P(\mathcal{M}_+|\mathbf{x}) = \frac{1}{1+e^{-\log\lambda(\mathbf{x})}} \tag{2.7.15}$$

Wählt man die multivariate Normalverteilung für $\mathbf{x}$,

$$f(\mathbf{x}) = C\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

C eine Normierungskonstante und $\boldsymbol{\mu}$ ein Mittelwertsvektor, so ist

$$\log\lambda = (\mathbf{x}-\boldsymbol{\mu}_+)'\Sigma_+^{-1}(\mathbf{x}-\boldsymbol{\mu}_+) - (\mathbf{x}-\boldsymbol{\mu}_-)'\Sigma_-^{-1}(\mathbf{x}-\boldsymbol{\mu}_-),$$

wobei $\boldsymbol{\mu}_+$ der Mittelwertsvektor für $\mathbf{x}$ ist, wenn $\mathcal{M}_+$ vorliegt, und $\boldsymbol{\mu}_-$ der entsprechende Mittelwertsvektor ist, wenn $\mathcal{M}_-$ vorliegt. Σ_+ und Σ_- sind die Varianz-Kovarianz-Matrizen unter den Bedingungen $\mathcal{M}_+$ bzw. $\mathcal{M}_-$. Für $\Sigma_+ = \Sigma_-$ liefert (2.7.15) dann eine logistische Regression, die linear in den Komponenten von $\mathbf{x}$ ist, und für $\Sigma_+ \neq \Sigma_-$ hängt die logistische Regression quadratisch von den Komponenten von $\mathbf{x}$ ab. Den linearen Fall erhält man ebenfalls, wenn man annimmt, dass $\mathbf{x} = x$ nur eine einzige Komponente hat, die x logistisch verteilt ist. Dieser Fall ist mit dem Fall $\Sigma_+ = \Sigma_-$ vergleichbar, weil $\mathbf{x}$ nun linear eingeht; welches Modell tatsächlich vorliegt, kann aus den geschätzten Parametern nicht erschlossen werden.

Natürlich ist eine Vielzahl anderer Modelle denkbar, vergl. Uebersax (1987). Welches Modell man auch immer man betrachten möchte, stets gilt, dass alle subjektiven Aspekte in der Sensitivität und Spezifität enthalten sind. Inhomogene Randverteilungen entstehen u. A. durch unterschiedliche Sensitivitäten und Spezifitäten, die *auch* durch verschiedene subjektive Annahmen über die Prävalenz zustande kommen können, – aber diese subjektiven θ^s-Werte können nicht direkt aus den Randverteilungen errechnet werden.

Urteilerübereinstimmung A und B seien zwei Diagnostiker oder (Be-)Urteiler. Ihre Diagnosen oder Urteile werden gut übereinstimmen, wenn ihre Sensitivitäten η_A, η_B und ihre Spezifitäten ϑ_A, ϑ_B hoch sind. Haben diese Parameter kleine Werte oder sind sie unterschiedlich, wird die Übereinstimmung weniger gut sein, weil bei zumindest einem der beiden Diagnostiker mit größerer Wahrscheinlichkeit geraten oder vielleicht auch systematisch fehlbeurteilt wird. Die Frage ist nun, ob ein sinnvolles Maß für die Übereinstimmung gefunden werden kann.

Ein Übereinstimmungsmaß wird von den Häufigkeiten a und d abhängen, da die Diagnostiker mit diesen Häufigkeiten die gleichen Urteile abgeben. Aber es ist möglich, dass in einigen Fällen die Übereinstimmung nur zufällig ist, weil bei zumindest einem der Urteiler die Diagnose eben nur zufällig mit der des anderen Diagnostikers übereinstimmt. Man wird deshalb versuchen, die zufälligen Übereinstimmungen wieder herauszurechnen, und das bedeutet, dass man den Anteil der zufälligen Übereinstimmungen bestimmen muß. Man kann nun zwei Fälle betrachten:

1. Einer der beiden Urteiler, etwa A, liefert Urteile, die als Goldstandard gelten: Dieser wird durch Urteile definiert, die als "wahr" akzeptiert werden, während die Diagnosen von B etwa auf der Basis einer Methode gefällt werden, die weniger aufwändig oder billiger ist als die Methoden, die den Goldstandard erzeugen. Man will die Diagnosen von B nur akzpetieren, wenn die Diagnosen auf dieser Basis möglichst perfekt mit denen des Goldstandards übereinstimmen. In diesem Fall können die Sensitivitäten und die Spezifitäten aus den Häufigkeiten der Tabelle **??** geschätzt werden, was zu einer deutlicheren Interpretation eines Übereinstimmungsmaßes führt.

2. Keiner der beiden Urteiler liefert Urteile, die als Goldstandard gewertet werden können. Dies ist ein in der Psychologie häufig auftretender Fall. Der Fall eines existierenden Goldstandards ist in diesem allgemeinen Fall als Spezialfall enthalten (denn auch Pathologen können irren).

Man kann auf die Idee kommen, dass die übliche χ^2-Statistik ein Maß für die Urteilerübereinstimmung sei: je besser die Übereinstimmung, desto größer wird der χ^2-Wert sein, etwa, wenn $b = d = 0$ ist, zudem hätte man auch eine inferenzstatistische Möglichkeit, die "reine" Zufälligkeit der Übereinstimmungen zu testen. Der χ^2-Wert signalisiert allerdings nur, ob es systematische Abweichun-

gen von den bei rein zufälligem Übereinstimmungen zu erwartenden Häufigkeiten gibt, – und diese systematischen Übereinstimmungen müssen nicht korrekte Urteile repräsentieren, sondern können auch auf systematische Fehlbeurteilungen hinweisen. Deswegen wird das χ^2 im Allgemeinen nicht als Maß für die Urteilerübereinstimmung verwendet.

Wie oben bereits angemerkt, repräsentieren die Häufigkeiten a und d die Übereinstimmungen der Diagnostiker. Allgemein sei der Anteil der übereinstimmenden Urteile durch P_0 gegeben: für K Klassen ist dann

$$P_0 = \sum_{k=1}^{K} p_{kk}, \tag{2.7.16}$$

wobei p_{kk} die Wahrscheinlichkeit bzw. relative Häufigkeit ist, mit der sowohl der Urteiler A als der Urteiler B die Klasse $\mathcal{C}_k$ für den zu beurteilenden Fall gewählt haben. In Bezug auf die Tabelle 5 ist einfach

$$P_0 = P_{++} + P_{--} = \frac{a}{N} + \frac{d}{N} = \frac{a+d}{N}. \tag{2.7.17}$$

P_0 ist dann auch der Ausgangspunkt für die gängigen Übereinstimmungsmaße. Die Unterschiede zwischen den Maßen ergeben sich durch die Art und Weise, in der die zufälligen Übereinstimmungen bestimmt werden. Die Maße werden im folgenden kurz vorgestellt: es sind

1. Das S-Maß von Bennet, Alpert und Goldstein (1954)
2. Das π-Maß von Scott (1955)
3. Das κ-(Kappa)-Maß con Cohen (1960).

Die Kritik an den Maßen S und π hat Cohen zu seinem Maß κ geführt, das zu den nahezu standardmäßig berechneten Maßen geworden ist. Dieses Maß wird hier deswegen ausführlich diskutiert werden. .

2.7.2 Die Maße von Bennet et al., Scott, und Cohen

Das S-Maß: Das erste Maß scheint von Bennet, Alpert und Goldstein (1954) vorgeschlagen worden zu sein. Es sei allgemein K die Anzahl der Antwortkategorien (im Falle von Tabelle **??** ist $K = 2$), und P_0 sei wie in (2.7.16) definiert. Dann definieren Bennet et al.

$$S = \frac{K}{K-1}\left(P_0 - \frac{1}{K}\right) \tag{2.7.18}$$

als Maß für die Übereinstimmung.

Bennet et al.'s Annahmen gehen auf ein Modell von Guttman (1946) zurück, in dem die Reliabilität eines (Test-)Items für eine Person definiert wird. Guttman definiert als *modale Wahrscheinlichkeit* (modal probability) die maximale Wahrscheinlichkeit, mit der eine Person in eine bestimmte Kategorie "sortiert" wird, und ist diese für alle Kategorien gleich groß, so ist die modale Wahrscheinlichkeit durch $1/K$ gegeben. Die Reliabilität für diese Person wird dann von Guttman als $R_i = \frac{K}{K-1}(P_i - \frac{1}{K})$ definiert, mit $1/K \leq P_i \leq 1$, und P_i ist die modale Wahrscheinlichkeit für die i-te Person. Offenbar ist die Größe S in (2.7.18) in direkter Analogie zu R_i definiert worden.

Im Falle "reiner Zufälligkeit" hat man

$$\hat{n}_{kk} = \frac{n_{k+}n_{+k}}{N}, \quad P_0 = \frac{1}{N}\sum_{k=1}^{K}\frac{n_{k+}n_{+k}}{N}$$

Nimmt man weiter an, dass jeder Beurteiler eine der Kategorien mit der Wahrscheinlichkeit $1/K$ – als "rein zufällig" – wählt, so ist die erwartete Randhäufigkeit für die k-te Kategorie

$$\mathbb{E}(n_{+k}) = \mathbb{E}(n_{k+}) = N\frac{1}{K} = \frac{N}{K},$$

und man erhält

$$\mathbb{E}(\hat{n}_{kk}) = \frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\left(\frac{n_{k+}n_{+k}}{N}\right) = \frac{1}{N}K\frac{N}{K}\frac{N}{K}\frac{1}{N} = \frac{1}{K},$$

wobei von $\mathbb{E}(n_{k+}n_{+k}) = \mathbb{E}(n_{k+})\mathbb{E}(n_{+k})$ wegen der Unabhängigkeitsannahme Gebrauch gemacht wurde. Also gilt bei rein zufälliger Übereinstimmung $P_0 = 1/K$, so dass dann $S = 0$. Sind andererseits $n_{ij} = 0$ für alle $i \neq j$ ($n_{11} = a$, $n_{12} = b$ etc in Tabelle **??**), so folgt $P_0 = 1$ und

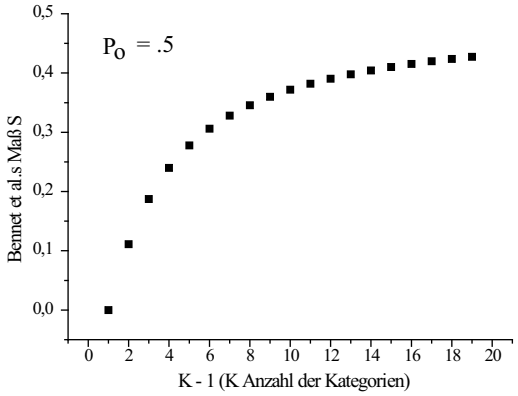
$$S = \frac{K}{K-1}\left(1 - \frac{1}{K}\right) = \frac{K}{K-1}\frac{K-1}{K} = 1,$$

so dass $0 \leq S \leq 1$. Bennet et al.s Maß S ist, für fixen Wert von P_0, nicht unabhängig von der Anzahl K der Kategorien; Abbildung 15 veranschaulicht diesen Sachverhalt. Eine weitere offene Frage ist, wie die üblicherweise geforderte Unabhängigkeit der Urteiler denn in P_0 untergebracht werden kann, ohne dass die Gleichung $\hat{n}_{kk} = (n_{k+}n_{+k})/N$ mit dieser Unabhängigkeit verbunden wird. Diese Fragen führten Scott (1955) zu seinem Maß π.

Scotts (1955) Maß π: Scott (1955) nahm die Abhängigkeit des Maßes S von der Anzahl K der Kategorien zum Anlaß, ein alternatives Maß π vorzuschlagen:

$$\pi = \frac{P_0 - P_e}{1 - P_e}. \tag{2.7.19}$$

Abbildung 15: Bennet et al.s Maß S; Abhängigkeit von der Anzahl der Kategorien



Hierin ist wieder $P_0 = \sum_k n_{kk}/N$ wieder der Anteil identischer Urteile, und P_e ist der erwartete Anteil identischer Urteile, die "rein zufällig" zustande kommen. Scott definiert

$$P_e = \sum_{k=1}^{K} p_k^2 \approx \sum_{k=1}^{K} \frac{n_{k+}n_{+k}}{N^2}, \tag{2.7.20}$$

wobei das $\approx$-Zeichen hier benutzt wurde, um anzuzeigen, dass $n_{+k}/N \approx p_k$ und $n_{k+}/N \approx p_k$. Dieser Ausdruck legt nahe, dass Scott implizit die Annahme gleicher Randverteilungen ("Homogenität der Randverteilungen") machte, also $n_{+k} \approx n_{k+}$ für $k = 1, \ldots, K$ postulierte; Gwet (2002) ist allerdings der Ansicht, dass diese Annahme, die Scott ja nicht explizit machte, Scott nicht unterstellt werden dürfe. Die Diskussion der Gwetschen Interpretation ist allerdings ein wenig akademisch und soll hier deswegen nicht weiter verfolgt werden.

(2.7.20) entspricht der Wahrscheinlichkeit, dass beide Urteiler unabhängig voneinander die gleiche Kategorie wählen. Für $P_0 \neq P_e$ könnte man folgern, dass die beiden Urteiler gelegentlich *nicht* unabhängig voneinander urteilen. Andererseits wird man natürlich fordern, dass die beiden Urteiler stets unabhängig voneinander disgnostizieren: würde man erlauben, dass es zu Absprachen zwischen den Diagnostikern kommt oder einer der beiden sich nach dem richtet, was der andere sagt, so würde die Unterscheidung von P_0 und P_e von vornherein sinnlos; überhaupt würde die Berechung eines Maßes für die Übereinstimmung überflüssig. Was hier mit Unabhängigkeit gemeint ist, muß genauer als *bedingte Unabhängigkeit* spezifiziert werden, die auch in den Fällen gegeben ist, wenn beide Urteiler zu einem gleichen Urteil kommen, ohne dass sich die Häufigkeit dieser übereinstimmenden Urteile gemäß (2.7.20) aus den Randsummen errechnen ließe. In Abschnitt 2.7.3 wird die Unterscheidung zwischen 'Unabhängigkeit' und 'bedingter Unabhängigkeit' explizit gemacht. Die Subtraktion von P_c wird of als "chance correction" interpretiert, als würden durch die Subtraktion von

P_c die zufällig korrekten Urteile eliminiert. Das ist nicht der Fall, worauf in den Bemerkungen Zur Rolle von P_c auf Seite 112 eingegangen wird.

Wichtig ist zunächst, dass sich das Maß π als, von K, der Anzahl der Kategorien, unabhängig erweist. Eine Kritik an Scotts π ergibt sich gleichwohl; sie kann aber in Zusammenhang mit der Variante, die diesem Maß von Cohen (1960) gegeben wurde, diskutiert werden. Es muß wieder angemerkt werden, dass der Begriff der Unabhängigkeit der Urteiler nicht explizit diskutiert wird.

Cohens Kappa κ (Kappa) (Cohen (1968)) ist ein Maß, dass mittlerweile nahezu standardmäßig berechnet wird.[32] Cohens κ gleicht auf den ersten Blick dem Scottschen π, aber Cohen kritisiert die Scottsche (implizite) Annahme gleicher Randverteilungen, so dass er ein anderes Maß für die Zufälligkeit der Übereinstimmung vorschlägt. Es seien

$$P_c = \sum_{k=1}^{K} p_{Ak} p_{Bk}, \tag{2.7.21}$$

wobei die

$$p_{Ak} \approx \frac{n_{k+}}{N}, \quad p_{Bk} \approx \frac{n_{+k}}{N}, \quad k = 1, \ldots, K$$

die Randverteilungen für Beurteiler A bzw. B sind (die Unterscheidung zwischen 'unabhängig' und 'bedingt unabhängig' ist hier wieder relevant, vergl. Abschnitt 2.7.3). Weiter sei wieder

$$P_0 = \sum_{k=1}^{K} \frac{n_{kk}}{N}.$$

Dann soll

$$\kappa = \frac{P_0 - P_c}{1 - P_c}. \tag{2.7.22}$$

ein Maß für die Urteilerübereinstimmung sein.

Eine Verallgemeinerung von Cohen's κ für mehr als zwei Urteiler ist von Fleiss (1971) vorgeschlagen worden.

Offenbar wird P_0 maximal, nämlich $P_0 = N/N = 1$, wenn die Beurteiler perfekt übereinstimmen, denn dann sind alle $n_{ij} = 0$ für $i \neq j$. Dann folgt $P_0 = 1$, damit $P_0 - P_c = 1 - P_c$ und also $\kappa = 1$. Umgekehrt werde angenommen, dass die Beurteiler rein zufällig sortieren. Dann gilt $\hat{n}_{kk} = n_{k+}n_{+k}/N$ und es folgt $P_0 = P_c$, d.h. $\kappa = 0$.

κ kann negative Werte annehmen. Dazu muß $P_0 < P_c$ gelten. Im Extrem ist $P_0 = 0$ so dass bei einem dichtomen Merkmal die beiden Beurteiler stets

[32] κ = griechisch Kappa. Der Buchstabe κ wird in diesem Text in erster Linie für Itemparameter (Schwierigkeit) benutzt. Cohen's κ ist allerdings ein eingeführter Begriff. Es wird im jeweiligen Kontext klar sein, was mit κ gemeint ist.

entgegengesetzt urteilen und der Spezialfall $n_{11} = n_{22} = 0$ eintritt; dann folgt $P_C = 1/2$ und damit

$$\kappa = -\frac{1/2}{1 - 1/2} = -1.$$

Gewichteter Kappa-Koeffizient Auf den ersten Blick erscheint κ ein vernünftiges Maß für die Übereinstimmung zu sein. Allerdings wird Nichtübereinstimmung über die Zufälligkeit der Sortierungen definiert, und dementsprechend die spezielle der Art Divergenz der Einschätzungen nicht in das κ-Maß ein. Diese Eigenschaft ist von Nachteil spätestens dann, wenn die Kategorien A, B, etc eine Rangordnung repräsentieren. Die Klassifikation in unmittelbar benachbarte Kategorien wird somit genau so bewertet wie eine Klassifikation in weit auseinander liegende Kategorien. κ unterschätzt dann die tatsächliche Übereinstimmung. Es liegt nahe, die Definition von κ zu modifizieren, indem man ein Gewichtung für die Kategorien einführt. Zunächst sei angemerkt, dass κ sich in der Form

$$\kappa = 1 - \frac{1 - p_0}{1 - p_c} = 1 - \frac{\sum_{i \neq j} n_{ij}/N}{\sum_{i \neq j} n_{i+} n_{j+}/N^2} = 1 - \frac{N \sum_{i \neq j} n_{ij}}{\sum_{i \neq j} n_{i+} n_{j+}} \tag{2.7.23}$$

schreiben läßt. Die n_{ij} und $n_{i+}n_{j+}$ lassen sich nun mit Gewichten w_{ij} versehen, dass ein modifizierter κ-Wert κ_w

$$\kappa_w = 1 - \frac{N \sum_{i \neq j} w_{ij} n_{ij}}{\sum_{i \neq j} w_{ij} n_{i+} n_{j+}} \tag{2.7.24}$$

entsteht. Bei ranggeordneten Kategorien und unter der Annahme gleichabständiger Kategoriengrenzen werden die Gewichte durch

$$w_{ij} = (i - j)^2 \tag{2.7.25}$$

definiert. Je größer also der Abstand $|i - j|$ der Kategorien K_i und K_j, desto größer wird w_{ij}, – die Gewichte wachsen sogar quadratisch mit dem Abstand der Kategorien.

Varianz von κ Wie jede Statistik ist der Wert von κ von der Stichprobe abhängig. So ergibt sich die Frage, von welchem Wert an κ signifikant von Null verschieden ist. Dazu muß man die Verteilung von κ unter $H_0 : \kappa = 0$ kennen. Everitt (1968) hat die Momente dieser Verteilung angegeben[33]. Fleiss, Cohen und Everitt (1969) haben aber Abschätzungen für den Standardfehler von κ für große Stichproben angegeben, die sehr genau der exakten Abschätzung durch Everitt (1968) entspricht:

$$\hat{\sigma}^2_\kappa = \frac{P_c(1 + P_c) - \sum_{k=1}^{K} P_{k+} P_{+k} (P_{k+} + P_{+k})}{n(1 - P_c)^2}. \tag{2.7.26}$$

[33]Der Artikel ist gegenwärtig nicht verfügbar, deswegen hier keine weiteren Angaben.

Nimmt man an, dass

$$\frac{\hat{\kappa}}{\hat{\sigma}_\kappa} \sim N(0,1),$$

so kann man H_0 in Bezug auf diese Annahme testen. Wie Banerjee et al. (1999) zu Recht anmerken, ist die Frage nach der Signifikanz aber im Allgemeinen von geringem Interesse, wenn die Beurteiler trainierte Experten sind und sie deshalb sowieso "signifikant" übereinstimmen.

Bewertung eines κ-Wertes Die Frage ist mehr, ob die Übereinstimmung "gut" oder "schlecht" ist. Landis und Koch (1977) haben diese Frage diskutiert und schlagen die folgende Klassifikation von κ-Werten vor: Die Autoren merken dazu

Tabelle 7: Landis & Koch (1977) Klassifikation von κ-Werten

κ	Stärke der Übereinstimmung
$< .00$	poor
.00 – .20	slight
.21 – .40	fair
.41 – .60	moderate
.61 – .80	substantial
.81 – 1.00	almost perfect

an, dass diese Klassifikation 'clearly arbitrary' seien und nur eine 'benchmark' darstellen. Aber sie ist für Praktiker anscheinend zu einer nicht weiter hinterfragten Einteilung und Interpretationshilfe geworden (im Wikipdia-Eintrag 'Cohens Kappa' (http://de.wikipedia.org/wiki/Cohens-Kappa) wird diese Einteilung nur vorgestellt ohne den Zusatz von Landis und Koch über die Beliebigkeit zu erwähnen. Man sollte also zurückhaltend mit dem Gebrauch dieser Kategorien sein, zumal Cohens κ durchaus kritisiert werden kann, wie im Folgenden noch verdeutlicht werden wird. Dazu wird zunächst eine etwas grundsätzliche Herleitung von Cohens κ vorgestellt, in die insbesondere auch die Frage nach der Unabhängigkeit der Urteiler explizit behandelt wird.

2.7.3 Herleitung von Cohens κ: $K = 2$

Kein Goldstandard Es wird zunächst der Fall zweier Urteiler betrachtet, von denen keiner den Goldstandard liefert. Mit A_+ und B_+ werden die "positiven" Urteile von A und B bezeichnet, wobei *positiv* bedeutet, dass von A bzw. B das bei einem "Fall" (einer Person) Vorhandensein des Merkmals $\mathcal{M}$ diagnostiziert wurde. Mit A_- und B_- werden die "negativen" Urteile von A bzw. B bezeichnet, d.h. die Diagnose, dass $\mathcal{M}$ nicht vorhanden ist. Dann

1. $\eta_A = P(A_+|\mathcal{M}_+)$, $\eta_B = P(B_+|\mathcal{M}_+)$ seien die *Sensitivitäten* von A bzw. B.
2. $\vartheta_A = P(A_-|\mathcal{M}_-)$ und $\vartheta_B = P(B_-|\mathcal{M}_-)$ seinen die *Spezifitäten* von A und B.

Diese Charakterisierungen entsprechen den Definition 1 und 2 auf Seite 87, wo auch die Prävalenz θ eingeführt worden ist.

Die Wahrscheinlichkeiten für A_+ und B_+ definieren die Randverteilungen und ergeben sich aus dem Satz der totalen Wahrscheinlichkeit (vergl. (2.6.119), Seite 87) als

$$P(A_+) = \eta_A\theta + (1-\vartheta_A)(1-\theta) \quad (2.7.27)$$
$$P(B_+) = \eta_B\theta + (1-\vartheta_B)(1-\theta). \quad (2.7.28)$$

Wegen $P(A_+) + P(A_-) = P(B_+) + P(B_-) = 1$ erhält man sofort $P(A_-) = 1 - P(A_+)$ und $P(B_-) = 1 - P(B_+)$. Natürlich sind auch die Randverteilungen $(P(A_+), P(A_-) = 1 - P(A_+))$ und $(P(B_+), P(B_-) = 1 - P(B_+))$ wegen (2.7.27) und (2.7.28) Funktionen der Parameter $\eta_A, \ldots, \vartheta_B$.

Es sei $C_+ = A_+ \cap B_+$, $C_- = A_- \cap B_-$. Gesucht ist nun ein Ausdruck für

$$P_0 = P(C_+) + P(C_-). \quad (2.7.29)$$

Nach dem Satz der totalen Wahrscheinlichkeit gilt

$$P(C_+) = P(C_+|\mathcal{M}_+)\theta + P(C_+|\mathcal{M}_-)(1-\theta). \quad (2.7.30)$$
$$P(C_-) = P(C_-|\mathcal{M}_+)\theta + P(C_-|\mathcal{M}_-)(1-\theta) \quad (2.7.31)$$

Wie in Abschnitt 2.6.11 bereits angedeutet muß man annehmen, dass die beiden Urteiler bedingt unabhängig voneinander urteilen:

Annahme: Bedingte Unabhängigkeit

$$P(C_+|\mathcal{M}_+) = P(A_+|\mathcal{M}_+)P(B_+|\mathcal{M}_+) = \eta_A\eta_B \quad (2.7.32)$$
$$P(C_+|\mathcal{M}_-) = P(A_+|\mathcal{M}_-)P(B_+|\mathcal{M}_-) = (1-\vartheta_A)(1-\vartheta_B) \quad (2.7.33)$$

und

$$P(C_-|\mathcal{M}_+) = P(A_-|\mathcal{M}_+)P(B_-|\mathcal{M}_+) = (1-\eta_A)(1-\eta_B) \quad (2.7.34)$$
$$P(C_-|\mathcal{M}_-) = P(A_-|\mathcal{M}_-)P(B_-|\mathcal{M}_-) = \vartheta_A\vartheta_B. \quad (2.7.35)$$

so dass

$$P(C_+) = \eta_A\eta_B\theta + (1-\vartheta_A)(1-\vartheta_B)(1-\theta) \quad (2.7.36)$$
$$P(C_-) = (1-\eta_A)(1-\eta_B)\theta + \vartheta_A\vartheta_B(1-\theta). \quad (2.7.37)$$

Mit $P_0 = P(C_+) + P(C_-)$ und P_c definiert wie in (2.7.21) (Seite 106) mit $K = 2$ ist

$$\kappa(\eta_A, \eta_B, \vartheta_A, \vartheta_B, \theta) = \frac{P_0 - P_c}{1 - P_c} \quad (2.7.38)$$

nun als Funktion der Sensitivitäten, der Spezifitäten und der Prävalenz explizit definiert. Der Wert von P_0 ist empirisch durch (2.7.16), Seite 103 gegeben, d.h. durch die Werte in der Diagonalen der Kontingenztabelle.

Kommentar: Die Annahme der bedingten Unabhängigkeit ist auf den ersten Blick plausibel und erscheint vielleicht sogar notwendig: existieren Abhängigkeiten zwischen den beiden Urteilern, so reflektiert die Übereinstimmung oder der Mangel davon nicht nur die jeweiligen Sensitivitäten und Spezifitäten und damit die unterschiedliche Informationslagen oder Diagnosestrategien, sondern eben auch die Abhängigkeitsbeziehung, an der man eigentlich nicht interessiert ist. Andererseits zeigt sich aber, dass die Annahme der bedingten Unabhängigkeit durchaus eine *starke*, soll heißen: nicht notwendig realistische Annahme ist. So kann es sein, dass das zu diagnostizierende Merkmal weniger oder stärker ausgeprägt ist. Verschiedene Tests oder verschiedene Urteiler sind sich häufig einig bei starker Ausprägung des Merkmals und weniger einig bei schwacher Ausprägung des Merkmals, und dieser Sachverhalt kann sich in einer Verletzung der Annahme bedingter Unabhängigkeit äußern (vergl. Hui & Zhou, 1998). □

Schätzbarkeit der Parameter: Es werde zunächst angenommen, dass kein Goldstandard vorliegt. Zunächst muß diskutiert werden, wieviele freie Parameter zu betrachten sind. Sicherlich die für die Urteiler charakteristischen Sensitivitäten und Spezifitäten, d.h. es gibt mindestens vier freie Parameter. Es gibt vier voneinander unabhängige Ausdrücke bzw. Gleichungen: die für $P(A_+)$, $P(B_+)$, $P(C_+)$ und $P(C_-)$; die Gleichungen für $P(A_-)$ und $P(B_-)$ ergeben sich aus denen für $P(A_+)$ und $P(B_+)$. θ ist ein für die Population charakteristischer Parameter. Selbst wenn er aus epidemiologischen Untersuchungen bekannt ist, wäre eine Schätzung des in der Stichprobe vorkommenden Anteils $\tilde{\theta}$ von $\mathcal{M}_+$-Fällen nützlich, da die Schätzungen für die Paramter $\eta_A, \cdots, \vartheta_B$ natürlich von $\tilde{\theta}$ abhängen. Darüber hinaus hängen sie von den tatsächlich beobachteten Symptomen $\mathbf{x}$ ab (vergl. (2.7.11), so dass die Stichprobenverteilung der Schätzungen von den bedingten Verteilungen (i) für $\mathbf{x}$, (ii) für $\tilde{\theta}$ abhängen, und schließlich gehen noch die stochastischen Aspekte des Urteilsprozesses, der $P(\mathcal{M}_+|\mathbf{x})$ bestimmt, ein. Angesichts der weiter unten beschriebenen Eigenschaften von κ kann man die Frage, ob es sich lohnt, die mathematischen Mühen der Abschätzung der Verteilungen der Parameter auf sich zu nehmen, nur mit großer Zurückhaltung betrachten.

Liegt ein Goldstandard vor, so gilt etwa $\eta_A = \vartheta_A = 1$. Aus (2.7.27) folgt dann $P(A_+) = \theta$, $P(C_+) = \eta_B\theta$, $P(C_-) = \vartheta_B(1-\theta)$. Man hat drei Gleichungen in den

drei Unbekannten θ, η_B und ϑ_B, so dass man die Schätzungen

$$\hat{\theta} = \hat{P}(A_+) \tag{2.7.39}$$

$$\hat{\eta}_B = \frac{\hat{P}(C_+)}{\hat{\theta}} = \frac{\hat{P}(C_+)}{\hat{P}(A_+)} \tag{2.7.40}$$

$$\hat{\vartheta}_B = \frac{\hat{P}(C_+)}{1-\hat{\theta}} = \frac{\hat{P}(C_+)}{1-\hat{P}(A_+)}, \tag{2.7.41}$$

für die Sensitivität und Spezifität des Diagnostikers B sowie der Prävalenz erhält, wobei mit $\hat{P}(A_+)$ und $\hat{P}(C_+)$ die Schätzungen von $P(A_+)$ und $P(C_+)$ aus den entsprechenden relativen Häufigkeiten der Tabelle gemeint sind.

Der Goldstandardfall wurde von Feuerman & Miller (2005) für ein dichotomes Merkmal behandelt. In Feuerman & Millers Definition von κ tritt die Prävalenz nicht mehr explizit auf und die Rolle der Prävalenz wird auch nicht weiter diskutiert, ebenso nicht in den Folgeartikeln Feuerman et al. (2007, 2008). Es wird im wesentlichen nur gezeigt, dass sich κ auch direkt aus den Schätzungen der Sensitivität und der Spezifität berechnen läßt – was sich mit ein wenig Algebra leicht zeigen läßt und worauf hier nicht weiter eingegangen zu werden braucht.

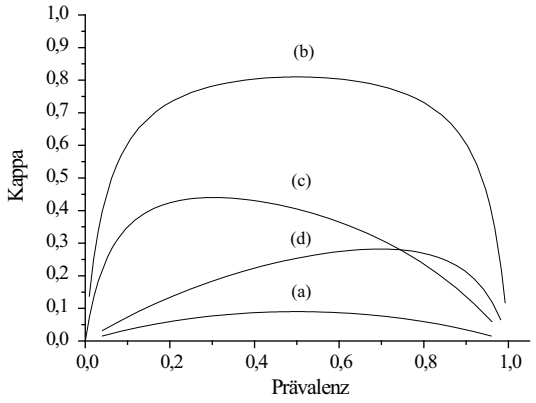
Die hier angegebenen Schätzungen für die Sensitivitäten und Spezifitäten werden direkt aus den Häufigkeiten der 2×2-Tabelle geschätzt. Die Eigenschaften der Schätzungen müssen hier nicht weiter diskutiert werden. Allerdings gibt es weitaus aufwändigere Methoden: Hui & Walter (1980) liefern Maximum-Likelihood-Schätzungen, McClish & Quade (1983) liefern Methoden der Schätzung der unbekannten Prävalenz. Hui & Zhou (1998) liefern Schätzungen auch für den Fall, dass *kein* Goldstandard, sondern bestenfalls ein *imperfekter Goldstandard* vorliegt; ebenfalls werden auch Bayesianische Schätzungen diskutiert (Andersen, 1997). Enø, Georgiadis & Johnson (2006) diskutieren ebenfalls allgemeine Schätzungen für Sensitivitäten, Spezifitäten und Prävalenz für den Fall, dass der wahre Zustand einer Person (oder Tieres) nicht bekannt ist. Allgemein gilt, dass die Schätzung der Parameter anhand von Maximum-Likelihood-Methoden etc numerisch wesentlich aufwändiger wird als bei den gängigen 'rough-and-ready'-Methoden.

Verhalten von κ: Wie in Anhang A.1 gezeigt wird, gilt für beliebige Sensitivitäten und Spezifitäten

$$\lim_{\theta \to 0} \kappa = \lim_{\theta \to 1} \kappa = 0, \tag{2.7.42}$$

vergl. Abb. 16. Dies bedeutet, dass nicht nur für sehr kleine, sondern auch für sehr große Werte von θ der Effekt von Rateprozessen dominant wird und damit κ sehr klein wird. Dies wiederum bedeutet, dass auch für große Sensitivitäten und Spezifitäten ein kleiner κ-Wert resultieren kann, ein kleiner κ-Wert mithin noch keinen Aufschluß über die tatsächlich vorhandenen Sensitivitäten und Spezifitäten liefert. Eine spezifischen Kombination von Sensitivitäten und Spezifitäten erzeugt einen spezifischen Verlauf der Abhängigkeit von κ von der Prävalenz θ.

Abbildung 16: Cohens κ als Funktion der Prävalenz für verschiedene Sensitivitäten und Spezifitäten: (a) $\eta_A = \eta_B = \vartheta_A = \vartheta_B = .65$, (b) $\eta_A = \eta_B = \vartheta_A = \vartheta_B = .95$, (c) $\eta_A = .5$, $\eta_B = .95$ $\vartheta_A = .95$, $\vartheta_B = .95$, (d) $\eta_A = \eta_B = .95$; $\vartheta_A = \vartheta_B = .5$



Bemerkenswert ist, dass κ als Funktion von θ für bestimmte Sensitivitäts- und Spezifitätskombinationen asymmetrisch verläuft, also für eine gegebene Kombination ihren Maximalwert für kleinere θ-Werte animmt (z.B. (c)) und für andere Kombinationen für größere θ-Werte (z.B. (d)). Homogene Randverteilungen erzeugen insbesondere für mittlere Prävalenzen, also $\theta \approx .5$, noch am besten zu interpretierende κ-Werte, insbesondere wenn die Urteiler über hohe Sensitivitäten und Spezifitäten verfügen. Nur muß man etwas über diese Werte wissen, um folgern zu können, dass man einen "guten" κ-Wert hat.

Liegen für θ keine Schätzungen aus unabhängigen epidemiologischen Untersuchungen vor, so sind die Parameter η_A bis ϑ_B, wie oben gezeigt, nicht identifizierbar; große Werte von κ (d.h. $\kappa > .8$) lassen große Sensitivitäten und Spezifitäten und einen eher mittleren Wert von θ vermuten. Für kleinere Werte von κ läßt sich kaum etwas sagen, eine Vielzahl von Kombinationen der Parameter ist möglich.

Feinstein und Cicchetti (1990) betrachten den Effekt unterschiedlicher Prävalenz: sie illustrieren anhand von 2×2-Tabellen, dass sich bei gleicher Übereinstimmung P_0 unterschiedliche κ-Werte ergeben und sprechen von Paradoxa, die die Definition von κ impliziere. Aufgrund der vorangegangenen Herleitung von κ als Funktion von fünf freien Parametern ist dieser Befund nicht verwunderlich und von Paradoxa kann kaum die Rede sein, so dass hierauf nicht näher eingegangen werden muß.

Anmerkung zur Bedeutung von P_c: Es ist die Frage aufgeworfen worden, ob die Bildung der Differenz $P_0 - P_c$ und die Normierung mit $1/(1 - P_c)$ überhaupt eine "Bereinigung" von zufälligen Effekten bedeutet. Uebersax spricht dement-

sprechend von 'The myth of chance-corrected agreement'

> "Despite its reputation as a chance-corrected agreement measure, kappa does not correct for chance agreement. Neither has the need for such adjustment been convincingly shown."
> (http://www.john-uebersax.com/stat/kappa2.htm)

Uebersax elaboriert das Argument nicht weiter, aber es ist leicht zu sehen, dass er Recht hat. Denn die Frage ist ja, was die "reine Zufälligkeit" P_c überhaupt bedeuten soll. Formal wird "reine Zufälligkeit" durch die Produktregel $P(A \cap B) = P(A)P(B)$ charakterisiert, d.h. durch die Gleichungen

$$P(A_+ \cap B_+) = P(A_+)P(B_+), \quad P(A_- \cap B_-) = P(A_-)P(B_-).$$

Wegen (2.7.27) und (2.7.28) hat man

$$\begin{aligned} P(A_+ \cap B_+) &= [\eta_A\theta + (1-\vartheta_A)(1-\theta)][\eta_b\theta + (1-\vartheta_B)(1-\theta)] \quad (2.7.43) \\ P(A_- \cap B_-) &= [(1-\eta_A)\theta + \vartheta_A(1-\theta)][(1-\eta_B)\theta + \vartheta_B(1-\theta)]. (2.7.44) \end{aligned}$$

Diese Gleichungen für $P(A_+ \cap B_+)$ und $P(A_- \cap B_-)$ zeigen, dass auch bei "reiner Zufälligkeit" die Sensitivitäten und Spezifitäten eingehen. Das ist nicht weiter verwunderlich, aber man könnte hoffen, dass die Komponenten, die zu $P(C_+|\mathcal{M}_-)$ und $P(C_-|\mathcal{M}_+)$ korrespondieren durch Subtraktion von P_c eliminiert werden, denn falsche, eventuell durch Raten zustande gekommene Urteile haben die Wahrscheinlichkeiten $P(C_+|\mathcal{M}_-)$ und $P(C_-|\mathcal{M}_+)$, und die "correction for chance" sollte vielleicht zumindest diese Urteile aus P_0 eliminieren. Nun wird P_0 durch $P_0 = P(C_+) + P(C_-)$ definiert, und $P(C_+)$ und $P(C_-)$ werden allgemein durch (2.7.30) und (2.7.31) definiert:

$$\begin{aligned} P(C_+) &= P(C_+|\mathcal{M}_+)\theta + P(C_+|\mathcal{M}_-)(1-\theta). \\ P(C_-) &= P(C_-|\mathcal{M}_+)\theta + P(C_-|\mathcal{M}_-)(1-\theta) \end{aligned}$$

Aber der Vergleich mit (2.7.43) und (2.7.44) lehrt, dass diese Wahrscheinlichkeiten durch Subtraktion von P_c *nicht* aus dem Ausdruck für P_0 eliminiert werden. Darüber hinaus enthalten die Wahrscheinlichkeiten $P(C_+|\mathcal{M}_+)$ und $P(C_-|\mathcal{M}_-)$, die durch die Sensitivitäten und Spezifitäten definiert sind, ebenfalls die Möglichkeit des Ratens: Es ist ja nicht ausgeschlossen, dass zwar $\mathcal{M}_+$ vorliegt, die urteilende Person $\mathcal{M}_+$ aber nicht identifiziert und einfach rät, dass $\mathcal{M}_+$ vorliegt. Die Betrachtung für $\mathcal{M}_-$ ist analog. Wie dies im einzelnen geschehen kann, könnte man in einem expliziten Modell des Urteilsprozesses formulieren. Solche Modelle sind aber im Allgemeinen gar nicht interessant, weil es einerseits verschiedene denkbare Modelle gibt, die andererseits in einer spezifischen diagnostischen Situation gar nicht überprüft werden können.

Die Rede von der "correction for chance" trifft also nicht den Kern der Definition von Scotts π oder Cohens κ. Die Subtraktion von P_c und die Division durch

$1 - P_c$ bewirkt einfach eine Normierung: streben die Sensitivitäten und Spezifitäten gegen 1, so strebt κ gegen 1, und streben sie gegen die Wahrscheinlichkeiten $P(A_+)$, $P(B_+)$, $P(A_-)$ und $P(B_-)$ so strebt κ gegen 0. Abhängigkeiten von $\mathcal{M}_+$ und $\mathcal{M}_-$ können im Übrigen auch negative Werte von π bezw. κ erzeugen. Die in Abbildung 16 illustrierte Abhängigkeit von der Prävalenz $P(\mathcal{M}_+) = \theta$ erschwert aber die Interpretation eines gegebenen κ-Wertes.

□

2.7.4 Herleitung von Cohens κ: $K > 2$

Vielfach soll eine Kategorisierung nicht nur bezüglich des Vorhandenseins oder Nichtvorhandenseins eines Merkmals vorgenommen werden, sondern Personen oder Objekte sollen bezüglich einer Menge von $K > 2$ Kategorien beurteilt werden. Eine Spezialfall dieses Falles ist, wenn die $\mathcal{C}_1, \ldots, \mathcal{C}_K$ die Aufteilung eines Kontinuums in Intervalle repräsentieren.

Es sei A_j^+ das Ereignis, dass Urteiler A ein Objekt der Klasse $\mathcal{C}_j$ zuordnet, und π_k sei die Wahrscheinlichkeit, dass das Objekt zur Klasse $\mathcal{C}_k$ gehört. Dann ist

$$P(A_j^+) = \sum_{k=1}^{K} P(A_j^+|\mathcal{C}_k)\pi_k = P(A_j^+|\mathcal{C}_j)\pi_j + \sum_{j \neq k} P(A_j^+|\mathcal{C}_k)\pi_k. \tag{2.7.45}$$

Es sei

$$\eta_{jk}^A = P(A_j^+|\mathcal{C}_k), \quad \vartheta_{jk}^A = P(A_j^-|\neg\mathcal{C}_k). \tag{2.7.46}$$

Inbesondere sei

$$\eta_{jj}^A = \text{Sensitivität v. A bez. } \mathcal{C}_j, \ \vartheta_{jj}^A = \text{ Spezifität v. A bez. } \mathcal{C}_j \tag{2.7.47}$$

Es ist $P_0 = \sum_k p_{kk} = \sum_k P(A_j^+ \cap B_j^+)$, und

$$P(A_j^+ \cap B_j^+) = P(A_j^+ \cap B_j^+|\mathcal{C}_j)\pi_j + P(A_j^+ \cap B_j^+|\neg\mathcal{C}_j)(1 - \pi_j) \tag{2.7.48}$$

Die bedingte Unabhängigkeit impliziert nun

$$P(A_j^+ \cap B_j^+) = P(A_j^+|\mathcal{C}_j)P(B_j^+|\mathcal{C}_j)\pi_j + P(A_j^+|\neg\mathcal{C}_j)P(B_j^+|\neg\mathcal{C}_j)(1 - \pi_j), \tag{2.7.49}$$

d.h.

$$P(A_j^+ \cap B_j^+) = \eta_{jj}^A\eta_{jj}^B\pi_j + (1 - \vartheta_{jj}^A)(1 - \vartheta_{jj}^B)(1 - \pi_j), \tag{2.7.50}$$

und

$$P_0 = \sum_{j=1}^{K} \eta_{jj}^A\eta_{jj}^B\pi_j + \sum_{j=1}^{K}(1 - \vartheta_{jj}^A)(1 - \vartheta_{jj}^B)(1 - \pi_j). \tag{2.7.51}$$

Die Berechung von P_c ist offenkundig. Die Frage ist, ob die Sensitivitäten und Spezifitäten geschätzt werden können. Es ist

$$P(A_j^+) = p(A_j^+|\mathcal{C}_j)\pi_j + P(A_j^+|\neg\mathcal{C}_j)(1-\pi_j) = \eta_{jj}^A\pi_j + (1-\vartheta_{jj}^A)(1-\pi_j) \quad (2.7.52)$$

$$P(B_j^+) = p(B_j^+|\mathcal{C}_j)\pi_j + P(B_j^+|\neg\mathcal{C}_j)(1-\pi_j) = \eta_{jj}^B\pi_j + (1-\vartheta_{jj}^B)(1-\pi_j) \quad (2.7.53)$$

für $j = 1, \ldots, K$. Es sind also K Parameter η_{jj}^A, K Parameter ϑ_{jj}^A und analog dazu die Parameter für B zu schätzen, – also 8 freie Parameter. Hinzu kommen – wegen $\sum_j \pi_j = 1$ – noch einmal $K-1$ freie Parameter (die Prävalenzen) π_j, so dass für die $K \times K$-Tabelle $3K-1$ freie Parameter zu schätzen sind. Allerdings scheinen die Parameter nicht identifzierbar zu sein, denn für alle $P(A_j^+ \cap B_k^+)$ kommen noch die η_{jk} und ϑ_{jk} hinzu.

Ein einfacher Ansatz, doch zu Schätzungen von Sensitivitäten zu kommen, besteht in der Dichotomisierung. Dazu wählt man eine Kategorie, etwa $\mathcal{C}_k$, und faßt alle übrigen Kategorien zu einer Kategorie $\neg\mathcal{C}_k$ zusammen. Man schätzt nun die Parameter für diese 2×2-Tabelle. Analog dazu kann man mit den anderen Kategorien verfahren. Die Frage, ob die Schätzungen irgendwelchen Optimalitätskriterien entsprechen, bleibt dabei vorerst offen.

2.7.5 Beurteilungen auf Ratingskalen

Der messtheoretische Status von Rating-Skalen ist oft nicht eindeutig: während die Skalen eine ordinale Ordnung durchaus wiedergeben können, ist es nicht leicht, festzustellen, ob auch in guter Näherung eine Intervallskalenstruktur vorliegt. Dies muß bei der Interpretation der Korrelationen zwischen den Schätzungen (Ratings) verschiedener Beurteiler berücksichtigt werden. Die Koeffizienten für Kategorial- oder Ordinaldaten ergeben sich zum Teil direkt aus dem Produkt-Moment-Koeffizienten

$$r_{xy} = \frac{Kov(X,Y)}{s_x s_y}; \quad (2.7.54)$$

so ergibt sich der ϕ-Koeffizient für zwei dichotome Variablen direkt aus der Formel für r_{xy}, wenn die Variablen Y und Y nur die Werte 0 oder 1 annehmen können. Die gleiche Aussage gilt für die punkt-biseriale Korrelation, bei eine Variable, etwa X, Intervallskalenniveau hat und die andere, Y, nur ein ordinales Niveau hat und die Y-Werte Rangzahlen sind. Für Rangkorrelationen wie Spearmans ρ ergibt sich der Koeffizient

Spearmans ρ:

$$\rho = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)} \quad (2.7.55)$$

ebenfalls aus (2.7.54), wenn für die X- und Y-Werte die entsprechenden Rangplätze eingegeben werden. Für inferenzstatistische Aussagen müssen natürlich die Verteilungen von ϕ bzw. ρ gewählt werden.

Kendalls τ: Anders liegt der Fall, wenn Kendalls τ verwendet werden soll: hier werden die Rangplätze r_i einer Variablen, etwa die von X, als Anker verwendet. Die korrespondierenden Rangplätze der Y-Werte seien s_i. Man betrachtet dann die Rangreihe $r_1, r_2, \ldots, r_n$ und vergleicht sie mit den Rängen der korrespondierenden s_i-Werte. Ist $r_i = s_i$ für alle $i = 1, \ldots, n$, so hat man einen perfekten Zusammenhang. Andernfalls gibt es *Inversionen* und *Proversionen*: eine Inversion liegt vor, wenn $s_i > s_j$ für $i < j$, und man hat eine Proversion, wenn $s_i < s_j$ für $i < j$. Je größer die Anzahl der Inversionen, desto kleiner ist der Zusammenhang zwischen X und Y. Es sei

$$k_{ij} = \begin{cases} 1, & \text{wenn } s_i > s_j, \ i < j \\ 0, & \text{wenn } s_i < s_j, \ i < j \end{cases} \tag{2.7.56}$$

und

$$I = \sum_{i<j} k_{ij} \tag{2.7.57}$$

die Gesamtzahl der Inversionen. Die Gesamtzahl der verschiedenen Paare von Rängen ist

$$\binom{n}{2} = \frac{n(n-1)}{2}.$$

Es sei P die Anzahl der Proversionen. Dann muß

$$P + I = \frac{n(n-1)}{2} \tag{2.7.58}$$

sein. Die Größe

$$W = P - I \tag{2.7.59}$$

heißt *Kendall-Summe*. Man hat

$$W = \begin{cases} n(n-1)/2, & I = 0 \\ -n(n-1)/2, & P = 0 \end{cases} \tag{2.7.60}$$

Für $I = 0$ ist der Zusammenhang zwischen den beiden Rangordnungen perfekt, für $P = 0$ ist er genau invers, d.h. W ist ein Maß für den Zusammenhang für die beiden Reihen. Durch Normierung erhält man einen Koeffizienten, der zwischen -1 und +1 liegt und damit zu einem Korrelationskoeffizienten korrespondiert:

$$\tau = \frac{W}{n(n-1)/2} = \frac{2W}{n(n-1)} \ \text{ Kendalls Tau.} \tag{2.7.61}$$

im Allgemeinen findet man für gegebene Rangdaten, dass Spearmans $\rho >$ Kendalls τ.

Kendalls W: Die bisher besprochenen Korrelationsmaße sind Reliabilitätsmaße für jeweils zwei Beurteiler. Will man ein Gesammtmaß für mehrere Experten haben, kann man Kendalls W berechnen: für M Urteiler hat man

$$W = \frac{12 \sum_{i=1}^{n} (r_i - \bar{r})^2}{M^2 (K^3 - K)}, \quad r_i = \sum_{j=1}^{b} r_{ij} \tag{2.7.62}$$

r_{ij} ist der Rangplatz des i-ten Objekts beim j-ten Urteiler, $\bar{r} = \sum_i r_i/n$, K die Anzahl der Kategorien, n die Anzahl der beurteilten Personen oder Objekte. Zwischen der durchschnittlichen (Spearmanschen) Rangkorrelation $\bar{\rho}$ und W besteht die Beziehung

$$\bar{\rho} = \frac{MW - 1}{M - 1} \tag{2.7.63}$$

bzw.

$$W = \frac{(M-1)\bar{\rho} + 1}{M}. \tag{2.7.64}$$

Intraklassen-τ: Für dieses Maß werden zunächst die x_i und die y_i-Werte in eine Rangreihe gebracht. Den beurteilten Personen oder Objekte ω_j werden dann die Ränge der beiden Beurteiler Hier sind R_1 und R_2 zwei Beurteiler, und $r_{(j)}$ ist der

	Personen/Objekte				
	ω_1	ω_2	ω_3	$\cdots$	ω_n
R_1	$r_{(1)}$	$r_{(2)}$	$r_{(3)}$	$\cdots$	$r_{(n)}$
R_2	$s_{(1)}$	$s_{(2)}$	$s_{(3)}$	$\cdots$	$s_{(n)}$

Rangplatz, den die person bzw, das Objekt ω_j beim Beurteiler R_1 erhalten hat, und $s_{(j)}$ ist der Rangplatz, den ω_j durch den Beurteiler R_2 erhalten hat. So ist etwa $r_{(1)} = 3$, wenn ω_1 durch R_1 auf den dritten Platz in der Rangreihe verwiesen wurde, und $s_{(1)} = 2$, wenn R_2 ω_1 auf den zweiten Platz verwiesen wurde. Es gilt also *nicht* notwendig

$$r_{(1)} < r_{(2)} < \cdots < r_{(n)},$$

und ebenso gilt *nicht* notwendig

$$s_{(1)} < s_{(2)} < \cdots < s_{(n)}.$$

Jetzt wird eine zweite Tabelle angelegt, in der die ω_j in eine Rangreihe gebracht werden:

$$\omega_{(1)} \prec \omega_{(2)} \prec \cdots \prec \omega_{(n)},$$

wobei $\omega_{(1)}$ diejenige Person oder dasjenige Objekt ist, für das $t_{(1)} = \min(r_{(j)}, s_{(j)})$ der kleinste aller $\min(r_{(j)}, s_{(j)})$-Werte ist, $\omega_{(2)}$ ist diejenige Person oder dasjenige Objekt, für das $t_{(2)} = \min((r_{(k)}, s_{(k)})$ der zweitkleinste Wert ist, etc. Für jedes $\omega_{(j)}$, $j = 1, \ldots, n$ wird nun die Anzahl der Proversionen P_j und Inversionen I_j relativ zu r_j-Werte bestimmt; analog dazu ddie Anzahl der Proversionen und Inversionen für den entsprechenden s-Wert. Dann ist

$$\tau_{ic} = \sum_{j=1}^{2n} (P_j - I_j) \tag{2.7.65}$$

ist dann der Intraklassen-τ-Wert.

2.7.6 Der Intraklassen-Korrelationskoeffizient

Haben die Ratings Intervallskalenniveau, so kann man u.U. die Frage nach der Reliabilität der Einschätzungen varianzanalytisch angehen. Mit ω_i, $i = 1, \ldots, n$ werden wieder die zu beurteilenden Personen oder Objekte bezeichnet. Alle Urteile werden auf einer Rating-Skala gefällt. Man kann dann, mit Shrout & Fleiss (1979), verschiedene Erhebungsansätze betrachten:

1. *Modell 1* Die ω_i werden jeweils von einer Menge M_i von Urteilern R_j, $j = 1, \ldots, K$, bewertet, die aus einer Population von Urteilern zufällig ausgewählt werden.

   Diese Versuchsanordnung ("Design") entspricht einer 1-faktoriellen ANOVA[34] mit Random-Effekten. Man hat die generelle Strukturgleichung

   $$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}. \tag{2.7.66}$$

   wobei a_i der Effekt von ω_i ist und b_j ist der Effekt von R_j; $(ab)_{ij}$ ist die Wechselwirkung von R_j und ω_i. Allerdings sind die Größen b_j und $(ab)_{ij}$ nicht schätzbar, so dass man sie zusammen mit e_{ij} zu einem Fehler

   $$w_{ij} = b_j + (ab)_{ij} + e_{ij} \tag{2.7.67}$$

   zusammenfassen muß. Damit hat man das Modell

   $$x_{ij} = \mu + a_i + w_{ij}. \tag{2.7.68}$$

2. *Modell 2* Es wird eine Menge M von K Urteilern R_j, $j = 1, \ldots, K$ zufällig aus einer Population gewählt, und jeder R_j beurteilt alle ω_i, d.h. R_j gibt n Urteile ab.

   Es gilt zunächst wieder die allgemeine Strukturgleichung (2.7.66) für 2-faktorielle Designs, allerdings kann nun der Effekt der Rater, also die b_j, geschätzt werden. Aber es gibt für die Kombination (R_j, ω_i) jeweils nur eine Messung, so dass die Wechselwirkungskomponente $(ab)_{ij}$ nicht geschätzt bzw. identifiziert werden kann. Wie bei solchen Designs üblich, muß angenommen werden, dass die Wechselwirkungen alle gleich Null sind, $(ab)_{ij} = 0$ für alle i, j. Damit hat man das Modell

   $$x_{ij} = \mu + a_i + b_j + e_{ij}. \tag{2.7.69}$$

3. *Modell 3* K Urteiler R_j, $j = 1, \ldots, K$ beurteilen alle ω_i, – aber man ist nur an den Urteilen *dieser* Menge M von Urteilern interessiert, d.h. man will nicht von M auf die Population möglicher Urteiler verallgemeinern. Die Urteiler gehen damit nicht wie bisher als *Random Factor*, sondern als *Fixed*

[34] Acronym für Analysis Of Variance, also Varianzanalyse

Factor ein. Der Faktor A mit den Effekten a_i ist dagegen nach wie vor ein *Fixed Factor*, und somit handelt es sich um ein 2-faktorielles Design mit gemischten Effekten.

Wieder gilt generell der Ansatz (2.7.66), nur ergeben sich andere Betrachtungen bezüglich der Interaktionen $(ab)_{ij}$. Für gegebenes ω_i muß angenommen werden, dass die Bedingung

$$\sum_{j=1}^{K}(ab)_{ij} = 0, \quad i = 1, \ldots, n \tag{2.7.70}$$

Wie in Scheffé (1959), Abschnitt 8.1 gezeigt wird, sind irgendzwei Wechselwirkungskomponenten $(ab)_{ij}$ und $(ab)_{i'j'}$ negativ miteinander korreliert. Wird, wie bei einem Design dieser Art üblich, angenommen, dass alle Kovarianzen zwischen den Interaktionen gleich groß, nämlich gleich c sind, so ergibt sich

$$c = -\frac{\sigma_b^2}{K-1}. \tag{2.7.71}$$

Bei den letzten beiden Fällen können die Daten in einer Matrix wie in Tabelle 8 zusammengefasst werden: Bevor die den Modellen entsprechenden Schät-

Tabelle 8: Einschätzungen von n Personen bzw. Objekten ω_i durch K Experten

Person	Experten			
ω_i	R_1	R_2	$\cdots$	R_k
1	x_{11}	x_{12}	$\cdots$	x_{1K}
2	x_{21}	x_{22}	$\cdots$	x_{2K}
$\vdots$			$\vdots$	
n	x_{n1}	x_{n2}	$\cdots$	x_{nK}

zungen für die Varianzkomponenten vorgestellt werden, wird der Intraklassen-Korrelationskoeffizient vorgestellt.

Der Intraklassen-Korrelationskoeffizient (IKK): Für ein Rating x_{ij} kann man vom Ansatz

$$x_{ij} = \mu_{ij} + w_{ij} = \mu + a_i + w_{ij}, \quad \begin{matrix} i = 1, \ldots, n \\ j = 1, \ldots, K \end{matrix} \tag{2.7.72}$$

ausgehen, wobei wieder w_{ij} ein zufälliger Fehler ist, der, wie oben angedeutet wurde, in weitere Komponenten zerlegt werden kann. Es sei $\mathbb{E}(\mu_{ij}) = \mu$, d.h. der Mittelwert der Mittelwerte sei μ. Dann ist $a_i = \mu_{ij} - \mu$ die Abweichung von μ_{ij} von μ, und man kann $\mu_{ij} = \mu + a_i$ schreiben. Da der Mittelwert der Abweichungen vom Mittelwert stets gleich Null ist, hat man $\sum_i a_i = 0$. a_i ist für ω_i eine Konstante, da

aber die ω_i zufällig in die Stichprobe gewählt werden kann a_i auch als Realisierung einer zufälligen Veränderlichen betrachtet werden. Deshalb wird der Faktor A mit den Ausprägungen a_i als Random Factor aufgefasst. Es werden nun einige, in der Theorie der Varianzanalyse übliche Annahmen gemacht:

$$\mathbb{E}(w_{ij}) = 0, \;\; \mathbb{E}(a_i w_{ij}) = 0, \;\; \mathbb{E}(w_{ij} w_{ik}) = 0, \tag{2.7.73}$$

so dass sich für den Erwartungswert und die Varianz von x_{ij} die Ausdrücke

$$\mathbb{E}(x_{ij}) = \mu + a_i, \quad \mathrm{Var}(x_{ij}) = \mathrm{Var}(a) + \mathrm{Var}(e) = \sigma_a^2 + \sigma_w^2 \tag{2.7.74}$$

ergeben.

Es werde nun die die Kovarianz von x_{ij} und x_{ik} betrachtet, also die Kovarianz zwischen der Beurteilung von ω_i durch die Experten R_j und R_k:

$$\mathrm{Kov}(x_{ij}, x_{ik}) = \mathbb{E}[(x_{ij} - \mu)(x_{ik} - \mu)]. \tag{2.7.75}$$

$\mathrm{Kov}(x_{ij}, x_{ik})$ ist zunächst als theoretischer Ausdruck gemeint. Wichtig ist, dass die Abweichungen vom allgemeinen Erwartungswert μ betrachtet werden. Setzt man nun die rechte Seite von (2.7.72) für die x_{ij} ein, so erhält man

$$\mathrm{Kov}(x_{ij}, x_{ik}) = \mathbb{E}(\mu + a_i + w_{ij} - \mu)(\mu + a_i + w_{ik} - \mu),$$

d.h.

$$\begin{aligned} \mathrm{Kov}(x_{ij}, x_{ik}) &= \mathbb{E}(a_i + w_{ij})(a_i + w_{ik}) \\ &= \mathbb{E}(a_i^2) + \mathbb{E}(a_i w_{ik}) + \mathbb{E}(a_i w_{ij}) \\ &= +\mathbb{E}(w_{ij} w_{ik}) = \mathbb{E}(a_i^2) \end{aligned}$$

d.h.

$$\mathrm{Kov}(x_{ij}, x_{ik}) = \sigma_a^2. \tag{2.7.76}$$

wegen 2.7.73. Für die Korrelation zwischen x_{ij} und x_{ik} hat man dann

$$\frac{\mathrm{Kov}(x_{ij}, x_{ik})}{\sqrt{\mathrm{Var}(x_{ij})}\sqrt{\mathrm{Var}(x_{ik})}}$$

Aber aus (**??**) hat man $\mathrm{Var}(x_{ij}) = \mathrm{Var}(x_{ik}) = \sigma_a^2 + \sigma_w^2$ unabhängig von ω_i, also hat man schließlich

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_w^2}. \tag{2.7.77}$$

ρ ist als *Intraklassenkorrelationskoeffizient* (IKK) bekannt. Es sei zunächst festgestellt, dass dieser Koeffizient unabhängig von den Indices i und j ist, obwohl sie noch in der Definition der Kovarianz $\mathrm{Kov}(x_{ij}x_{ik})$ auftauchen. Sie verschwinden, weil die Kovarianz eben $\mathrm{Kov}(x_{ij}, x_{ik}) = \sigma_a^2$ unabhängig von den i und j ist, ebenso wie die Varianz der x_{ij}. Die Bezeichung von ρ als *Intrakassen*-Korrelationskoeffizient geht wohl auf Fisher (1938/1956) zurück: betrachtet man

z.B, die Korrelation zwischen Brüdern bezüglich eines Merkmals, so kann sie u.a. vom Alterunterschied zwischen den Brüdern abhängen. Vernachlässigt man derartige Variable und betrachtet nur die Klasse der Brüder, so berechnet man die Korrelation eben nur innerhalb der Klasse der Brüder, – also als Intraklassenkorrelation. Natürlich hat man stets $\sigma_a^2 \geq 0$ und $\sigma_w^2 \geq 0$, so dass $\rho \geq 0$, d.h. ρ kann keine negativen Werte annehmen. Andererseits kann ρ nicht größer als 1 werden, wenn nämlich für $\sigma_a^2 > 0$ die Fehlervarianz σ_w^2 gegen Null geht, denn

$$\lim_{\sigma_w^2 \to 0} \frac{\sigma_a^2}{\sigma_a^2 + \sigma_w^2} \to \frac{\sigma_a^2}{\sigma_a^2} = 1, \tag{2.7.78}$$

oder, für $\sigma_w^2 > 0$,

$$\lim_{\sigma_a^2 \to \infty} \frac{\sigma_a^2}{\sigma_a^2 + \sigma_w^2} = 1 \tag{2.7.79}$$

denn nach Division des Zählers und Nenners durch σ_a^2 hat man

$$\frac{\sigma_a^2}{\sigma_a^2 + \sigma_w^2} = \frac{1}{1 + \sigma_w^2/\sigma_a^2}$$

und $\lim_{\sigma_a^2 \to \infty} \sigma_w^2/\sigma_a^2 = 0$, so dass

$$0 \leq \rho \leq 1. \tag{2.7.80}$$

Dem entspricht die Tatsache, dass ρ ein Varianzanteil ist, d.h. als Anteil der Varianz der a_i-Werte an der Gesamtvarianz $\sigma_a^2 + \sigma_w^2$. In der Tat stellt die Gleichung (2.7.72) ja eine Regressionsgleichung dar:

$$x_{ij} = \mu + a_i + w_{ij} = \alpha a_i + \beta + w_{ij},$$

mit den Regressionskoeffizienten $\alpha = 1$ und $\beta = \mu$. Setzt man $\hat{x}_{ij} = a_i + \mu$, so ist $\mathrm{Var}(\hat{x}_{ij}) = \sigma_a^2$, während $\mathrm{Var}(x_{ij}) = \sigma_a^2 + \sigma_w^2$ ist. Bei der üblichen linearen Regression ist

$$\frac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)} = r^2 = 1 - \frac{\sigma_w^2}{\mathrm{Var}(y)}$$

als Determinationskoeffizient bekannt. Angewandt auf die x_{ij} hat man

$$\frac{\mathrm{Var}(\hat{x}_{ij})}{\mathrm{Var}(x_{ij})} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_w^2} = \rho, \tag{2.7.81}$$

d.h. der Korrelationskoeffizient ρ ist gleichzeitig ein Determinationskoeffizient, nämlich das Quadrat der Korrelation für die Regression von x_{ij} auf die a_i. Für diesen Korrelationskoeffizienten gilt ja, wie aus der Regressionsrechnung bekannt ist,

$$r = \alpha \frac{\sigma_a}{\sigma_{x_{ij}}} = \frac{\sigma_a}{\sqrt{\sigma_a^2 + \sigma_w^2}}$$

und damit $r^2 = \rho$.

ρ_{kj} ist, als Korrelation zwischen den Beurteilern, ein Reliabilitätskoeffizient. Der Koeffizient wird, für $\sigma_e^2 > 0$, groß, wenn σ_a^2 groß wird (vergl. (3.2.101)), d.h. also, wenn die Unterschiede zwischen den a_i groß werden. Man bedenke aber, dass für inferenzstatistische Betrachtungen die Interpretation von ρ als Reliabilitätskoeffizient die für die Varianzanalyse charakteristische Annahme homogener Varianzen gelten muß, die sich in der Aussage $\mathrm{Var}(w_{ij}) = \sigma_w^2$ für alle i, j äußert.

Damit ist der Intraklassenkoeffizient prinzipiell als Reliabilitätskoeffizient erklärt worden. Es muß nun die modellabhängige Form des IKK diskutiert werden. Im Nenner von (2.7.77) steht die Gesamtvarianz von x_{ij}:

$$\mathrm{Var}(x_{ij}) = \mathrm{Var}(a_i) + \mathrm{Var}(w_{ij}) = \sigma_{ges}^2 = \sigma_a^2 + \sigma_w^2,$$

wobei σ_{ges}^2 für 'Gesamtvarianz' steht. Die verschiedenen ANOVA-Modelle unterscheiden sich hinsichtlich der Zerlegung von σ_w^2.

Anwendung der ANOVA-Modelle: Im Modell 1 wird σ_w^2 gar nicht weiter aufgeteilt und ρ nimmt den in (2.7.77) angegebenen Wert an:

$$\rho = \rho_1 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_w^2} \tag{2.7.82}$$

Die Formel für ρ ist nur der systematischen Darstellung wegen wiederholt worden, damit deutlich wird, was von jetzt ab unter ρ_1 verstanden werden soll.

Für das Modell 2 hat man nach (2.7.69)

$$x_{ij} = \mu + a_i + b_j + e_{ij},$$

d.h.

$$\sigma_{ges}^2 = \sigma_a^2 + \sigma_b^2 + \sigma_e^2. \tag{2.7.83}$$

Mithin gilt

$$\rho_2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma_e^2}. \tag{2.7.84}$$

Im Modell 3 schließlich ist der Faktor B mit den Ausprägungen b_j ein Fixed Factor; die b_j liefern also keine Varianzkomponente, oder $\sigma_b^2 = 0$. Mithin ist

$$\sigma_{ges}^2 = \sigma_a^2 + \sigma_e^2, \tag{2.7.85}$$

und man hat

$$\rho_3 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}. \tag{2.7.86}$$

Die Unterschiede zwischen den ρ_1, ρ_2 und ρ_3 ergeben sich durch die Schätzungen für die Varianzkomponenten.

Anmerkungen: Die Mittelwertsunterschiede zwischen den Beurteilern gehen in den "Messfehler" w ein – der IKK ρ heißt dann auch *unjustiert*. Wird die Fehlervarianz von den mittleren Unterschieden der Urteiler bereinigt, so wird die

Fehlervarianz zu $\sigma_a^2 + \sigma_b^2 + \sigma_e^2$, wobei σ_e^2 jetzt die "echte" Fehlervarianz ist. Der IKK heißt nun *justiert*. Die Entscheidung für die Art des IKK hängt von der Art der Interpretation der Schätzskala ab. IKK_{unjust} wird dann verwendet, wenn die absoluten Werte der Schätzungen betrachtet werden. IKK_{just} wird verwendet, wenn die Einschätzungen unabhängig vom Urteiler interpretiert werden; die Varianz zwischen den Schätzern wird dann als Fehlervarianz angesehen.

Beispiel: Zwei Lehrer beurteilen Schüler im gleichen Fach, aber einer der Lehrer benoted konsistent eine Note schlechter als der andere. Die Noten können dann nicht absolut interpretiert werden, da der mittlere Unterschied (ein Notenwert) als "Fehler" betrachtet werden muß. Dementsprechend ist IKK_{unjust} das richtige Reliabilitätsmaß. Die Note eines Schülers sagt nicht nur etwas über seine Leistung, sondern auch über den Beurteiler, also den Lehrer aus. Will man andererseits von jedem Lehrer die k besten Schüler einstellen, so ist der mittlere Unterschied zwischen den Noten irrelevant, da eben von beiden Lehrern die k besten Schüler (etwa für einen Spezialkurs) ausgesucht werden. In diesem Fall wird man IKK_{just} berechnen. Wird ein gewöhnlicher Produktmomentkorrelationskoeffizient als Reliabilitätskoeffizient berechnet, so wird der mittlere Unterschied zwischen den Urteilern *stets* herausgerechnet, dh er entspricht stets IKK_{just}. Bei der Berechung eines IKK kann der Unterschied zwischen justiertem und unjustiertem Koeffizienten gemacht werden. □

In der Tabelle 9 werden die Koeffizienten noch einmal zusammengefasst:

Tabelle 9: Typen von Intraklassenkorrelationen

Typ		σ_{ges}^2	ρ
unjustiert	1-faktoriell	$\sigma_a^2 + \sigma_w^2$	$\rho_1 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_w^2}$
unjustiert	2-faktoriell	$\sigma_a^2 + \sigma_b^2 + \sigma_e^2$	$\rho_2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_b^2+\sigma_e^2}$
justiert	2-faktoriell	$\sigma_a^2 + \sigma_b^2 + \sigma_e^2$	$\rho_3 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$

Schätzungen der Varianzkomponenten: Generell wird bei einer ANOVA die Gesamtvarianz (hier der x_{ij}-Werte) in Komponenten, die auf die Faktoren zurückgehen, und in Komponenten, die auf einen Fehler zurückgehen aufgespalten, d.h. es werden die zu den Varianzen korrespondierenden Quadratsummen zerlegt. Im einfachen, 1-faktoriellen Fall hat man die Zerlegung

$$QS_{ges} = QS_{zw} + QS_{inn},$$

wobei QS für 'Quadratsumme' steht, der Index zw für 'zwischen' den experimentellen Gruppen und inn für 'innerhalb' der Gruppen. Die QS_{inn} repräsentiert den Fehler, QS_{zw} den Effekt der Stufen des Faktors oder der Faktoren im mehrfaktoriellen Fall. Da im Modell 1 der 1-faktorielle Fall mit einem Random Factor betrachtet wird, hat man[35]

$$QS_{inn} = \sum_{i,j}(a_i - \bar{a})^2 + \sum_{i,j}(w_{ij} - \bar{w})^2 \tag{2.7.87}$$

$$QS_{zw} = \sum_{i} n_i(a_i - \bar{a})^2 + \sum_{i}(n_i(w_i - \bar{w})^2 \tag{2.7.88}$$

Wegen $\sum_{i,j}(a_i - \bar{a})^2 = \sum_i n_i(a_i - \bar{a})^2$ gehen die a_i sowohl in QS_{inn} wie in QS_{zw} ein. $\bar{a} <$ und $\bar{w}$ sind die *Stichprobenmittelwerte* der a_i und der w_{ij}.

Tabelle 10: Quadratsummen der ANOVAs; $f = K/(K-1)$, $\theta_a^2 = \sum_i a_i/(K-1)$

Quelle	df	Mod. 1	Mod. 2	Mod. 3
MS_{zw}	$n-1$	$K\sigma_a^2 + \sigma_w^2$	$K\sigma_a^2 + \sigma_b^2 + \sigma_e^2$	$K\sigma_a^2 + \sigma_e^2$
MS_{inn}	$n(K-1)$	σ_w^2	$\sigma_a^2 + \sigma_b^2 + \sigma_e^2$	$\theta_a^2 + f\sigma_b^2 + \sigma_e^2$
$MS_{Urteiler}$	$K-1$	–	$n\sigma_a^2 + \sigma_b^2 + \sigma_e^2$	$n\theta_a^2 + f\sigma_b^2 + \sigma_e^2$
MS_{res}	$(n-1)(K-1)$	–	$\sigma_b^2 + \sigma_e^2$	$f\sigma_b^2 + \sigma_e^2$

$$\rho_1 = \frac{MS_{zw} - MS_{inn}}{MS_{zw} + (K-1)MS_{inn}} \tag{2.7.89}$$

$$\rho_2 = \frac{MS_{zw} - MS_{res}}{MS_{zw} + (K-1)MS_{res} + K(MS_U - MS_{res})/n} \tag{2.7.90}$$

$$\rho_3 = \frac{MS_{zw} - MS_{res}}{MS_{zw} + (K-1)MS_{res}} \tag{2.7.91}$$

Shrout & Fleiss (1979) geben Konfidenzintervalle für die ρ_j an, die man unter den üblichen Voraussetzungen der Varíanzanalyse berechnen kann. Die können nützlich sein, wenn die Koeffizienten so klein sind, dass man an ihrer Signifikanz zweifeln kann. Dann zeigen sie allerding in jedem Fall, dass die Interrater-Reliabilität so klein ist, dass den Einschätzungen nicht sehr getraut werden kann. Eine SPSS-orientierte Darstellung findet man in Wirtz & Caspar (2002).

2.7.7 Die Hauptachsentransformation von Ratings

Die Beurteilerübereinstimmung kann als Korrelation zwischen den Einschätzungen durch verschiedene Urteiler bestimmt werden. Gegeben seien Daten wie in

[35] Eine Herleitung findet man u.a. im Skriptum 'Einführung in die Varianzanalyse'.

der Tabelle 8. Die folgenden Überlegungen zeigen, dass Korrelationen deutlich kleiner als 1 auf eine Mehrdimensionalität der Beurteilten ω_i verweisen.

Es sei Z die (n, K)-Matrix der Einschätzungen x_{ij} der Person oder des Objekts ω_i durch den Beurteiler R_j; Z gehe aus der Matrix X der Einschätzungen x_{ij} durch Spaltenstandardisierung hervor. Für Z (wie natürlich auch für X, aber hier wird nur die von Z betrachtet) gilt stets die Singularwertzerlegung (SVD – *Singular Value Decomposition*)

$$Z = Q\Lambda^{1/2}P', \tag{2.7.92}$$

wobei Q und P orthonormale Matrizen sind; Q ist die Matrix der Eigenvektoren von ZZ', P ist die Matrix der Eigenvektoren von $Z'Z$ und Λ ist die Diagonalmatrix der von Null verschiedenen Eigenwerte, von ZZ' und $Z'Z$ (ZZ' und $Z'Z$ haben identische Eigenwerte). Wegen der unvermeidlichen Messfehler sind die Spalten von Z linear unabhängig, so dass der Rang s von Z gleich $\min(n, K)$ ist; überlicherweise ist $K < n$, so dass $s = K$. Die Orientierung der Spaltenvektoren von P ist gleich der Orientierung der Hauptachsen des durch $Z'Z$ definierten Ellipsoids; $ZP = Q\Lambda^{1/2}$ bedeutet die Rotation der ursprünglichen Koordinatenachsen derart, dass die neuen Koordinaten $L = Q\Lambda^{1/2}$ mit diesen Hauptachsen zusammenfallen; die Spaltenvektoren von L sind orthogonal, so dass der Übergang zu diesen Koordinaten einer Transformation der Hauptachsen entspricht. Die Spalten von L enthalten die Koordinaten der Personen auf unkorrelierten latenten Dimensionen. $A = P\Lambda^{1/2}$ enthält die "Ladungen" der Variablen – hier: der Rater $R_1, \ldots, R_K$ – auf diesen latenten Dimensionen.

Sowohl Q als auch P werden durch die Daten Z bestimmt; (2.7.92) impliziert

$$ZP = Q\Lambda^{1/2}, \quad Z'Q' = P\Lambda^{1/2}, \tag{2.7.93}$$

und man kann sagen, dass wegen dieser wechselseitigen Abhängigkeit – Q wird durch Multiplikation von Z mit P berechnet, und P wird durch Multiplikation von Z' mit Q berechnet – P und Q nicht unabhängig voneinander sind. Unter bestimmten Bedingungen kann man aber erwarten, dass diese beiden Matrizen nicht voneinander abhängen, wie im Folgenden elaboriert wird.

Nach (2.7.92) sind die Ratings x_{ij} durch zwei Parametersätze spezifiziert: die Matrix Q enthält die Ausprägungen eine bestimmten Menge von Merkmalen – die hier als latente Merkmale betrachtet werden können – , und P enhält, in der Sprache der Testtheorie, Aufgabenparameter, die hier als Parameter der Rater auftauchen. Um es ein wenig verquer zu formulieren: Eine beurteilte Person muß gewissermaßen die Aufgabe lösen, von einem Rater beurteilt zu werden, d.h. der Gewichtung der Merkmale durch den Rater möglichst gut zu entsprechen. Denn (2.7.92) besagt ja, dass ein Rating z_{ij} – die Beurteilung der i-ten Person durch den j-ten Rater – durch ein Skalarprodukt gegeben ist:

$$z_{ij} = q_{i1}a_{j1} + \ldots + q_{iK}a_{jK}, \quad a_{jt} = p_{jt}\sqrt{\lambda_t} \tag{2.7.94}$$

$p_{jt}\sqrt{\lambda_t}$ entspricht einer Gewichtung des t-ten latenten Merkmals durch R_j. Nach der Cauchy-Schwarzschen Ungleichung ist z_{ij} maximal relativ zur Länge der Vektoren

$$(q_{i1}, \ldots, q_{iK})', \quad (a_{j1}, \ldots, a_{jK})',$$

wenn $q_{it} = \alpha a_{jt}$ für alle $t = 1, \ldots, K$, $\alpha \in \mathbb{R}$ eine Konstante.

Es gibt zwei Aspekte von (2.7.92), die im Zusammenhang mit der Frage der Urteilerübereinstimmung betrachtet werden können:

1. Die Singularwertzerlegung (2.7.92) impliziert $Z'Z = P\Lambda P'$, d.h. die Korrelationen ($R = Z'Z/n$) sind durch die "Ladungen" $A = P\Lambda^{1/2}$ bestimmt, wobei eine Ladung $\alpha_{jt} = p_{jt}\sqrt{\lambda_t}$ den "Wert" des Raters R_j auf der t-ten latenten Dimension angibt. Man kann diesen Wert als Gewicht, mit dem Rater R_j das t-te latente Merkmal in sein Urteil eingehen läßt, betrachten. Beurteilt ein Rater R_j die zu beurteilenden Personen "sine ira et studio", also "objektiv und neutral", so wird man erwarten, dass die α_{jt}, $t = 1, \ldots, K$ eben unabhängig von den jeweils beurteilten Personen sind. Für verschiedene Stichproben von beurteilten Personen sollte also die Matrix A in guter Näherung konstant sein. Da (2.7.92) impliziert, dass $A = Z'Q$, wird also gefordert, dass die Matrizen Z für die jeweiligen Stichproben so aufgebaut sind, dass die personenspezifischen Aspekte sich im Produkt $Z'Q$ gerade aufheben. Ist dies nicht der Fall, existieren Wechselwirkungen zwischen Rater und beurteilter Person, d.h. der Rater beurteilt bestimmte Merkmale einer Person in Abhängigkeit von personspezifischen Eigenschaften – was nicht sein sollte.

2. Gute Urteilerübereinstimmung impliziert Eindimensionalität der Datenmatrix. Dazu werde für den Moment angenommen, dass die Schätzungen fehlerfrei sind, und die Korrelationen zwischen den Ratern R_j und R_k gleich 1 sind (für alle $j \neq k$). Da die $\mathbf{Z}_j$ und $\mathbf{Z}_k$ dieselbe Länge haben, ist dann $\mathbf{Z}_j = \mathbf{Z}_k$ für alle j, k. Nach (2.7.92) ist $ZA = Q$. Es sei $\mathbf{q}_1$ der erste Spaltenvektor in Q. Dann folgt

   $$\mathbf{Z}_1 a_{11} + \mathbf{Z}_2 a_{21} + \cdots + \mathbf{Z}_K a_{1K} = \mathbf{Z}_1(a_{11} + \cdots + a_{1K}) = \beta_1 \mathbf{Z}_1 = \mathbf{q}_1,$$

   denn die $\mathbf{Z}_j$ sind ja alle identisch. Eine analoge Betrachtung führt zu $\beta_2 \mathbf{Z}_1 = \mathbf{q}_2$. Aber $\mathbf{q}_1$ und $\mathbf{q}_2$ müssen orthogonal sein, d.h. $\mathbf{q}_1'\mathbf{q}_2 = \beta_1\beta_2 \mathbf{Z}_1'\mathbf{Z}_1 = 0$, was nur möglich ist, wenn $\beta_2 = 0$, d.h. es gibt nur eine Dimension (dies folgt natürlich schon unmittelbar aus der Identität der $\mathbf{Z}_j$ und $\beta_1 \mathbf{Z}_1 = \mathbf{q}_1$, – alle weiteren latenten Dimensionen sind ebenfalls proportional zu $\mathbf{Z}_1$ und haben damit die gleiche Orientierung wie $\mathbf{q}_1$, und dies bedeutet ja gerade, dass es nur eine Dimension gibt.

   Dieser Befund bedeutet nicht, dass die Rater nur ein Merkmal beurteilen, $\mathbf{q}_1$ definiert nur einen 1-dimensionalen Teilraum in einem m-dimensionalen Merkmalsraum, und m kann gleich 1 oder gleich irgendeiner großen Zahl

sein. Die Eindimensionalität bedeutet, dass die Gewichtung dieser Merkmale bis auf eine auf alle Merkmale gleichermaßen wirkende Proportionalitätskonstante für alle Rater identisch ist. Wird ein mehr als 1-dimensionaler Raum für die Repräsentation der Rater benötigt, so verweist dieser Befund auf unterschiedliche Gewichtungen der Merkmale durch die verschiedenen Rater.

Der fehlerfreie Fall ist natürlich eine Fiktion, aber wenn die Korrelationen zwischen den Ratern hinreichend groß sind, kann man zumindest eine approximative Eindimensionalität annehmen.

2.7.8 Korrespondenzanalyse und Urteilerübereinstimmung

Maße wie Cohens κ hängen von freien Parametern ab, die eine Interpration des Maßes erschweren, wenn diese Parameter nicht unabhängig vom Maß geschätzt werden können. Es stellt sich die Frage, ob ein einzelnes Maß ein sinnvoller Indikator für die Urteilerübereinstimmung ist.

Für die Analyse der Urteilerübereinstimmung besteht eine Alternative zu einem Maß wie κ in der Korrespondenzanalyse. Notwendiger Ausgangspunkt für diese Analyse ist allerdings eine Kontingenztabelle mit mindestens $K = 4$, die allerdings in vielen Untersuchungen gegeben ist. Die Korrespondenzanalyse besteht in einer Analyse der Residuen

$$x_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}, \quad \hat{n}_{ij} = \frac{n_{i+}n_{+j}}{N}; \tag{2.7.95}$$

Es wird üblicherweise argumentiert, dass die x_{ij} die systematischen Anteile in den Urteilskombinationen enthalten. Aber wie bei Scotts π oder Cohens κ bedeutet die Subtraktion von $\hat{n}_{ij}$ nicht, dass die n_{ij} von irgendwie zufällig zustande gekommenem Urteilen bereinigt werden, – schon weil die Strukturen, die systematische Abhängigkeiten in den n_{ij} erzeugen, natürlich auch in den Randsummen n_{i+} und n_{+j} enthalten sind (vergl. Gln (2.7.43) und (2.7.44)). Die Gleichung (4.9.74) stellt zunächst einmal nur eine Beziehung zum χ^2 der Tabelle her.

Durch eine Hauptachsentransformation der Matrix $X = (x_{ij})$ werden latente Variablen bestimmt, die so skaliert werden, dass die Varianzanteile pro latenter Variable in unabhängige χ^2-Anteile transformiert werden. Man betrachtet also so viele latente Variable, wie nötig sind, um das Gesamt-χ^2 zu erklären. Das Resultat der Korrespondenzanalyse ist ein *Biplot*, in dem die Kategorien $\mathcal{C}_1, \ldots, \mathcal{C}_K$ durch Punkte repräsentiert werden, und zwar einmal dem Gebrauch durch den Urteiler A entsprechend, und zum anderen dem Gebrauch des Urteilers B entsprechend. Liegen die Punkte etwa von $\mathcal{C}_k$ nahe beieinander, so wird $\mathcal{C}_k$ von beiden Urteilern in ähnlicher Weise gesehen, liegen sie weiter auseinander, so werden unterschiedlich Sichtweisen reflektiert. Allerdings können die (euklidischen) Distanzen zwichen zwei Punkten nur dann direkt auf Unterschiede hin interpretiert werden,

wenn sich die Punkte auf denselben Urteiler beziehen. Die euklidischen Distanzen zwischen zwei Punkten, die zu unterschiedlichen Urteilern korrespondieren, können nur über das Skalarprodukt der Vektoren, die vom Koordinatenursprung zu den Punkten zeigen sinnvoll gedeutet werden. Sind $\mathbf{u}$ und $\mathbf{v}$ zwei Vektoren mit der gleichen Anzahl von Komponenten, so gilt für das Skalarprodukt

$$\mathbf{u}'\mathbf{v} = \cos\phi \|\mathbf{u}\| \|\mathbf{v}\|, \tag{2.7.96}$$

wobei ϕ der Winkel zwischen den Vektoren ist und $\|\mathbf{u}\|$ und $\|\mathbf{v}\|$ die Längen der Vektoren sind. Da $\cos\phi \to 1$ für $\phi \to 0$ folgt, dass $\mathbf{u}'\mathbf{v}$ bei gegebener Länge der Vektoren maximal wird, wenn der Winkel zwischen ihnen gleich Null ist. Sofern die latenten Variablen Merkmale repräsentieren, durch die die Kategorien definiert werden, bedeutet also die Tatsache, dass zwei Punkte auf einer Geraden liegen, die durch den 0-Punkt des Koordinatensystems führt, dass die Kategorien durch die Ausprägungen (u_{11}, u_{12}) und (u_{21}, u_{22}) der latenten Merkmale definiert sind mit $u_{21} = au_{11}, u_{22} = au_{12}$ – die Ausprägungen unterscheiden sich nur durch eine beiden Komponenten gemeinsame Proportionalitätskonstante a. Für $\phi \to 1$ strebt $\cos\phi$ gegen Null und die beiden Punkte sind die Endpunkte orthogonaler Vektoren. Dies bedeutet, intuitiv gesprochen, dass die beiden Kategorien bezüglich ihrer Ausprägung auf den latenten Dimensionen nicht miteinander korrelieren. Eine ausführliche Darstellung der Korrespondenzanalyse wird in Mortensen (2010) gegeben.

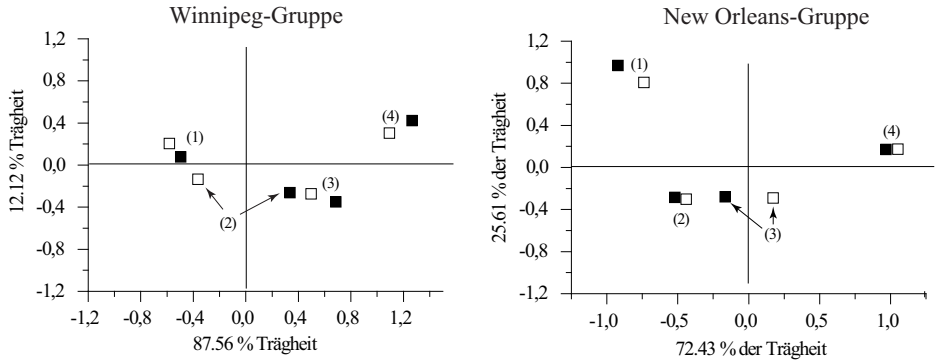
Beispiel 2.7.2 Die Anwendung der Korrespondenzanalyse läßt sich anhand der Daten von Westlund und Kurland (1953) illustrieren, die von Landis & Koch (1977) in Bezug auf das Cohensche κ diskutiert werden. Westlund et al. untersuchten die Diagnosen von Ärzten bezüglich der Möglichkeit der Multiplen Sklerose (MS). Verschiedene Ärzte beurteilten die Symptome von Patienten, indem sie eine von 4 möglichen Kategorien angaben: (1) Sicherlich MS, (2) Wahrscheinlich MS, (3) Möglicherweise MS (Chancen 50 : 50), (4) unwahrscheinlich bzw. sicherlich nicht MS. Es gab zwei Stichproben von Patienten: eine stammte aus Winnipeg (Manitoba), die andere aus New Orleans (Lousiana). Die Winnipeg-Stichprobe wurde zuerst von Ärzten aus Winnipeg beurteilt, und dann wurden die Sympmptome von Ärzten in New Orleans beurteilt. Die Ergebnisse wurden in einer 4×4-Tabelle zusammengefasst: die Zeilenkategorien waren die Katgegorien (1) bis (4), wie sie von den New Orleans-Ärzten benutzt wurden, die Spaltenkategorien waren die gleichen Kategorien, aber wie sie von den Winnipeg-Ärzten verwendet wurden. In analoger Weise ergab sich eine Tabelle für die New Orleans-Patienten, die sich aus den Urteilen der New Orleans-Ärzten (Spalten) und den Winnipeg-Ärzten (Zeilen) ergab. Ganz offensichtlich sind sich die Ärzte nicht immer darüber einig, welche Kategorie die korrekte ist, denn wären sie es, müßten alle Zellen der Tabellen außer denen in der Diagonale leer sein. Die Frage ist, ob die Abweichungungen von perfekter Übereinstimmung rein zufälliger Natur sind oder ob sie einen systematischen Trend enthalten. Diese Frage wurde von Landis

Tabelle 11: Beurteilung zweier Patientengruppen bezüglich der Möglichkeit von Multipler Sklerose. Zeilen der Winnipeg-Gruppe: Beurteilung durch die New Orleans-Neurologen, Zeilen der New Orelans-Gruppe: Beurteilung durch die Winnipeg-Neurologen.

	Winnipeg-Gruppe					New Orleans-Gruppe			
New Orl.	Winnipeg-Neurologe				New Orl.	Winnipeg-Neurologe			
Neurol.	(1)	(2)	(3)	(4)	Neurol.	(1)	(2)	(3)	(4)
(1)	38	5	0	(1)	(1)	5	3	0	0
(2)	33	11	4	0	(2)	3	11	4	0
(3)	10	14	5	6	(3)	2	13	3	4
(4)	3	7	3	10	(4)	1	2	4	14

und Koch (1977) diskutiert. Außer zufälligen Unterschieden in der Beurteilung können sich einerseits die Stichproben voneinander unterscheiden, andererseits können sich die Neurologen aus Winnipeg und die aus New Orleans systematisch in ihren Diagnosen unterscheiden. Auf die statistischen Betrachtungen von Landis und Koch soll hier aus Platzgründen nicht weiter eingegangen werden; statt dessen wird das Ergebnis von Korrespondenzanalysen der beiden Tabellen vorgestellt (vergl. Abschnitt 4.9.8, wo eine kurze Beschreibung des Verfahrens gegeben wird). Die χ^2-Werte beider Tabellen sind hochsignifikant, was eine etwas triviale Feststellung ist, da die Häufigkeiten in den Tabellen deutlich zeigen, dass die beiden Urteilergruppen die jeweiligen Fälle zumindest ähnlich beurteilen. Die Korrespondenzanalyse liefert nun eine graphische Darstellung der Relationen zwischen den Zeilenkategorien einerseits und den Spaltenkategorien andererseits. Die erste Dimension (x-Achse) ist diejenige Dimension, die den größten Teil des χ^2-Wertes einer Tabelle erklärt. Auf dieser Dimension werden die Kategorien (1) bis (4), also von "sicherlich" bis "sicherlich nicht" abgebildet, was zu erwarten ist. In der Winnipeg-Gruppe werden 87.6 % des Gesamt-χ^2 durch diese Dimension erklärt. Bei der zweiten Dimension kann es sich um ein Artefakt ("horse shoe effect") handeln. Bei der New Orleans-Gruppe werden immerhin noch 72.43 % des Gesamt-χ^2 durch die erste Dimension erklärt; wiederum könnte die zweite Dimension ein Artefakt sein. Das Wesentliche an beiden Analysen ist aber, dass die Punkte für die "Eigen"beurteilung (schwarze Kästchen, also die Winnipeg-Patienten durch Winnipeg-Neurologen, die New Orleans-Patienten durch New Orleans-Neurologen) stets dicht beieinander liegen, – mit der Ausnahme der Kategorie (2)(wahrscheinlich MS) bei der Winnipeg-Gruppe, die bei der Fremdbeurteilung dichter an die Kategorie (1) und bei der Eigenbeurteilung dichter an die Kategorie (3) gerät, und bei der New Orleans-Gruppe ist es die Kategorie (3) (möglicherweise MS), die bei der Eigenbeurteilung näher an der Kategorie (2) liegt, während bei der Fremdbeurteilung durch Winnipeg-Neurologen die beiden Kategorien mehr voneinander getrennt erscheinen. □

Abbildung 17: Korrespondenzanalysen der Daten aus der Tabelle 11. Offene Kästen in der Abbildung für die Winnipeg-Gruppe: Beurteilung durch die New Orleans-Neurologen, in der Abbildung für die New Orleans-Gruppe: Beurteilung durch die Winnipeg-Neurologen. Schwarze Kästen: Beurteilung der Winnipeg-Gruppe durch Winnipeg-Neurologen, der New Orleans-Gruppe durch New Orleans-Neurologen. Trägheit: entspricht dem χ^2-Anteil, der durch die Dimension erklärt wird. (1):sicherlich MS, (2): wahrscheinlich MS, (3): möglicherweise MS, (4): unwahrsch. bis sicherlich keine MS. Winnipeg-Daten: $\chi^2 = 64.4$, New Orleans-Daten: $\chi^2 = 44.07$ (beide hoch signifikant).



Das folgende Beispiel illustriert den Fall zweier Klassifikationen, von denen eine als Goldstandard interpretiert wird:

Beispiel 2.7.3 Es soll die Möglichkeit, Erkrankungen der Cervix (Gebärmutterhals) anhand von Bildern zu diagnostizieren, die mittels der *Optical Coherence Tomography* (OCT) gewonnen wurden. Zu diesem Zweck wurden einerseits Gewebeproben entnommen, die nach Maßnahme des histologischen Befundes klassifiziert wurden, und von den gleichen Proben wurden OCT-Bilder gefertigt, die von einer unabhängigen Gutachterin ("Expertin") klassifiziert wurden. Die Klassen waren: 0 = gesund, 10 = Entzündung, CIN 21, CIN 22, CIN 23. Die Klassen CIN 21, CIN 22 und CIN 23 repräsentieren, ranggeordnet nach Ausprägung, Vorformen einer Krebserkrankung. Die Urteile der Histologin (Pathologin) sowie der OCT-Expertin wurden in der folgenden Tabelle zusammengefasst:[36] Es zeigte sich, dass die verschiedenen Kategorien mit unterschiedlicher Häufigkeit identisch verwendet wurden, wie die Abbildung 18 (a) zeigt. Akzeptiert man die Urteile der Histologin als Goldstandard, so bedeuten die Daten aus (a), dass die OCT-Expertin für die Kategorien 0 und 23 relativ häufig "korrekte" Urteile fällt. Für die Kategorien 21 und 22 sinkt der Anteil "korrekter" Urteile auf $\approx .6$, und bei der

[36]Die Daten wurden freundlicherweise von Frau cand. med. Rebecca Gaschler zur Verfügung gestellt.

Tabelle 12: Kontingenztabelle der Urteile (i) anhand des histologischen Befundes (Zeilen) und des OCT-Befundes (Spalten). ($\chi^2 = 261.437$, $df = 16$, $p = .000$, Cohens $\kappa = .65$, Sensitivität = 91%, Spezifität = 78%)

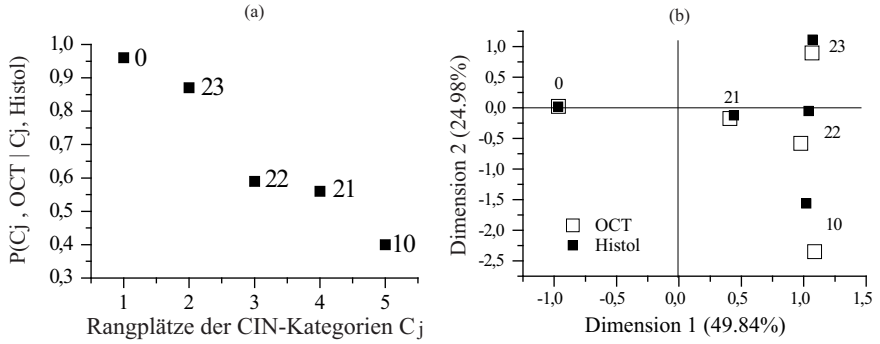
		OCT					
		0	21	22	23	10	Σ
Hist	0	67	3	0	0	0	70
	21	3	13	4	3	0	23
	22	0	1	10	6	0	17
	23	0	2	1	20	0	23
	10	0	2	5	2	6	15
	Σ	70	21	20	31	6	148

Kategorie 10 (Entzündung) sinkt sie gar auf $\approx .4$. Es ergab sich $\kappa = .65$, wobei die Sensitivität der OCT-Expertin bei .91 liegt und ihre Spezifität .78 beträgt. Angesichts der Daten in Abbildung 18 (a) wird deutlich, dass die Sensitivitäts- und Spezifitätswerte eine relativ geringe Aussagekraft haben.

Weitere Einsicht in die Struktur der Urteilsprozesse liefert die auf die Tabelle 12 angewandte Korrespondenzanalyse. Die Tabelle enthält relativ viele 0-Häufigkeiten, so dass angemerkt werden muß, dass die Analyse in erster Linie illustrierenden Charakter hat. Die Abbildung 18 (b) zeigt den resultierenden Biplot. Man kann davon ausgehen, dass es mindestens 2 latente Dimensionen gibt, die den Urteilsprozess bestimmen: diese Dimensionen erkären ca 75% des Gesamt-χ^2 der Tabelle. Die erste Dimension scheint in erster Linie die Schwere der Erkrankung, vermutlich aber die Deutlichkeit, mit der eine Erkrankung diagnostiziert werden kann zu repräsentieren. Die Kategorien 22, 23 und 10 würden dann als "gleich als 'krank' diagnostizierbar" gelten, wobei allerdings eine differenzierende Bewertung bezüglich der zweiten Dimension in die Betrachtung einzubeziehen ist. Von der Kategorie 23 ausgehend zunehmend über die Kategorie 22 zur Kategorie 10 scheint die OCT-Expertin die zweite Dimension als geringer ausgeprägt zu sehen. Die hier repräsentierten Befunde korrespondieren zu Berichten über ausgedehnte Diskussionen über die unterschiedliche Bewertung insbesondere der Kategorie 10.

Die sehr ähnlichen Werte für die bedingten Wahrscheinlichkeiten "korrekter" Urteile für die Kategorien 21 und 22 gehen entweder zurück auf eine Ähnlichkeit dieser Kategorien derart, dass das Erscheinungsbild von Geweben, die zu einer dieser Kategorien gehören, oft keine eindeutigen Entscheidungen zuläßt (visuelles "Rauschen"), oder dass es systematische Unterschiede zwischen den beiden Experten (Histologin versus OCT-Expertin) gibt. Anhand der relativen Häufigkeiten kann nicht zwischen diesen Möglichkeiten unterschieden werden. Der Biplot legt jedoch nahe, dass die Hypothese der Verrauschtheit der Proben und der zu ihnen

Abbildung 18: (a) Übereinstimmung der OCT-Expertin mit der Histologin: Dargestellt werden die Schätzungen von bedingten Wahrscheinlichkeiten $P(C_j, OCT|C_j, Histol)$ von CIN-Kategorisierungen C_j anhand eines OCT-Bildes gegeben die gleiche Kategorisierung durch die Histologin (natürlich ohne explizite Kenntnis der histologischen Diagnose). Die CIN-Kategorien sind nach den Werten dieser bedingten Wahrscheinlichkeiten ranggordnet worden. Für die Kategorien 0 (gesund) und 23 ist die Übereinstimmung recht gut, für die Kategorien 21 und 22 sinkt sie auf $\approx .6$ und für die Kategorie 10 (Entzündung) sinkt sie auf weniger als .4. (b) Korrespondenzanalyse der Klassifikationen anhand histologischer und OCT-Befunde; die beiden Dimensionen erklären ca 75% des Gesamt-χ^2, der Tabelle. Die unterschiedlichen Einschätzungen insbesondere der CIN 10 beruhen offensichtlich auf unterschiedlichen Einschätzungen des durch die Dimension II repräsentierten Merkmals.



korrespondierenden Bilder am ehesten auf die Kategorie 21 zutrifft, die geringe Einigkeit der beiden Experten also auf Zufälligkeiten in der Beurteilung zurückzuführen ist, nicht aber auf konzeptuelle Unterschiede in der Bewertung, denn *im Mittel* wird die Kategorie 21 von beiden Experten als im gleichen Ausmaß durch das Merkmal charakterisiert gesehen, das durch die erste latente Dimension abgebildet wird. Konzeptuelle Unterschiede deuten sich zunächst bezüglich der Kategorie 23 an, die einerseits noch gering zu sein scheinen und die darüber hinaus als nur wenig verrauscht erscheint (die Beurteilungen stimmen in ca 87.5% der Fälle überein). Systematische, also konzeptuelle Unterschiede in der Bewertung erscheinen deutlich ausgeprägter bei der Kategorie 22 und am deutlichsten bei der Kategorie 10. Die zweite Dimension scheint im wesentlichen zwischen krebsartigen Erkrankungen einerseits und der Entzündung andererseits zu differenzieren. Die OCT-Expertin hat demnach im Vergleich zur Histologin eine Tendenz, eher in Richtung der Kategorie 'Entzündung' zu klassifizieren. Es ist klar, dass ein einzelnes Maß wie κ differenziertere Betrachtungen wie diese nicht zuläßt. □

2.7.9 Log-lineare Modelle und Urteilerübereinstimmung

Wie auch bei der Korrespondenzanalyse geht es darum, *Muster* von Übereinstimmungen und systematischen Abweichungen in den Urteilen verschiedener Beobachter zu entdecken. Während die Korrespondenzanalyse im Wesentlichen ein deskriptives Verfahren ist, das eine visuelle Repräsentation dieser Systematik liefert, erlaubt die log-lineare Analyse die inferenzstaistische Bewertung von Interaktionskomponenten – ihr Nachteil ist die mangelnde Anschaulichkeit.

Urteilen die Experten, Diagnostiker etc stochastisch unabhänig voneinander, so sollten sich die Häufigkeiten n_{ij} der Kontingenztabelle gemäß dem allgmeinen Satz $P(A \cap B) = P(A)P(B)$ für unabhängige Ereignisse A und B vorhersagen lassen: Ist A_i das (zufällige) Ereignis, dass Beurteiler A die i-te Kategorie für einen Fall gewählt hat, und ist B_j das zufällige Ereignis, dass Beurteiler B für den gleichen Fall die j-te Kategorie gewählt hat, so ist n_{ij}/N die Schätzung für $\hat{P}(A_i \cap B_j)$ und es sollte

$$\hat{P}(A_i \cap B_j) = \hat{P}(A_i)\hat{P}(B_j) = \frac{n_{i+}}{N}\frac{n_{+j}}{N}$$

für den Fall stochastischer Unabhängigkeit der Urteiler gelten. Für n_{ij} ergibt sich daraus die "Vorhersage"

$$\hat{n}_{ij} = \frac{n_{i+}n_{+j}}{N}, \quad N = \sum_i n_{i+} = \sum_j n_{+j}. \tag{2.7.97}$$

Nun sind die e-Funktion und der Logarithmus (zur Basis e) Umkehrfunktionen zueinander, d.h. für beliebiges x gilt $x = e^{\log x}$. Dann gilt insbesondere

$$\hat{n}_{ij} = e^{\log \hat{n}_{ij}},$$

und für den Fall stochastischer Unabhängigkeit hat man

$$\log \hat{n}_{ij} = \log n_{i+} + \log n_{+j} - \log N = u + u_i + v_j. \tag{2.7.98}$$

Dies ähnelt der Strukturgleichung für eine 2-dimensionale Varianzanalyse *ohne* Wechselwirkung: Die $\log \hat{n}_{ij}$ lassen sich bei stochastischer Unabhängigkeit in zwei "Haupteffekte" u_i und v_j und einen globalen Effekt u zerlegen. Gilt die stochastische Unabhängigkeit *nicht*, so muß man noch einen zusätzlichen Term δ_{ij} schätzen, der die Wechselwirkung zwischen den Kategorien A_i und B_j repräsentiert:

$$\log \hat{n}_{ij} = u + u_i + v_j + \delta_{ij} \tag{2.7.99}$$

δ_{ij} repräsentiert einen systematischen Teil des Urteilsverhaltens. Solche Teile müssen, wenn die Annahme stochastischer Unabhängigkeit nicht gerechtfertigt ist, nicht für alle (i, j) existieren, und man wird ein Modell mit miminaler Anzahl von $\delta_{ij} \neq 0$ suchen. Die gängigen Statistikpakete enthalten Module, mit denen das bestpassende Modell gefunden werden kann. Ein solches Modell kann dann

hilfreich sein, z.B. systematische Fehlbeurteilungen bei einem Urteiler relativ zu einem "Goldstandard" (etwa der Befund der Pathologie versus der Inspektion durch einen Chirurgen bei der Diagnose von Krebserkrankungen) aufzudecken, die nur bei bestimmten Kategorien auftreten; globale Maße wie κ erlauben eine derartige Analyse nicht.

Log-lineare Modelle zur Analyse von Urteilsverhalten sind etwa von Tanner und Young (1985), Agresti (1992) und Graham (1995) vorgeschlagen worden.

2.7.10 Logistische Regression und die Rolle von Kovariaten

Methoden, wie etwa die logistische Regression, erlauben es, bestimmte Hypothesen über die Art systematischer Trends, falls diese existieren, inferenzstatistisch anzugehen. Zuerst scheint dieser Ansatz von Coughlin, Pickel, Goodman und Wilkens (1992) vorgeschlagen worden zu sein. Lipsitz, Parzen, Fitzmaurice und Klar (2003) weisen aber darauf hin, dass bei der Arbeit von Coughlin et al. die Möglichkeit zufälliger Übereinstimmung der Urteile verschiedener Beurteiler nicht adäquat berücksichtigt worden sei und demonstrieren die Implikation dieser Vernachlässigung, nämlich die systematische Verzerrung der Schätzungen der Regressionskoeffizienten, durch die ein falsches Gewicht auf die angebliche Wirksamkeit bestimmter Prädiktoren geworfen wird, d.h. es werden Pseudoabhängigkeiten als genuine, durch irgendwelche Merkmale erzeugte Abhängigkeiten interpretiert. Sie schlagen deswegen einen 2-stufigen Ansatz vor, in dem ein "offset" geschätzt wird, der die zufällige Übereinstimmung berücksichtigt. Die Autoren betrachten als Beispiel die Beurteilung von Verhaltensweisen von Kindern als "psychophathologisch". Eine solche Beurteilung kann durch Eltern erfolgen, oder durch Lehrer. Diese beiden Gruppen von Beurteilern können sich systematisch voneinander unterscheiden, und der Effekt der Zugehörigkeit zu einer dieser beiden Gruppen kann über die logistische Regression modelliert werden. Natürlich kann es auch systematische Unterschiede zwischen den beurteilten Gruppen geben; bei dem in der Tabelle 11 gezeigten Beispiel kann es ja sein, dass die in Winnipeg beurteilten Personen sich systematisch von der Gruppe aus New Orleans unterscheiden. Es kann aber noch eine dritte Variable hinzukommen, nämlich die Zeit, die zwischen den Beurteilungen durch Eltern und Lehrern vergangen ist.

Lipsitz et al. gehen zunächst aus von der

Annahme: Die Beurteilungen durch verschiedene Beurteiler sind unabhängig voneinander, d.h. die Beurteilungen erfolgen ohne Kenntnis der Beurteilungen in der jeweils anderen Gruppe.

Um die Rolle zufälliger und systematischer Effekte modellieren zu können,

werden zunächst die folgenden Indikatorvariablen eingeführt:

$$Y_{ir} = \begin{cases} 1, & i\text{-te Person erhält "positive" Beurteilung durch Urteiler } r, \\ 0, & i\text{-te Person erhält "negative" Beurteilung durch Urteiler } r \end{cases} \tag{2.7.100}$$

wobei $i = 1, \ldots, m$ und $r = 1, 2$. Weiter sei

$$Y_i = \begin{cases} 1, & \text{die Urteiler stimmen bei der } i\text{-ten Person überein}, \\ 0, & \text{die Urteiler stimmen bei der } i\text{-ten Person nicht überein} \end{cases} \tag{2.7.101}$$

Der Wert von Y_i kann durch die Werte der Y_{ir} ausgedrückt werden:

$$Y_i = Y_{i1}Y_{i2} + (1 - Y_{i1})(1 - Y_{i2}). \tag{2.7.102}$$

Denn wenn beide Urteiler "positiv" urteilen, ist $Y_{i1} = Y_{i2} = 1$ und natürlich ist dann auch $Y_{i1}Y_{i2} = 1$. Gleichzeitig gilt $1 - Y_{i1} = 1 - Y_{i2} = 0$, und $Y_i = 1 + 0 = 1$. Urteilen beide "negativ", so ist $Y_{i1}Y_{i2} = 0$ und $1 - Y_{i1} = 1 - Y_{i2} = 1$, so dass wiederum $Y_i = 1$ ist. Urteilen sie aber unterschiedlich, so ist eine der beiden Variablen $Y_{ir} = 0$ und die beiden Produkte werden notwendig gleich Null, also folgt $Y_i = 0$.

Urteilen die Urteiler – wie vorausgesetzt – unabhängig voneinander, so gilt für die Wahrscheinlichkeit einer Übereinstimmung[37]

$$\begin{aligned} P(Y_i = 1) &= P(Y y_{i1} = Y_{i2} = 1 \vee Y_{i1} = Y_{i2} \\ &= P(Y_{i1} = Y_{i2} = 1) + P(Y_{i1} = Y_{i2} = 0) \end{aligned} \tag{2.7.103}$$

denn die Ereignisse $Y_{i1} = Y_{i2} = 1$ und $Y_{i1} = Y_{i2} = 0$ können nicht zugleich auftreten, d.h.

$$P(Y_{i1} = Y_{i2} = 1 \wedge Y_{i1} = Y_{i2} = 0) = 0.$$

Setzt man nun $p_{ir} = P(Y_{ir} = 1)$ und nimmt die Unabhängigkeit der Beurteilungen an, so ist wegen

$$P(Y_{i1} = Y_{i2} = 1) = P(Y_{i1} = 1 \wedge Y_{i2} = 1) = P(Y_{i1} = 1)P(Y_{i2} = 1) = p_{i1}p_{i2}.$$

Analog dazu gilt

$$P(Y_{i1} = Y_{i2} = 0) = (1 - p_{i1})(1 - p_{i2}).$$

Damit erhält man für die Wahrscheinlichkeit einer zufälligen Urteilerübereinstimmung

$$P(Y_i = 1) = p_{i1}p_{i2} + (1 - p_{i1})(1 - p_{i2}). \tag{2.7.104}$$

[37] $\vee$ steht für das einschließende 'oder', $\wedge$ für 'und'.

Beispiel 2.7.4 Zur Illustration werde $p_{i1} = p_{i2} = .1$ angenommen. Dann ist die Wahrscheinlichkeit einer zufälligen Übereinstimmung gleich

$$P(Y_i = 1) = .1^2 + .9^2 = .82.$$

Gilt andererseits $p_{i1} = p_{i2} = .5$, so erhält man

$$P(Y_i = 1) = .5^2 + .5^2 = .5.$$

Man kann nun auch den Fall betrachten, dass zwar die Urteiler das in Frage stehende Merkmal mit gleicher Wahrscheinlichkeit festellen ("Psychophatisches Verhalten") und diese Wahrscheinlichkeit für eine Gruppe – z.B. Altersgruppe – typisch ist, aber für verschiedene Gruppen – verschiedene Altersgruppen – verschieden ist. Gilt also bei einer jüngeren Gruppe von Kindern $p_{i1} = p_{i2} = .1$, bei einer älteren (etwa pubertierenden) aber $p_{i1} = p_{i2} = .5$, so zeigen die Rechnungen, dass die Wahrscheinlichkeit *zufälliger* Übereinstimmung in beiden Gruppen verschieden ist. □

Die Rolle von Kovariaten: Der Wert der Wahrscheinlichkeit $P(Y_i = 1)$ kann sowohl von der beurteilten (i-ten) Person, aber auch von den Urteilern abhängen. Die i-te Person kann durch einen Vektor $\mathbf{x}_i$ charakterisiert werden, dessen Komponenten "Dummy"-Variablen sein können, etwa $x_{i1} = 1$, wenn die Person weiblich ist, und $x_{i1} = 0$, wenn sie männlich ist. Eine zweite Komponente kann ihr Alter angeben, etc. Die Urteiler ("Rater") stammen aus einer Population von Urteilern und können sich ebenfalls voneinander unterscheiden, etwa hinsichtlich ihrer Erfahrung (Jahre gutachterlicher Tätigkeit), hinsichtlich ihres Geschlechts, das einen systematische Einfluß auf die Urteile haben kann, etc. Man kann hierfür urteilerspezifische Vektoren $\mathbf{x}_{ir}$ einführen, wobei hier $r = 1, 2$ gelte. Für das Ereignis $Y_i = 1$ kann nun ein logistischer Regressionsansatz gemacht werden. Es sei

$$p_i = P(Y_i = 1|\mathbf{x}_i, \mathbf{x}_{i1}, \mathbf{x}_{i2}).$$

Man hat dann das logistische Modell

$$\text{Logit}(p_i) = \log \frac{p_i}{1 - p_i} = \mathbf{x}_i'\beta_0 + \mathbf{x}_{i1}'\beta_1 + \mathbf{x}_{i2}'\beta_2. \tag{2.7.105}$$

Die β_j sind zu schätzende Regressionsgewichte. Die Hypothese

$$H_0 : \ \beta_0 = \beta_1 = \beta_2 = 0$$

besagt, dass die Urteilerübereinstimmungen nur zufällig sind, es also keine systematischen Einflüsse (i) der beurteilten Personen und (ii) der urteilenden Personen gibt. Die Wahrscheinlichkeiten p_{ir} in (2.7.104) können explizit in der Form

$$p_{ir} = P(Y_{ir} = 1|\mathbf{x}_{i1}, \mathbf{x}_{i2})$$

geschrieben werden. Hängen die p_{ir} tatsächlich von den Kovariaten $\mathbf{x}_{ir}$ ab, so kann $p_i = P(Y_i = 1)$ mit $\mathbf{x}_i$ und den $\mathbf{x}_{ir}$ kovariieren ($\beta_j \neq 0$), obwohl die Übereinstimmung durch Zufall zustande kommt:

Beispiel 2.7.5 *(***Illustration**) Es soll gezeigt werden, dass ein Regressionsgewicht $\neq 0$ geschätzt werden kann, obwohl die Übereinstimmung rein zufällig und ohne systematische Komponenten ist.

Dazu sei $x_i = 1$, wenn die i-te Person zur jüngeren Gruppe gehört, und $x_i = 0$, wenn sie zur älteren Gruppe gehört. Das logistische Modell (2.7.105) nimmt dann die Form

$$\text{Logit}(p_i) = \beta_0 + \beta_1 x_i. \tag{2.7.106}$$

Insbesondere gelte $\beta_0 = 0$ und $\beta_1 = -2.20$. Ist eine ältere Person ($x_i = 1$) "positiv" beurteilt worden, so ist das diesem Ereignis entsprechende Logit

$$\text{Logit}[p_i(x_i = 1)] = -2.20 x_i = -2.2,$$

und (2.7.106) impliziert

$$p_i(X_i = 1) = e^{-2.2} = .5,$$

und für eine jüngere Person findet man

$$p_i(x_i = 0) = .1,$$

für beide Urteiler. Man erhält, wie schon in der ersten Illustration berechnet, $P(Y_i = 1) = .82$ für die jüngeren Personen, und $P(Y_i = 1) = .5$ für die älteren Beurteilten. Betrachtet man nun den Quotienten der Logits, so hat man

$$\frac{\text{Logit}[P_i(x_i = 1)]}{\text{Logit}[P_i(x_i = 0)]} = \beta_1 = 4.56.$$

Man erhält eine von Null verschiedene Schätzung für das Regressionsgewicht β_1, also für den Altersfaktor, obwohl die Übereinstimmung rein zufällig ist! □

Um von Null verschiedene Schätzungen für Regressionsparameter zu vermeiden, wenn die Übereinstimmungen nur zufällig sind, schlagen Lipsitz et al. eine Modifikation des Ansatzes der logistischen Regression vor. Dazu betrachte man die in (2.7.104) gegebene Wahrscheinlichkeit einer zufälligen Übereinstimmung $P(Y_i = 1)$, und definiere

$$\eta_i = \text{Logit}[P(Y_i = 1)]. \tag{2.7.107}$$

Das logistische Modell (2.7.105) kann nun ersetzt werden durch das Modell

$$\text{Logit}[p_i(x_i)] = \eta_i + \beta_0 + \beta_1 x_i. \tag{2.7.108}$$

Die Schreibweise $p_i(x_i)$ für p_i soll andeuten, dass p_i eine Funktion von x_i ist. Ebenso ist $\eta_i = \eta_(x_i)$ eine Funktion von x_i. Man findet dann

$$\left(\frac{p_i(1)/(1 - p_i(1))}{p_i(0)/(1 - p_i(0))}\right) \frac{\eta_i(0)}{\eta_i(1)} = e^{\beta_1}. \tag{2.7.109}$$

Hierin ist $p_i(1)/(1 - p_i(1))$ die "Wettchance" (Odds) für die "positive" Beurteilung für einer jüngeren Person, und $(1 - p_i(0))$ sind die Odds für eine "positive"

Beurteilung einer älteren Person. Der Quotient dieser Odds und dem Quotienten $\eta_i(1)/\eta_i(0)$, charakterisiert die Chancen einer positiven Beurteilung und jüngeren versus älteren Probanden, wenn die Übereinstimmung nur zufällig ist. Für $\beta_1 = 0$ ist der Quotient in (2.7.109) gleich 1, d.h.

$$\frac{p_i(1)/(1-p_i(1))}{p_i(0)/(1-p_i(0))} = \frac{\eta_i(1)}{\eta_i(0)},$$

d.h. der Quotient auf der linken Seite entspricht dem für zufällige Übereinstimmung.

Die Größe η_i muß geschätzt werden. Dazu muß p_{ir} – die Wahrscheinlichkeit, dass die i-te Person eine "positive" Bewertung durch den r-ten Urteiler erhält – geschätzt werden. Dazu wird das Regressionsmodell

$$\text{Logit}(p_{ir}) = \mathbf{x}'_{ir}\gamma_{1r} + \mathbf{x}'_i\gamma_{2r}, \quad r = 1, 2 \tag{2.7.110}$$

angesetzt. Urteilen die beiden Rater mit ungleichen Wahrscheinlichkeiten, so ist $(\gamma_{11}, \gamma_{21}) \neq (\gamma_{12}, \gamma_{22})$, andernfalls $(\gamma_{11}, \gamma_{21}) = (\gamma_{12}, \gamma_{22})$.

Die Schätzungen für die β- und γ-Parameter erhält man in einem 2-Schrittverfahren:

1. Über eine logistische Regression erhält man eine Schätzung für Y_{ir} mit den Prädiktoren $(\mathbf{x}_i, \mathbf{x}_{ir})$ und damit für $\hat{p}_{ir}$.

2. Man erhält eine Schätzung

$$\hat{\eta}_i = \text{Logit}[\hat{p}_{i1}\hat{p}_{i2} + (1-\hat{p}_{i1})(1-\hat{p}_{i2}). \tag{2.7.111}$$

   Damit wendet man eine logistische Regression für Y_i auf $(\mathbf{x}_i, \mathbf{x}_{i1}, \mathbf{x}_{i2})$ an, wobei $\hat{\eta}_i$ eingesetzt wird. Dies liefert die Schätzungen $(\beta_1, \hat{\beta}_2, \hat{\beta}_3)$.

Lipsitz et al. diskutieren ihren Ansatz in Bezug auf die Diagnose von "Psychopathology" bei (i) 5 – 8 Jahre alten Kindern, und (ii) 9 – 11 Jahre alten Kindern und prüfen insbesondere das Modell

$$\text{Logit}(p_i) = \hat{\eta}_i + \beta_0 + \beta_1\text{AGE}_i + \beta_2\text{Gender}_i + \beta_3\text{TIMELAG}_i. \tag{2.7.112}$$

Die Details sollen hier nicht weiter ausgeführt werden, interessierte Leser können sie in der Arbeit von Lipsitz et al. finden.

2.7.11 Latente Klassen und diagnostische Übereinstimmung

Bei diesem Ansatz wird von der Idee Gebrauch gemacht, dass manifeste – also irgendwie gemessene – Variable oft in Bezug auf latente, also nicht direkt gemessene Variable interpretiert werden, wobei im Allgemeinen die Anzahl der latenten Variablen geringer als die Anzahl der manifesten Variablen ist. Die Faktorenanalyse ist das bekannteste Modell für latente Variable, dessen Anwendung

aber auf metrische Variablen begrenzt ist. Die in diagnostischen Situationen benutzten Antwortkategorien sind aber oft nicht metrisch, sondern nur kategorial. Die folgende Tabelle gibt einen Überblick über Methoden, von Daten auf latente Variablen zu schließen: Es sei wieder $\mathbf{x} = (x_1, \ldots, x_p)'$ der Vektor, der die Sym-

Tabelle 13: Verfahren zu Identifikation latenter Variablen

	Manifeste Var.	
Lat. Variablen	metrisch	kategorial
metrisch	Faktorenanalyse	latent trait analysis; Fakt'analyse von kateg. Variablen
kategorisch	latent profile analysis	latent class analysis
	analysis of mixtures	

ptome repräsentiert, und $\mathbf{y} = (y_1, \ldots, y_q)'$ sei ein Vektor, dessen Komponenten latente Variablen abbilden. $f(\mathbf{x})$ sei die Dichtefunktion für $\mathbf{x}$, und $h(\mathbf{y})$ sei die Dichtefunktion für $\mathbf{y}$. $g(\mathbf{x}|\mathbf{y})$ sei die bedingte Dichte für $\mathbf{x}$, gegeben $\mathbf{y}$. Dann gilt ganz generell

$$f(\mathbf{x}) = \int_{R_y} h(\mathbf{y}) g(\mathbf{x}|\mathbf{y}) d\mathbf{y} \tag{2.7.113}$$

Abhängigkeiten zwischen den manifesten Variablen werden durch die latenten Variablen erzeugt. Ist eine vollständige Menge latenter Variablen gegeben, so sind die x_i statistisch unabhängig, wenn die latenten Variablen konstant gehalten werden, d.h. die Abhängigkeiten zwischen den manifesten Variablen werden durch die Variation der latenten Variablen erzeugt. Diese Deutung der Wirkung latenter Variablen ist die

Annahme der lokalen Unabhängigkeit: Sie spielt in der nicht-klassischen Testtheorie eine bedeutende Rolle. Dann gilt

$$g(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{p} g_i(x_i|\mathbf{y}). \tag{2.7.114}$$

Bartholomew (1987), p. 5, weist darauf hin, dass man diese Annahme nicht empirisch testen kann, da es keine Möglichkeit gibt, $\mathbf{y}$ festzuhalten: In jedem Fall haben die latenten Variablen Werte, die man *per definitionem* nicht direkt beobachten kann, so dass eben kein Test der Annahme durchgeführt werden kann. Der Begriff der lokalen Unabhängigkeit ist denn auch eher eine Art, den Begriff der Vollständigkeit der latenten Variablen zu definieren: Gilt die bedingte Unabhängigkeit der x_i gemäß (2.7.114) nicht, so ist eben die angenommene Zahl latenter

Variablen noch nicht vollständig. Das Ziel ist die Repräsentation von $f(\mathbf{x})$ durch

$$f(\mathbf{x}) = \int_R h(\mathbf{y}) \prod_{i=1}^{p} g_i(x_i|\mathbf{y}) d\mathbf{y}, \tag{2.7.115}$$

wobei also der Ausdruck für $g(\mathbf{x}|\mathbf{y})$ in (2.7.113) durch die rechte Seite von (2.7.114) ersetzt wurde. Diese Notation läßt den Wert q der Anzahl der latenten Variablen verschwinden und ist insofern trügerisch. Jedenfalls wird der kleinste Wert von q gesucht, für den (2.7.115) gilt.

Die x_i sind, im kategorialen Fall, Indikatorvariablen:

$$x_i(k) = \begin{cases} 1, & \text{Fall } i \text{ ist } k\text{-ter Kategorie zugeordnet worden} \\ 0, & \text{sonst} \end{cases} \tag{2.7.116}$$

(2.7.115) wird dann ersetzt durch einen Ausdruck der Form

$$f(\mathbf{x}) = \sum_{k=1}^{K} \eta_k \prod_{i=1}^{N} \pi_i(k)^{x_i(k)}. \tag{2.7.117}$$

wobei η_k die a priori-Wahrscheinlichkeit für die k-te Klasse ist und $\pi_i(k)$ die Wahrscheinlichkeit einer Klassifikation des Falles i in die k-te Klasse ist.

Eine Anwendung auf die Urteilerübereinstimmung findet man z.B. in Übersax & Grove (1990). Das Hauptproblem ist die Schätzung der Parameter; die Diskussion wird hier übergangen.

3 Klassische Testtheorie (KTT)

3.1 Die Annahmen der Klassischen Testtheorie

3.1.1 Gulliksens Axiome

In Abschnitt 2.4 wurde der Gesamtteswert (Gleichung (2.4.1), Seite 22) intuitiv als Messwert im betrachteten Test eingeführt. Die Frage ist, wie ein solcher Testwert interpretiert werden kann; dazu werden in der KTT die Größen Reliabilität und Validität eingeführt. Die Herleitung dieser Größen basiert auf Axiomen, wie sie zuerst von Gulliksen (1950) zusammengestellt wurden. Es läßt sich zeigen, dass diese Axiome sich aus der Annahme $X = \tau + \varepsilon$ (vergl. Gleichung (3.1.1) weiter unten) herleiten lassen und insofern also keine Axiome sind (Lord & Novick (1968), Fischer (1974)). Eine solche Herleitung wird in diesem Skriptum in Abschnitt 3.1.2 skiziiert.

Die Notation der folgenden Darstellung ist an Lord & Novick (1968), Kapitel 2, orientiert. Demnach bezeichnet a eine Person aus der Population $\mathcal{P}$, g ein Item

I_g, $g = 1, \ldots, n$. X_{ag} Testwert der Person $a \in \mathcal{P}$ für das g-te Item[38]. Wird eine beliebige Person aus $\mathcal{P}$ betrachtet, ersetzen Lord & Novick den Index a durch einen Stern, insbesondere schreiben sie X_{*g}, wenn die Antwort einer beliebigen Person auf das Item I_g angezeigt werden soll.

> **Annahmen der KTT:** X_{ag}, und damit X_{*g}, $a = 1, \ldots, m$ und $g, h = 1, \ldots, n$ sind zufällige Veränderliche, für die folgenden Erwartungswerte ($\mathbb{E}$), Varianzen ($\mathbb{V}$) und Kovarianzen ($\mathbb{K}$) existieren mögen.
>
> $$\begin{aligned} X_{*g} &= \tau_{*g} + \varepsilon_{*g} && (3.1.1)\\ \mathbb{E}(\varepsilon_{*g}) &= 0 && (3.1.2)\\ \mathbb{K}(\varepsilon_{*g}, \tau_{*g}) &= 0 && (3.1.3)\\ \mathbb{K}(\varepsilon_{*g}, \tau_{*h}) &= 0 && (3.1.4)\\ \mathbb{K}(\varepsilon_{*g}, \varepsilon_{h*}) &= 0 && (3.1.5) \end{aligned}$$
>
> Hierin sind τ_{*g}, τ_h die *wahren Werte* der Person $*$ für das g-te bzw. h-te Item, und ε_{*g}, ε_{*h} sind die *Messfehler* bei Messungen mit diesen Items. Die Annahmen repräsentieren eine *Messfehlertheorie* des Testens. □

Aus den Annahmen (3.1.1) und (3.1.2) folgt

$$\mathbb{E}(X_{*g}) = \tau_{*g}. \tag{3.1.6}$$

Der wahre Wert ist demnach einfach der Erwartungswert der zufälligen Veränderlichen X_{*g}. Die Annahme (3.1.2) drückt u. a. aus, dass ein möglicherweise existierender systematischer Fehler in den wahren Wert absorbiert wird: hätte man etwa $\mathbb{E}(\varepsilon_{*g}) = b_g \neq 0$, so erhielte man

$$\mathbb{E}(X_{*g}) = \mathbb{E}(\tau_{*g}) + \mathbb{E}(\varepsilon_{*g}) = \tau_{*g} + b_g.$$

Setzt man $\tau'_{*g} = \tau_{*g} + b_g$, $\varepsilon'_{*g} = \varepsilon_{*g} - b_g$, so erhält man $X_{*g} = \tau'_{*g} + \varepsilon'_{*g}$. Benennt man nun τ'_{*g} und ε'_{*g} wieder in τ_{*g} und ε_{*g} um, so ist man wieder bei (3.1.1) und (3.1.2).

Erläuterungen:

1. Die Aussagen (3.1.1) bis (3.1.5) wurden zuerst von Gullicksen (1950) als "Axiome" der KTT eingeführt. Tatsächlich läßt sich zeigen, dass die Aussagen (3.1.2) bis (3.1.5) unter bestimmten, relativ allgemeinen Randbedingungen aus der Annahme (3.1.1) folgen. Herleitungen wurden von Lord &

[38] Die Notation ist leicht verändert: Lord & Novick schreiben für einen Score X_{ga}, wobei a eine Person, g einen Test (gewöhnlich ein Item, das Item I_g) kennzeichnet, während hier X_{ag} geschrieben wird. Der Grund dafür ist, dass Scores oft in einer Marix zusammengefasst werden, wobei einer eingebürgerten Konvention zufolge die Zeilen die Personen, die Spalten die Tests bzw. Items repräsentieren. Da der erste Index eines Matrixelements stets die Zeile, der zweite die Spalte kennzeichnet, in der das Element steht, kommt es zu der Schreibweise X_{ag}.

Novick (1968), Fischer (1970?) und Holland & Hoskins (2003) (s. Abschnitt 3.1.2 unten) geliefert.

2. Die Annahme (3.1.1) entspricht den Annahmen, die den üblichen statistischen Verfahren (Allgemeines Lineares Modell (ALM), also z. B. die Regressionsanalyse) unterliegt.

3. Die Axiome der KTT enthalten keine explizite Annahme über die Wahrscheinlichkeitsverteilung der Fehler ε_{*g}. In Lord & Novick (1968) wird ohne weitere Diskussion davon ausgegangen, dass ε_{*g} und insbesondere auch der Summenscore X_a zumindest approximativ (wegen des Zentralen Grenzwertsatzes) normalverteilt sei. Diese Annahme erlaubt die übliche Berechnung eines Konfidenzintervalles für normalverteilte Variablen für die Schätzung $\hat{\tau}_a$ des wahren Wertes des Summenscores (Gesamttestwerts) einer Person.

   Die Annahme der Normalverteilung für ε_{*g} kann sicherlich nicht in aller Allgemeinheit gemacht werden. Man betrachte dazu den wichtigen Fall dichotomer Items, bei denen $x_{*g} \in \{0, 1\}$ eine Indikatorvariable ist, d.h. $x_{ag} = 1$ zeigt an, dass die Person a das Item I_g "positiv" beantwortet bzw. die Aufgabe gelöst hat. Folgt man der Annahme (3.1.1) und schreibt $x_{*g} = \tau_{*g} + \varepsilon_{*g}$, so folgt $\varepsilon_{*h} = x_{*g} - \tau_{*g}$, d.h.

$$\varepsilon_{*g} = \begin{cases} 1 - \tau_{*g}, & x_{*g} = 1 \\ \\ -\tau_{*g}, & x_{*g} = 0 \end{cases}$$

   Pro Person und Item ist der Fehler also jeweils nur zweier Werte fähig, der mithin nicht normalverteilt sein kann. Sinnvoller ist es, eine unterliegende (und insofern "latente") Variable ξ anzunehmen derart, dass

$$\begin{aligned} x_{*g} &= 1, \quad \text{wenn } \xi > S_g \\ x_{*g} &= 0, \quad \text{wenn } \xi \leq S_g \end{aligned} \tag{3.1.7}$$

   Hier ist S_g ein Schwellenwert für die Ausprägung des Merkmals, das gemessen werden soll: ist das Merkmal zum Zeitpunkt der Messung stärker ausgeprägt als S_g, so wird das Item "positiv" beantwortet; dieser Fall wird durch $x_{*g} = 1$ indiziert. Andernfalls wird das Item "negativ" beantwortet, und dies wird durch $x_{*g} = 0$ indiziert. ξ ist eine zufällige Veränderliche, die normalverteilt sein kann. Je größer etwa der Erwartungswert bei vorgegebener Varianz, desto größer die Wahrscheinlichkeit des Ereignisses $\xi > S_g$. Diese Betrachtung führt allerdings aus der KTT hinaus und in die nichtklassischen (Item-Response) Testtheorien hinein, die in Kapitel 4 behandelt werden.

   Bei einer hinreichend großen Anzahl dichotomer Items läßt sich oft der Summenscore $X_a = \sum_g x_{ga}$ durch eine Normalverteilung approximieren.

Für den Fall, dass I_g eine Ratingskala repräsentiert, ist die Annahme der Normalverteilung nicht immer sinnvoll: werden etwa die Kategorien

$$(-3, -2, -1, 0, 1, 2, 3)$$

vorgegeben und entspricht eine extreme Position den Ansichten einer Person (-3, -1, oder 2 bzw. 3), so bewirken Schätzfehler eher eine Abweichung zur Mitte der Skala, d.h. die Fehlerverteilungen werden asymmetrisch sein. Bei dichotomen Items und Ratingskalen wird die Normalverteilungsannahme also bestenfalls approximativ für den Gesamtwert gelten.

Insgesamt läßt sich sagen, dass die Gullicksen-Axiome sicherlich einer Vorstellung von Messung entsprechen, die unter bestimmten Bedingungen sinnvoll sein kann, wenn es etwa vernünftig ist, die Unabhängigkeit des Messfehlers von der tatsächlichen Merkmalsausprägung anzunehmen und wenn in diesem Zusammenhang eine symmetrische Verteilung der Fehler angenommen werden kann. Ob diese Annahmen ohne Weiteres auch für Fragebogen und Leistungstests gelten können, muß allerdings in Frage gestellt werden. Hierauf wird weiter unten noch eingegangen. Zuerst soll noch der Begriff des wahren Wertes weiter elaboriert bzw. expliziert werden.

Replikationen und Messwiederholungen: Das Testen oder Befragen einer Person entspricht der zufälligen Auswahl einer Person aus der Population, zu der sie gehört. Ist die gewählte Person gerade die Person a, so ist $X_{*g} = X_{ag}$ und $\tau_{*g} = \tau_{ag}$. Da die Person a zufällig gewählt wurde, wird also auch das, was für die Person charakteristisch ist, zufällig gewählt. Für a ist der Erwartungswert τ_{ag} charakteristisch, mithin ist τ_{ag} eine Realisierung der zufälligen Veränderlichen τ_{*g}. Für die zufällige Veränderliche τ_{*g} eine Verteilungsfunktion erklären:

$$G_{*g}(\tau_{*g}) = P(\tau_{*g} \leq \tau). \tag{3.1.8}$$

G_{*g} ist demnach die Wahrscheinlichkeit, zufällig eine Person mit einem wahren Wert τ_{ag} kleiner, höchstens gleich τ zu finden. Für den zufälligen Fehler ε_{*g} läßt sich analog die Verteilungsfunktion

$$H_{*g} = P(\varepsilon_{*g} \leq \varepsilon). \tag{3.1.9}$$

H_{*g} ist die Wahrscheinlichkeit, dass der Fehler der Messung bei der zufällig gewählten Person kleiner, höchstens gleich ε ist. Wählt man eine Anzahl von Personen zufällig aus $\mathcal{P}$ aus und die Wahrscheinlichkeit der wiederholten Auswahl einer Person dabei gleich Null, so stellen die entsprechenden Messungen *Replikationen* von Messungen dar. Der Erwartungswert

$$\tau_g = \mathbb{E}(\tau_{*g}) \tag{3.1.10}$$

ist der Mittelwert über *alle* τ_{ag}. Der Erwartungswert des Fehlers bei der k-ten Replikation werde mit $\mathbb{E}_k(\varepsilon_{ag,k})$ bezeichnet.

Von *Messwiederholungen* ist die Rede, wenn für eine gegebene Person $a \in \mathcal{P}$ die Messungen den Erwartungswert τ_{ag} haben. Nur die Fehler ε_{ag} sind zufällige Größen.

Operationalistische und Platonische Charakterisierung: Die Definition (3.1.6) als erwarteter Wert des "Scores" X ist eine *operationale* oder *operationalistische* Definition; auf die Bedeutung des Ausdrucks 'wahrer Wert' wird nicht weiter eingegangen. Der Begriff 'wahrer Wert' wird nur auf die Stichprobe von Item, die den Test ausmachen, bezogen. Wird das gleiche Merkmal mit einer anderen Stichprobe von Items gemessen, kann sich ein anderer wahrer Wert ergeben. Ein Bezug auf einen wahren Wert, wie er durch den Parameter θ in der Definition 2.6.1, Seite 31, eingeführt wurde, wird in der KTT nicht geliefert, – auch wenn dies möglich wäre. Für eine gegebene Stichprobe von Items $I_1, \ldots, I_n$ wäre dann τ eine Funktion von θ, oder man könnte

$$\tau_a = \tau(\theta_a; I_1, \ldots, I_n) \tag{3.1.11}$$

schreiben. Natürlich wird auch nicht auf eine unterliegende Variable Bezug genommen, so dass τ_a als Erwartungswert dieser unterliegenden Variablen gedeutet werden könnte, denn die Wahrscheinlichkeit einer Antwort wird ja nicht weiter modelliert.

Sutcliffe (1965) hat die operationale Definition von der von ihm so genannten *platonischen* Definition unterschieden. Lord und Novick (1968) erläutern diese Definition anhand des Beispiels eines *chicken sexers*, einer Person also, die feststellen soll, ob ein Küken männlich oder weiblich ist. Demnach hat etwa ein Hähnchen den wahren Wert 1, ein Hühnchen den wahren Wert 0, – oder umgekehrt, die Zuordnung dieser Zahlen ist beliebig, sie muß nur im Folgenden beibehalten werden. Der chicken sexer vergibt den Score $X = 1$, wenn er der Ansicht ist, das Küken sei ein Hähnchen, und $X = 0$ im Fall, dass er der Meinung ist, es sei ein Hühnchen. Die Klassifikation ist nicht fehlerfrei, korrekte Klassifikationen werden nur mit bestimmten Wahrscheinlichkeiten getroffen. Von diesen Überlegungen ausgehend könnte eine Verbindung zur Itemfunktion hergestellt werden, zumal es eine zufällig fluktuierende Zustandsvariable der getesteten Person ist, deren Wert die Antwort bestimmt. Im Rahmen der KTT wird dieser Ansatz jedoch normalerweise nicht aufgenommen.

3.1.2 Die KTT als Spezialfall einer Item-Response-Theorie

In diesem Abschnitt wird ein Ansatz von Holland & Hoskens (2003) vorgestellt, demzufolge sich die KTT als Spezialfall einer Item-response-Theorie darstellen läßt; der Abschnitt ist in erster Linie von theoretischem Interesse und liefert u.a. eine Herleitung der Aussagen (3.1.2), (3.1.3) und des Standarschätzfehlers. Holland et al.'s Ansatz wird nicht vollständig präsentiert, da einige der Entwicklungen über den Rahmen dieses Skriptums hinausgehen.

Holland et al. betrachten zunächst zwei Tests, die die Rohdaten $\mathbf{X}$ und $\mathbf{Y}$ liefern; dies sind zwei Zufallsvektoren, die von einer einzigen Testperson stammen. Weiter sei $G \subseteq \mathcal{P}$ eine Subpopulation von $\mathcal{P}$. G kann ein Geschlecht (männlich, weiblich) bedeuten, oder alle Menschen über 30 Jahre alt, etc. s_x und s_y bezeichnen *scoring functions*, die resultierenden Scores werden mit S_x und S_y bezeichnet: $S_x = s_x(\mathbf{X})$, $S_y = s_y(\mathbf{Y})$. s_x bzw. s_y kann die Anzahl korrekt gelöster Aufgaben sein oder eine irgendwie gewichtete Summe der einzelnen Antworten im Test. Für die scoring functions wird nicht mehr als Eindeutigkeit gefordert; S_x und S_y sind zufällige Veränderliche über $\mathcal{P}$, die die Scores für zufällig gewählte Personen aus $\mathcal{P}$ liefern.

θ_x und θ_y seien latente Variable. Es werden vier Annahmen gemacht:

1. **Eindimensionalität:** $\theta_x, \theta_y \in \mathbb{R}$, d.h. θ_x und θ_y sind keine Vektoren sondern repräsentieren jeweils nur ein Merkmal.

2. **Unabhängigkeit I:** Für beliebige Teilpopulation $G \subseteq \mathcal{P}$,

$$P(\mathbf{X}, \mathbf{Y}|\theta_x, \theta_y, G) = P(\mathbf{X}, \mathbf{Y}|\theta_x, \theta_y), \tag{3.1.12}$$

   d.h. die gemeinsame Verteilung von $\mathbf{X}$ und $\mathbf{Y}$ hängt nur von θ_x und θ_y ab, nicht aber von einer speziellen Subgruppe der Population.

3. **Unabhängigkeit II:**

$$P(\mathbf{X}, \mathbf{Y}|\theta_x, \theta_y) = P(\mathbf{X}|\theta_x, \theta_y)P(\mathbf{Y}|\theta_x, \theta_y), \tag{3.1.13}$$

   d.h. $\mathbf{X}$ und $\mathbf{Y}$ sind bedingt stochastisch unabhängig; diese Annahme entspricht der lokalen stochastischen Unabhängigkeit.

4. **Unabhängigkeit III:**

$$P(\mathbf{X}|\theta_x, \theta_y) = P(\mathbf{X}|\theta_x), \quad P(\mathbf{Y}|\theta_x, \theta_y) = P(\mathbf{Y}|\theta_y). \tag{3.1.14}$$

   Diese Annahme entspricht ebenfalls der lokalen stochastischen Unabhängigkeit, fordert aber darüber hinaus, dass die beiden latenten Variablen die Bedingung der "simple structure" wie sie in der Theorie Faktorenanalyse definiert wird erfüllen.

Für die wahren Werte τ_x und τ_y soll gelten

$$\tau_x = \tau_x(\theta_x) = \mathbb{E}(S_x|\theta_x), \quad \tau_y = \tau_y(\theta_y) = \mathbb{E}(S_y|\theta_y). \tag{3.1.15}$$

Wegen (3.1.12) (Unabhängigkeit I) gilt dann auch

$$\tau_x = \mathbb{E}(S_x|\theta_x, G), \quad \tau_y = \mathbb{E}(S_y|\theta_y, G) \tag{3.1.16}$$

für beliebiges $G \subseteq \mathcal{P}$. Analog zu den *item characteristic functions* (Itemfunktionen in diesem Skript) heißen die $\tau_x(\theta_x)$ und $\tau_y(\theta_y)$ *test characteristic functions*

von $\mathbf{X}$ und $\mathbf{Y}$. Es wird angenommen, dass sie stetige und streng monotone Funktionen von θ_x bzw. θ_y sind.

Es mögen nun die üblichen Annahmen der KTT gelten, wobei X_{ag} oder X_a durch S_X ersetzt wird; das ist nicht weiter wesentlich, betont nur, dass S_x irgendeine *scoring function* ist. Für Y_{ag}, Y_a wird S_y geschrieben. Generell hat man dann

$$S_x = \tau_x + \varepsilon_x, \tag{3.1.17}$$

wobei der Fehlerterm durch

$$\varepsilon_x = S_x - \tau_x \tag{3.1.18}$$

definiert wird. Weiter gilt

$$\mathbb{E}(S_x|\theta_x) = \tau_x, \tag{3.1.19}$$

und die bedingte Varianz von S_x wird durch

$$\mathbb{V}(S_x|\tau_x) = \sigma^2_{S_x}(\tau_x) \tag{3.1.20}$$

definiert. Dann gilt

Satz 3.1.1 *Es gilt*

$$\begin{aligned} \mathbb{E}(\varepsilon_x|\tau_x) &= 0 && (3.1.21)\\ \mathbb{V}(\varepsilon_x|\tau_x) &= \mathbb{V}(S_x|\tau_x) = \sigma^2_{S_x}(\tau_x) && (3.1.22)\\ \sigma^2_{\varepsilon_x|G} &= \mathbb{V}(\varepsilon_x|G) = \mathbb{E}(\sigma^2_{S_x}(\tau_x|G) && (3.1.23) \end{aligned}$$

Beweis: Es ist

$$\mathbb{E}(\varepsilon_x|\tau_x) = \mathbb{E}(S_x - \tau_x|\tau_x) = \mathbb{E}(S_x|\tau_x) - \mathbb{E}(\tau_x|\tau_x) = \tau_x - \tau_x = 0.$$

Weiter hat man

$$\mathbb{V}(\varepsilon_x|\tau_x) = \mathbb{V}(S_x - \tau_x|\tau_x) = \mathbb{V}(S_x|\tau_x) = \sigma^2_{s_x}(\tau_x),$$

und schließlich

$$\mathbb{V}(\varepsilon_x|G) = \mathbb{E}(\mathbb{V}(\varepsilon_x|G)) + \mathbb{V}(\mathbb{E}(\varepsilon_x|G, \tau_x)) = \mathbb{E}(\sigma^2_{S_x}(\tau_x|G)) + \mathbb{V}(0|G),$$

so dass

$$\mathbb{V}(\varepsilon_x|G = \mathbb{E}(\sigma^2_{S_x}(\tau_x)|G).$$

□

Anmerkung: Die Aussage (3.1.21) impliziert

$$\mathbb{E}(\varepsilon_x|G) = 0, \quad \text{für irgend ein } G \subseteq \mathcal{P}. \tag{3.1.24}$$

Satz 3.1.2 *Es gilt*

$$\mathbb{K}(\varepsilon_x, \tau_x|G) = 0. \tag{3.1.25}$$

Beweis: Man hat

$$\begin{aligned}\mathbb{K}(\varepsilon_x, \tau_x|G) &= \mathbb{E}(\varepsilon_x\tau_x|G) - \mathbb{E}(\varepsilon_x|G)\mathbb{E}(\tau_x|G) - \mathbb{E}(\varepsilon_x|G)\mathbb{E}(\tau_x|G)\\ &= \mathbb{E}[\mathbb{E}(\varepsilon_x\tau_x|\tau_x, G)|G] - 0\\ &= \mathbb{E}[\tau_x\mathbb{E}(\varepsilon_x|\tau_x, G)|G] = 0.\end{aligned}$$

□

Anmerkungen: (3.1.25) ist die Aussage (3.1.3), die hier bewiesen wird. (3.1.25) impliziert sofort

$$\mathbb{K}(\varepsilon_x, S_x|G) = \mathbb{K}(\varepsilon_x, \varepsilon_x|G) = \mathbb{V}(\varepsilon_x|G) = \sigma^2_{S_x}(\tau_x) \tag{3.1.26}$$

$$\mathbb{K}(S_x, \tau_x|G) = \mathbb{K}(\tau_x, \tau_x|G) = \mathbb{V}(\tau_x|G) = \sigma^2_{\tau_x|G} \tag{3.1.27}$$

$$\sigma^2_{S_x|G} = \sigma^2_{\varepsilon_x|G} + \sigma^2_{\tau_x|G} \tag{3.1.28}$$

In der KTT ist $\sigma_{\varepsilon_x|G}$ der unbedingte StandardMessfehler, wie im folgenden Abschnitt noch gezeigt werden wird, während die Gleichung (3.1.22) bedeutet, dass $\sigma_{\varepsilon_x}(\tau_x) = \sigma_{S_x}(\tau_x)$ der bedingte StandardMessfehler ist. $\sigma_{\varepsilon_x|G}$ ist eine Zusammenfassung von $\sigma_{S_x}(\tau_x)$. □

3.2 Reliabilität und ihre Abschätzung

3.2.1 Zum Begriff der Reliabilität

Auf Seite 9 ist bereits angedeutet worden, was unter Reliabilität zu verstehen sein soll: die Fehlervarianz des Tests soll klein sein im Vergleich zur Gesamtvarianz des Tests. Diese intuitive Definition läßt sich in eine formale Definition überführen, da ja anhand des Messwerts der wahre Wert "vorausgesagt werden soll: Die Gleichung $X = \tau + \varepsilon$ ist ja eine lineare Regressionsgleichung, wie man sieht, wenn man sie in der üblichen Form einer Regressionsgleichung schreibt:

$$x = \beta\tau + \alpha + \varepsilon,$$

wobei aber $\beta = 1$ und $\alpha = 0$ gelten soll. Es sei $\sigma^2_\tau = \mathbb{V}(\tau)$, $\sigma^2_x = \mathbb{V}(X)$, $\sigma^2_\varepsilon = \mathbb{V}(\varepsilon)$. Für die Korrelation zwischen den τ- und den X-Werten gilt allgemein

$$\rho_{x\tau} = \beta\frac{\sigma_\tau}{\sigma_x},$$

und wegen $\beta = 1$ hat man

$$\rho^2_{x\tau} = \frac{\sigma^2_\tau}{\sigma^2_x}. \tag{3.2.1}$$

Andererseits gilt für $\beta = 1$

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_\varepsilon^2,$$

woraus nach Divsion durch σ_x^2

$$1 = \frac{\sigma_\tau^2}{\sigma_x^2} + \frac{\sigma_\varepsilon^2}{\sigma_x^2}$$

folgt, so dass (3.2.1)

$$\rho_{x\tau}^2 = \frac{\sigma_\tau^2}{\sigma_x^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_x^2} \tag{3.2.2}$$

ergibt. Für $\sigma_\varepsilon^2/\sigma_x^2 \to 0$ folgt also $\rho_{x\tau}^2 \to 1$, und für $\sigma_\varepsilon^2/\sigma_x^2 \to 1$ folgt, $\rho_{\tau X}^2 \to 0$. Der Determinationskoeffizient $\rho_{\tau X}^2$ drückt also genau das aus, was man intuitiv unter Reliabilität versteht. Dementsprechend hat man die

Definition 3.2.1 *Die Reliabilität eines Tests ist durch den Quotienten* $\mathbb{V}(\tau)/\mathbb{V}(X)$*, d.h. durch*

$$\mathit{Rel}(X) \stackrel{\mathit{def}}{=} \frac{\mathbb{V}(\tau)}{\mathbb{V}(X)} = \rho_{x\tau}^2 \tag{3.2.3}$$

definiert.

Diese Definition spezifiert zwar eine Maßzahl für die Reliabilität, erlaubt aber noch nicht, sie auch zu berechnen, da die τ-Werte nicht kannt sind, – also kann auch $\mathbb{V}(\tau)$ nicht unmittelbar berechnet werden. Um $\mathrm{Rel}(X)$ berechnen zu können, muß eine zu (3.2.3) äquivalente Formel gefunden werden.

Intuitive Betrachtungen zur Bestimmung von $\mathbb{V}(\tau)$**:** Man stelle sich vor, es seien zwei Tests mit Messwerten X bzw. X' gegeben, die beide dasselbe Merkmal erfassen. Die Korrelation zwischen X und X' könnte unter bestimmten Randbedingungen eine Abschätzung von $\mathbb{V}(\tau)$ liefern. Denn es sei $X = \tau + \varepsilon_x$ und $X' = \tau' + \varepsilon_{x'}$. Man muß nun bedenken, dass hier τ und τ' zufällige Veränderliche sind, die bei der zufälligen Wahl einer Person $a \in \mathcal{P}$ die Werte τ_a und τ_a' annehmen. Weiter sind $\mathbb{E}(X) = \mathbb{E}(\tau) = \tau_x$, und $\mathbb{E}(X') = \mathbb{E}(\tau') = \tau_{x'}$ die jeweiligen Erwartungswerte von τ bzw. τ', und $\mathbb{E}(\varepsilon_x) = \mathbb{E}(\varepsilon_{x'}) = 0$. Für die Korrelation $\rho(X, X')$ erhält man

$$\rho(X, X') = \rho_{xx'} = \frac{\mathbb{K}(X, X')}{\sigma_x \sigma_{x'}},$$

mit

$$\mathbb{K}(X, X') = \mathbb{E}[(X - \mathbb{E}(X))(X' - \mathbb{E}(X')] = \mathbb{E}(XX') - \mathbb{E}(X)\mathbb{E}(X')$$

und

$$\sigma_x^2 = \mathbb{V}(X) = \mathbb{V}(\tau) + \mathbb{V}(\varepsilon_x), \quad \sigma_{x'}^2 = \mathbb{V}(X') = \mathbb{V}(\tau') + \mathbb{V}(\varepsilon_{x'}).$$

Weiter ist

$$\mathbb{E}(XX') = \mathbb{E}[(\tau + \varepsilon_x - \tau_x)(\tau' + \varepsilon_{x'} - \tau_{x'})] = \mathbb{E}(\tau\tau') - \tau_x\tau_{x'},$$

wie man durch Ausmultiplizieren und Berücksichtigung von $\mathbb{E}(\varepsilon) = \mathbb{E}(\varepsilon') = 0$ leicht überprüft. Man betrachte nun den speziellen Fall, dass die Tests so geartet sind, dass die Bedingungen

1. $\tau = \tau'$, und
2. $\mathbb{V}(\varepsilon) = \mathbb{V}(\varepsilon')$

gelten; die beiden Tests heißen dann *parallel* (vergl. Abschnitt 3.2.2, Definition 3.2.3). Dann ist auch $\tau_x = \tau_{x'} = \mathbb{E}(\tau)$ und man hat

$$\mathbb{E}(XX') = \mathbb{E}(\tau^2) - \mathbb{E}^2(\tau) = \mathbb{V}(\tau) \tag{3.2.4}$$

und

$$\sigma_x \sigma_{x'} = \sigma_x^2 = \mathbb{V}(X).$$

Daraus folgt die Beziehung

$$\rho_{xx'} = \frac{\mathbb{V}(\tau)}{\mathbb{V}(X)} = \rho_{x\tau}^2 = \mathrm{Rel}(X), \tag{3.2.5}$$

wobei $\mathrm{Rel}(X)$ die Reliabilität des Tests X bedeutet. Die Korrelation zwischen den Tests X und X' ist also gleich der Reliabilität, – vorausgesetzt die genannten Bedingungen gelten. Da die Korrelation $\rho_{xx'}$ tatsächlich berechnet – zumindest an einer Stichprobe abgeschätzt – werden kann, läßt sich die Reliabilität berechnen, obwohl die Varianz $\mathbb{V}(\tau)$ nicht explizit bekannt ist. Man bemerke, dass $\rho_{xx'} = \rho_{x\tau}^2$ gilt; die Korrelation zwischen zwei parallelen Tests ist gleich dem Determinationskoeffizienten für die Beziehung zwischen den Messwerten X und den zuegehörigen wahren Werten τ.

Die genannten Bedingungen sind allerdings in der Tat speziell, und ob sie für zwei Tests zutreffen oder nicht, muß empirisch entschieden werden. Dazu werden die folgenden Begriffe definiert.

3.2.2 Kongenerische, τ-äquivalente und parallele Tests

Die Notwendigkeit, für die Reliabilität eine Maßzahl angeben zu können, hat durch eine intuitive Betrachtung auf den Begriff paralleler Tests geführt: sind zwei Tests parallel, so liefert der Determinationskoeffizient $\rho_{xx'} = \rho_{x\tau}^2 = \mathbb{V}(\tau)/\mathbb{V}(X)$ ein Maß für die Reliabilität. Allerdings ist die Forderung, dass die Tests mit den Messwerten X und X' parallel sein müssen, um die Berechnung der Reliabilität zu ermöglichen, einigermaßen radikal. Denn man muß zu einem gegebenen Test mit den Messwerten X zunächst einmal einen Paralleltest konstruieren, also einen Test mit den gleichen Varianzen $\mathbb{V}(\varepsilon)$ und überdies gleichen Erwartungswerten. Die Frage ist also, ob die Forderung nach Parallelität nicht relaxiert werden kann.

Eine erste Verallgemeinerung wurde von Novick & Lewis (1967) vorgeschlagen. Sie führten den Begriff der *τ-äquivalenten Messung* ein, vergl.auch Lord

& Novick (1968) (p. 47). Dies sind Messungen mit identischen wahren Werten, aber möglicherweise verschiedenen Varianzen; viele der relevanten Statistiken der Testtheorie ergeben sich bereits für τ-äquivalente Messungen.

Eine weitere Verallgemeinerung wurde von Jöreskog (1970, 1971) vorgenommen. Gegeben seien wieder zwei Tests mit Messwerten X und X', die bis auf Messfehler das gleiche Merkmal erfassen und bei denen sich die wahren Werte nur durch eine lineare Transformation unterscheiden. Solche Tests heißen in der Testtheorie *kongenerisch*[39]. Die formale Definition solcher Tests ist

Definition 3.2.2 *Es seien $X_1, \ldots, X_n$ zufällige Veränderliche, die die Testscores für n Tests $T_1, \ldots, T_n$ repräsentieren. Es gelte wieder $X_j = \tau_j + \varepsilon_j$, $j = 1, \ldots, n$, wobei $\mathbb{E}(X_j) = \tau_j$ der wahre Wert im j-ten Test sei und $\mathbb{K}(\tau_j, \varepsilon_j) = 0$ und $\mathbb{K}(\varepsilon_j, \varepsilon_k) = 0$ für alle j und $j \neq k$ gelte. Weiter repräsentiere τ_j den wahren Wert einer Person im Test T_j; τ_j ist eine zufällige Veränderliche in dem Sinne, in dem die Person zufällig aus $\mathcal{P}$ gewählt wird. Gilt nun*

$$\tau_j = \nu_j + \lambda_j \eta, \quad j = 1, 2, \ldots, n \tag{3.2.6}$$

für, so heißen die n Tests kongenerisch.

Tests sind also kongenerisch, wenn sie das gleiche Merkmal messen und die wahren Werte jeweils lineare Transformationen der Merkmalsausprägung η sind; die Forderung $\mathbb{K}(\tau_j, \varepsilon_j) = 0$ ist ja eigentliche keine Forderung, sondern eine Konsequenz der Definition von ε_j als Abweichung des Messwerts von τ_j, aber die Forderung $\mathbb{K}(\varepsilon_j, \varepsilon_k) = 0$ ist eine echte Forderung. Die folgenden Klassen von Tests können als Spezialfälle kongenerischer Tests aufgefasst werden:

Definition 3.2.3 *Die Tests $T_1, \ldots, T_n$ mit den Scores $X_1, \ldots, X_n$ mit $X_j = \tau_j + \varepsilon_j$, $j = 1, \ldots, n$ seien kongenerisch.*

1. *Es gelte $\lambda_j = 1$, $\nu_j \neq 0$. Weiter gelte $\mathbb{K}(\varepsilon_j, \varepsilon_k) = 0$. Dann heißen die Tests* essentiell τ-äquivalent.

2. *Es gelte $\lambda_j = 1$ und $\nu_j = 0$, d.h. es gelte $\tau_j = \tau_k$. Dann heißen die Tests T_j und T_k τ-äquivalent. Gilt weiter $\mathbb{V}(\varepsilon_j) = \mathbb{V}(\varepsilon_k)$ und $\mathbb{K}(\varepsilon_j, \varepsilon_k) = 0$, so heißen T_j und T_k* parallel.

Läßt man etwa die Forderung nach gleichen Varianzen fallen, so fordert man zunächst einmal nur τ-Äquivalenz; die τ-Skalen der beiden Tests sind aber immer noch gleich. Fordert man nur die essentielle τ-Äquivalenz, dürfen sich die τ-Skalen noch um eine additive Konstante unterscheiden. Die Forderung nach unabhängigen Fehlern, also $\mathbb{K}(\varepsilon_j, \varepsilon_k) = 0$, muß aber aufrechterhalten werden.

[39] Wohl eine eingedeutschte Form des englischen *congeneric* mit der Bedeutung *gleiartig, gleich geartet*; weder im Duden, noch im Wahrig noch in anderen Fremdwörterbüchern wird das Wort kongenerisch aufgeführt.

Sind X_j und X_k kongenerische Messungen, so sind sie durch eine lineare Transformation aufeinander bezogen, d.h.es gilt

$$X_j = \alpha X_k + \beta, \tag{3.2.7}$$

für geeignet gewählte Zahlen α und β.

Denn da X_j und X_k kongenerisch sind, gilt

$$X_j = \nu_j + \lambda_j \eta \tag{3.2.8}$$
$$X_k = \nu_k + \lambda_k \eta \tag{3.2.9}$$

und η repräsentiert das von beiden Tests erfaßte Merkmal. Man kann beide Gleichungen nach η auflösen:

$$\eta = \frac{X_j - \nu_j}{\lambda_j}, \quad \eta = \frac{X_k - \nu_k}{\lambda_k}. \tag{3.2.10}$$

Da die beiden Quotienten gleich η sind, sind sie auch untereinander gleich:

$$\frac{X_j - \nu_j}{\lambda_j} = \frac{X_k - \nu_k}{\lambda_k}. \tag{3.2.11}$$

Multipliziert man beide Seiten mit λ_j und addiert ν_j, so erhält man

$$X_j = \frac{\lambda_j}{\lambda_k}(X_k - \nu_k) + \nu_j = \frac{\lambda_j}{\lambda_k} X_k + \nu_j - \frac{\lambda_j}{\lambda_k}\nu_k \tag{3.2.12}$$

Setzt man

$$\alpha = \lambda_j/\lambda_k, \quad \beta = \nu_j - \lambda_j/\lambda_k, \tag{3.2.13}$$

so hat man die lineare Beziehung (3.2.7). □

Ein paar einfache Beispiele mögen den Begriff der kongenerischen Messungen veranschaulichen.

Beispiel 3.2.1 (Temperaturmessungen) Temperaturmessungen sind zwar keine psychologischen Messungen, aber sie können helfen, eine eher inhaltliche Vorstellung von kongenerischen Messungen zu bekommen. Bekanntlich liefern sowohl die Celsius- wie auch die Fahrenheit-Skala Messwerte auf einer Intervallskala, die durch eine lineare Beziehung der Art

$$X[C^o] = \alpha Y[F^o] + \beta \tag{3.2.14}$$

aufeinander bezogen sind; $X[C^o]$ sind die Messungen in Celsius, $Y[F^o]$ sind die Messungen in Fahrenheit. Natürlich kann man auch die Fahrenheit-Messungen als lineare Funktion der Celsius-Messungen ausdrücken. Beide Messungen sind Messungen der Temperatur η, nur auf verschiedenen Skalen. Man könnte η etwa als Messung auf der Kelvin-Skala interpretieren. Dann gibt es jedenfalls Skalenparameter derart, dass

$$X[C^o] = \nu_x + \lambda_x \eta, \quad Y[F^o] = \nu_y + \lambda_y \eta,$$

aus der sich die lineare Beziehung (3.2.14) herleiten läßt analog zur Herleitung von (3.2.7). Die Messungen in Celsius und Fahrenheit sind kongenerisch.

Die Parameter ν_x und λ_x ergeben sich aus der Entscheidung Celsius', den Nullpunkt seiner Skala mit dem Gefrierpunkt des Wassers gleichzusetzen, und die Skaleneinteilung so zu wählen, dass $100C^o$ dem Siedepunkt des Wassers engtspricht. Es sei η_o die Temperatur in Kelvin, bei der Wasser gefriert, und η_s sei die Temperatur, bei der Wasser siedet. Die Differenz $\eta_s - \eta_o$ entspricht der auf der Celsius-Skala, also ist $\eta_s - \eta_o = 100$. Offenbar folgt dann $\lambda_x = 1$. Fahrenheits Entscheidung war, den Gefrierpunkt bei $32F^o$ und den Siedepunkt bei $212F^o$ zu setzen (warum auch immer!). Daraus errechnet man $\alpha \approx .529$ und $\beta \approx -16.931$ für die Umrechnung von Fahrenheit in Celsius, und für die Umrechnung von Celsius in Fahrenheit erhält man $\alpha' = 1/\alpha$, $\beta' = -\beta/\alpha$.

Hätte Fhrenheit ebenfalls $\lambda_y = 1$ gewählt, so hätte man für die Fahrenheit-Skala für den Gefrierpunkt 32^0 [F], aber $\nu_y + \lambda_y\eta_s = \nu_y + \eta_s = 32 + 100 = 132$ [F]. Celsius- und Fahrenheit-Messungen wären dann essentiell τ-äquivalent. Hätte Fahrenheit außerdem noch $\nu_y = \nu_x$ gewählt, so wären die Messungen τ-äquivalent. Sie wären aber noch nicht notwendig parallel. Denn Fahrenheit hätte sein Messgerät eventuell aus anderem Material gebaut, das auf Schwankungen der Umgebung stärker (oder weniger) reagiert als das Gerät von Celsius. Dann hätten die Messfehler unterschiedliche Varianzen; obwohl man davon ausgehen kann, dass die Messfehler von Celsius und die von Fahrenheit unabhängig voneinander sind, hätten die beiden unterschiedliche Varianzen, ihre Tests wären also nicht parallel.

□

Beispiel 3.2.2 (Psychophysiologische Messungen) Psychischer Stress soll einmal über Hormonkonzentrationen im Speichel (X) und zum anderen über den galvanischen Hautwiderstand (Y) gemessen werden. η sei der erlebte Stress. Die erste Frage ist, ob die Messungen X und Y überhaupt kongenerisch sind, denn kongenerische Messungen bedeuten ja, dass sowohl X als auh Y lineare Funktionen von η sind. So ist denkbar, dass der galvanische Hautwiderstand Y eine nichtlineare Funktion vo η ist. Y könnte eine positiv beschleunigte Funktion von η sein, so dass etwa $Y = y_0e^{ax}$ mit $a > 0$ gilt, y_0 eine Konstante, oder Y ist negativ beschleunigt und strebt gegen einen Maximalwert, wie etwa $y = 1 - e^{-ax}$, $a > 0$. Die Annahme einer linearen Beziehung bedarf also sicherlich einer empirischen Überprüfung.

Es werde nun angenommen, die Messungen X und Y seien kongenerisch. Dann ist auch die Beziehung zwischen den X- und Y-Werten linear; erweist sie sich als nichtlinear, sind Zweifel an der Kongenerität angebracht. Gilt insbesondere $\lambda_x = \lambda_y$ aber $\nu_x \neq \nu_y$, so sind die Messungen überdies essentiell τ-äquivalent. Dies bedeutet, dass die X- und Y-Messungen einen gegebenen Bereich (η_0, η_s) in gleicher Weise aufteilen und sich nur um eine additive Konstante unterscheiden: die Differenz der Hormonkonzentrationen $\Delta X = X_2 - X_1$ reflektiert dann

den gleichen Unterschied von Stress wie die Differenz $\Delta Y = Y_2 - Y_1$, sofern ΔX und ΔY numerisch gleich sind. Gilt überdies $\nu_x = \nu_y$, so sind die Messungen τ-äquivalent. Sofern aber die Messfehler bei den Hormonmessungen einerseits und den Widerstandsmessungen andererseits unterschiedlich streuen (die Messinstrumente können ja nicht identisch sein), sind die Messungen nicht parallel. Erst wenn diese Streuungen bzw. Varianzen gleich groß sind, sind die Messungen auch parallel. □

Die Frage ist nun, ob nicht nur für parallele, sondern allgemein für kongenerische Tests, zumindest für die Spezialfälle τ-äquivalenter oder essentiell τ-äquivalenter Tests die Reliabilität numerisch angegeben werden kann. Tatsächlich ist dies möglich, wie Jöreskog (1971) gezeigt hat. Es sei (für den Moment) η die wahre Ausprägung des Merkmals, und die wahren Werte in tatsächlich applizierten Tests T_j seien τ_j, $j = 1, \ldots, \tau_n$. die sich als Transformationen von η ergeben. η ist eine zufällige Veränderliche insofern, als sie zwar eine charakteristische Konstante für eine gegebene Person ist, sich die verschiedenen Personen aber hinsichtlich ihres η-Wertes unterscheiden; mit der zufälligen Wahl einer Person erhält man eben zufällig einen bestimmten η-Wert (den man freilich nicht kennt). Jöreskog argumentiert zunächst, dass die Skala η der wahren Werte der Ausprägung des gemessenen Merkmals stets so angelegt werden kann, dass

$$\mathbb{E}(\eta) = 0, \quad \mathbb{V}(\eta) = 1. \tag{3.2.15}$$

Die Verteilung der θ-Werte in der Population soll also den Erwartungswert Null und die Varianz 1 haben; mit dieser Forderung wird nur der Nullpunkt und die Varianz der Skala festgelegt. Aus

$$X_j = \tau_j + \varepsilon_j = \nu_j + \lambda_j \eta + \varepsilon_j$$

(also $\tau_j = \nu_j + \lambda_j \eta$, d.h. τ_j ist eine lineare Transoformation von η) folgt nun

$$\mathbb{E}(X_j) = \nu_j + \lambda_j \mathbb{E}(\theta)$$

und damit

$$\mathbb{V}(X_j) = \mathbb{V}(\lambda_j \eta) = \lambda_j^2 \mathbb{V}(\eta),$$

d.h. wegen (3.2.15)

$$\mathbb{E}(X_j) = \nu_j, \quad \mathbb{V}(X_j) = \lambda_j^2. \tag{3.2.16}$$

Die Reliabilität des Tests T_j ist dann

$$\rho^2_{X_j \eta} = \frac{\lambda_j^2}{\mathbb{V}(X_j)} = \frac{\lambda_j^2}{\lambda_j^2 + \mathbb{V}(\varepsilon_j)}, \quad j = 1, \ldots, n. \tag{3.2.17}$$

Diese Definition entspricht der Definition (3.2.3).

Natürlich ergibt sich nun die Frage, wie die Parameter λ_j und $\mathbb{V}(\varepsilon_j)$ geschätzt werden können, um zu einer Schätzung des Reliabilitätskoeffizienten $\rho^2_{X_j\eta}$ zu gelangen. Jöreskog liefert eine Lösung auf der Basis der Faktorenanalyse. Dazu geht man davon aus, dass nach Voraussetzung *nur ein* Merkmal gemessen wird, nämlich das durch θ repräsentierte. Fasst man nun die Tests zu einem Vektor $\vec{x} = (X_1, \ldots, X_n)'$ zusammen, die λ_j zu einem Vektor $\vec{\lambda} = (\lambda_1, \ldots, \lambda_n)'$, definiert man $\vec{\nu} = (\nu_1, \ldots, \nu_n)$, so kann man die Gleichungen $X_j = \tau_j + \varepsilon_j$ und (3.2.6) zu einer Vektorgleichung[40]

$$\vec{x} = \vec{\nu} + \eta\vec{\lambda} + \vec{\varepsilon} \tag{3.2.18}$$

zusammenfassen. Die Kovarianzen zwischen den X_j sind

$$\mathbb{K}(X_j, X_k) = \mathbb{E}[(X_j - \mathbb{E}(X_j)(X_k - \mathbb{E}(X_k)] = \lambda_j\lambda_k + \mathbb{E}(\varepsilon_j\varepsilon_k),$$

denn $\mathbb{E}(X_j) = \nu_j$, $\mathbb{V}(X_j) = \lambda_j^2 + \sigma_j^2$, wegen $\mathbb{V}(\tau) = 1$ und $\sigma_j^2 = \mathbb{V}(\varepsilon_j)$. Die Kovarianzen der Fehler sind $\mathbb{K}(\varepsilon_j, \varepsilon_k) = \mathbb{E}(\varepsilon_j\varepsilon_k) - \mathbb{E}(\varepsilon_j)\mathbb{E}(\varepsilon_k) = 0$ für $j \neq k$. Die Matrix der Kovarianzen ist also

$$\vec{x}\vec{x}' = \begin{bmatrix} \lambda_1^2 & \lambda_1\lambda_2 & \cdots & \lambda_1\lambda_n \\ \lambda_2\lambda_1 & \lambda_2^2 & \cdots & \lambda_2\lambda_n \\ & & \vdots & \\ \lambda_n\lambda_1 & \lambda_n\lambda_2 & \cdots & \lambda_n^2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \tag{3.2.19}$$

Da nur ein Merkmal gemessen wird, sollte es nur einen Faktor geben. Die Faktorenanalyse (Maximum-Likelihood-Schätung) liefert dann die Schätzungen $\hat{\lambda}_j^2$ und die $\hat{\sigma}_j^2$ als Residuen, und die Schätzungen der Reliabilitäten sind dann

$$\hat{\rho}^2_{x_j\eta} = \frac{\hat{\lambda}_j^2}{\hat{\lambda}_j^2 + \hat{\sigma}_j^2}. \tag{3.2.20}$$

Die $\hat{\rho}^2_{x_j\eta}$ entsprechen den Kommunalitäten der Tests auf dem ersten Faktor. Man erinnere sich, dass die Faktorladung eines Tests die Korrelation des Tests mit dem entsprechenden Faktor ist. Der Faktor ist hier die Variable η. $\rho_{x_j\eta}$ kann also als Faktorladung des Tests X_j auf dem Faktor interpretiert werden, der die Variable η repräsentiert.

Test der Annahme der Kongenerität: Die Schätzungen (3.2.20) liefern also Reliabilitätsschätzungen ohne die einschränkende Parallelitätsannahme. Die Interpretation (3.2.19) der Kovarianzmatrix und die resultierende Schätzung der Reliabilität (3.2.20) gelten natürlich nur, wenn die Tests kongenerisch sind. Eine Faktorenanalyse der Matrix $\vec{x}\vec{x}'$ läßt sich aber stets rechnen, aber die Frage ist, ob (3.2.19) tatsächlich gilt. So ist z.B. nicht klar, ob die Annahme $\mathbb{K}(\varepsilon_j, \varepsilon_k) = 0$ tatsächlich gelten muß, ob sie also im konkreten Fall als erfüllt angesehen werden

[40] vergl. http://www.uwe-mortensen.de/Skripten: Vektoren und Matrizen

kann. Aber (3.2.19) liefert auch gleichzeitig einen Test der Hypothese der Kongenerität. Man prüft leicht nach, dass (3.2.19) die Gleichheit bestimmter Quotienten von Kovarianzen impliziert, nämlich :

$$\frac{\mathbb{K}(X_i, X_j)}{\mathbb{K}(X_j, X_k)} = \frac{\mathbb{K}(X_i, X_l)}{\mathbb{K}(X_j, X_k)}, \text{ für } i \neq k, i \neq l, j \neq k, j \neq l \tag{3.2.21}$$

Der mehrdimensionale Fall: Die Fokussierung auf eine 1-dimensionale Faktorlösung entspricht der üblichen Forderung der Homogenität eines auf Reliabilität geprüften Tests. In der Tat ist es aber möglich, die Reliabilität kongenerischer Tests auch für den heterogenen Fall, also den Fall eines mehrdimensionalen Merkmals, zu bestimmten (Lucke, 2005). Die Darstellung erfordert allerdings Kenntnisse der Matrixalgebra und überschreitet damit den Rahmen dieses Skriptums.

3.2.3 Arten der Reliabilität

Dies kann geschehen, indem man den gleichen Test zu zwei verschiedenen Zeitpunkten an den gleichen Personen administriert und dann die Korrelation für die Messwertpaare (X, X') berechnet; – dies ist die Re-Test-Reliabilität. Das Problem hierbei ist, dass (i) die gleichen Probanden für die beiden Testdurchführungen zur Verfügung stehen müssen und Gedächtniseffekte ausgeschlossen werden müssen. Eine andere Möglichkeit besteht darin, einen gegebenen Test in zwei Hälften aufzuteilen und die beiden Hälften als Paralleltests aufzufassen. Man spricht dann von der Split-Half-Reliabilität. Es zeigt sich aber, dass die Reliabilität von der Länge des Tests abhängt, so dass die Split-Half-Reliabilität eine Unterschätzung der Reliabilität liefert, die dann korrigiert werden muß. Die dritte Möglichkeit ist, zwei tatsächlich parallele Tests zu konstruieren und die Messungen in diesen Tests dann zur Abschätzung von $\rho_{xx'}$ zu benützen. Die Details dieser verschiedenen Vorgehensweisen werden in den folgenden Abschnitten entwickelt. Hier soll noch ein Ausdruck für $\mathbb{V}(\varepsilon)$ gegeben werden, der sich aus $\rho_{xx'}$ ableiten läßt. Damit ergibt sich auch ein Ausdruck für $\mathbb{V}(\tau)$; lass sich individuelle Fehler und wahre Werte nicht angeben, so können doch die Varianzen dieser Größen angegeben werden, womit dann die Möglichkeit gegeben ist, aufgrund von Messwerten Bereiche für die jeweiligen wahren Werte anzugeben.

3.2.4 Die Spearman-Brown-Formeln für $n > 1$ Komponenten

Die Betrachtungen bezogen sich bis jetzt nur auf einen "Score" X in einem Test, ohne darauf einzugehen, wie dieser Score gebildet wird. Im einfachsten Fall repräsentiert X eine Antwort auf ein einzelnes Item. im Allgemeinen besteht aber ein Test aus n Items, wobei n entweder fest vorgegeben ist oder erst während der Testdurchführung bestimmt wird. Reliabilität und Validität eines Tests hängen vom Wert von n ab. Es soll hier auf die wesentlichen Aspekte des Effekts der Zusammensetzung von Tests eingegangen werden.

Am Anfang steht der Fall, dass der Test aus zwei Komponenten besteht, die allerdings ihrerseits aus mehreren Komponenten (Items) bestehen können. Es resultiert eine Formel, die die Reliabilität eines Tests auf der Basis der Split-Half-Reliabilität bestimmt werden kann.

Es seien Y_1 und Y_2 zwei Testkomponenten mit den wahren Werten τ_1 und τ_2 sowie den Fehlern ε_1 und ε_2, und es gelte

$$X = Y_1 + Y_2. \tag{3.2.22}$$

Offenbar gilt für den wahren Wert und den Fehler von X

$$\begin{aligned} \tau &= \mathbb{E}(X) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2) = \tau_1 + \tau_2, && (3.2.23)\\ \varepsilon &= \varepsilon_1 + \varepsilon_2. && (3.2.24) \end{aligned}$$

Für die Varianzen erhält man

$$\begin{aligned} \mathbb{V}(\tau) &= \mathbb{V}(\tau_1) + \mathbb{V}(\tau_2) + 2\mathbb{K}(\tau_1, \tau_2),\\ &= \mathbb{V}(\tau_1) + \mathbb{V}(\tau_2) + 2\rho(\tau_1, \tau_2)\sqrt{\mathbb{V}(\tau_1)\mathbb{V}(\tau_2)} && (3.2.25)\\ \mathbb{V}(\varepsilon) &= \mathbb{V}(X) = \mathbb{V}(\varepsilon_1) + \mathbb{V}(\varepsilon_2). && (3.2.26) \end{aligned}$$

Für die Kovarianz der Fehler gilt nach Satz **??** $\mathbb{K}(\varepsilon_1, \varepsilon_2) = 0$.

Spearman-Brown-Formel für einen Test doppelter Länge: Die Tests T_1 und T_2 mit den Messungen Y_1 und Y_2 seien parallel. Dann gilt $\tau_1 = \tau_2 = \tau$, $\mathbb{V}(Y_1) = \mathbb{V}(Y_2) = \mathbb{V}(Y)$, und $\mathbb{V}(\varepsilon_1) = \mathbb{V}(\varepsilon_2)$ und die vorangehenden Gleichungen vereinfachen sich zu

$$\begin{aligned} \mathbb{V}(X) &= 2\mathbb{V}(Y_1)(1 + \rho_{y_1y_2}) = 2\mathbb{V}(Y)(1 + \rho_{yy'}) && (3.2.27)\\ \mathbb{V}(\tau) &= 4\mathbb{V}(\tau_1) && (3.2.28)\\ \mathbb{V}(\varepsilon) &= 2\mathbb{V}(\varepsilon_1). && (3.2.29) \end{aligned}$$

Die Gleichungen folgen einfach durch Nachrechnen. Sicherlich ist

$$\mathbb{V}(X) = \mathbb{V}(Y_1)+\mathbb{V}(Y_2)+2\rho(Y_1, Y_2)\sqrt{\mathbb{V}(Y_1)\mathbb{V}(Y_2)} = 2\mathbb{V}(Y)+2\rho(Y_1, Y_2)\mathbb{V}(Y),$$

da ja nach Voraussetzung $\mathbb{V}(Y_1) = \mathbb{V}(Y_2) = \mathbb{V}(Y)$. Dann folgt aber sofort

$$\mathbb{V}(X) = 2\mathbb{V}(Y)(1 + \rho(Y_1, Y_2)).$$

Die Gleichung (3.2.28) folgt aus (A.4.2) weil für parallele Tests $\mathbb{V}(\tau_1) = \mathbb{V}(\tau_2)$ und $\rho(\tau_1, \tau_2) = 1$. Die Gleichung (3.2.29) folgt aus der Voraussetzung $\mathbb{V}(\varepsilon_1) = \mathbb{V}(\varepsilon_2)$ und $\mathbb{K}(\varepsilon_1, \varepsilon_2) = 0$. □

Die Reliabilität des zusammengesetzten Tests läßt sich nun über die Reliablität der Tests T_1 bzw. T_2 ausdrücken. Es ist ja

$$\rho_{xx'} = \frac{\mathbb{V}(\tau_1 + \tau_2)}{\mathbb{V}(X)} = \frac{4\mathbb{V}(\tau_1)}{2\mathbb{V}(Y)(1 + \rho_{yy'})},$$

wegen (3.2.28) und (3.2.27). Weiter ist $\rho_{yy'}$ durch $\rho_{yy'} = \mathbb{V}(\tau_1)/\mathbb{V}(Y)$ gegeben, so dass

$$\rho_{xx'} = \frac{2\rho_{yy'}}{1+\rho_{yy'}} \tag{3.2.30}$$

folgt. Dies ist die *Spearman-Brown-Formel für einen Test doppelter Länge*. Die Reliabilität des zusamemngesetzten Tests wird hier anhand der Reliabilität der Komponenten, die als parallele Tests vorausgesetzt werden, bestimmt. Sind die Komponenten gerade die Split-Half-Hälften, so kann man anhand von (3.2.30) die Reliabilität des ursprünglichen Tests bestimmen. Eine Illustration der Beziehung (3.2.30) findet man in Abb. 19, (A), Seite 159.

Lord & Novick (1968), p. 84, geben eine weitere Herleitung für die Reliabilität eines zusammengesetzten Tests. Gegeben seien zwei parallele Tests mit den Scores Y und Y'; die Reliabilität eines Tests ist dann durch die Korrelation $\rho_{yy'}$ gegeben. Gesucht ist die Reliabilität eines Tests doppelter Länge. Dazu seien vier parallele Tests mit Messwerten $Y_1, \ldots, Y_4$ gegeben, die zu zwei wiederum parallelen Tests $X = Y_1 + Y_2$ und $X' = Y_3 + Y_4$ zusammengefasst werden. Die gesuchte Reliabilität ist die Reliabilität von X, wie sie durch die Korrelation $\rho_{xx'}$ gegeben ist.

Es sei

$$\rho_{yy'} = \rho(Y_j, Y_k) \text{ für alle } j, k \tag{3.2.31}$$

wegen der Parallelität der Y_j, $j = 1, \ldots, 4$. Es gilt nun wieder (3.2.30), also

$$\rho_{xx'} = \frac{2\rho_{yy'}}{1+\rho_{yy'}}.$$

Darüber hinaus läßt sich zeigen, dass

$$\rho_{xx'} \geq \rho_{yy'}. \tag{3.2.32}$$

gilt.

Beweis: Man hat

$$\rho_{xx'} = \frac{\mathbb{K}(Y_1 + Y_2, Y_3 + Y_4)}{\sigma_{12}\sigma_{34}}, \tag{3.2.33}$$

wobei $\sigma_{12} = \sqrt{\mathbb{V}(Y_1 + Y_2)}$ und $\sigma_{34} = \sqrt{\mathbb{V}(Y_3 + Y_4)}$ ist. Es ist

$$\mathbb{V}(Y_1+Y_2) = \mathbb{V}(Y_1)+\mathbb{V}(Y_2)+2\rho(Y_1,Y_2)\sqrt{\mathbb{V}(Y_1)\mathbb{V}(Y_2)} = 2\mathbb{V}(Y)+2\rho(Y_1,Y_2)\mathbb{V}(Y),$$

d.h.

$$\mathbb{V}(Y_1 + Y_2) = 2\mathbb{V}(Y)(1 + \rho(Y_1, Y_2)),$$

da ja $\mathbb{V}(Y_1) = \mathbb{V}(Y_2) = \mathbb{V}(Y)$. Analog dazu findet man

$$\sqrt{\mathbb{V}(Y_3) + \mathbb{V}(Y_4)} = 2\mathbb{V}(Y') + 2\rho(Y_3, Y_4)\mathbb{V}(Y') = 2\mathbb{V}(Y')(1 + \rho(Y, Y')),$$

wegen $\mathbb{V}(Y_3) = \mathbb{V}(Y_4) = \mathbb{V}(Y')$, und natürlich $\mathbb{V}(Y) = \mathbb{V}(Y')$, $\rho(Y_1, Y_2) = \rho(Y_3, Y_4) = \rho(Y, Y')$. Dann folgt

$$\sigma_{12}\sigma_{34} = \sqrt{4\mathbb{V}(Y)\mathbb{V}(Y')(1 + \rho(Y, Y'))^2} = 2\mathbb{V}(Y)(1 + \rho(Y, Y')).$$

Für die Kovarianz erhält man

$$\begin{aligned}\mathbb{K}(Y_1+Y_2,Y_3+Y_4) &= \mathbb{E}[(Y_1+Y_2)(Y_3+Y_4)]-\mathbb{E}(Y_1+Y_2)\mathbb{E}(Y_3+Y_4)\\ &= \mathbb{E}(Y_1Y_3)+\mathbb{E}(Y_1Y_4)+\mathbb{E}(Y_2Y_3)+\mathbb{E}(Y_2Y_4)\\ &\quad -\mathbb{E}(Y_1)\mathbb{E}(Y_3)-\mathbb{E}(Y_1)\mathbb{E}(Y_4)-\mathbb{E}(Y_2)\mathbb{E}(Y_3)-\mathbb{E}(Y_2Y_4)\\ &= 4(\mathbb{E}(YY')-\mathbb{E}(Y)\mathbb{E}(Y'))\\ &= 4\rho_{yy'}\sigma_y\sigma_{y'}=4\rho_{yy'}\mathbb{V}(Y) \qquad (3.2.34)\end{aligned}$$

Dann folgt für $\rho_{xx'}$

$$\rho_{xx'}=\frac{4\rho_{yy'}\mathbb{V}(Y)}{2\mathbb{V}(Y)(1+\rho(Y,Y'))}=\frac{2\rho_{yy'}}{1+\rho_{yy'}}.$$

□

Die Behauptung (3.2.32) sieht man wie folgt ein. Die beiden Komponenten haben jeweils die Reliabilität $\rho_{yy'}$, der zusammengesetzte Test hat die Reliabilität $\rho_{xx'}$. Es ist stets $0\le\rho_{yy'}\le 1$. Angenommen, es gelte

$$\frac{2\rho_{yy'}}{1+\rho_{yy'}}\le\rho_{yy'}.$$

Dann folgt

$$\frac{2}{1+\rho_{yy'}}\le 1,\ \text{d.h. } 2\le 1+\rho_{yy'},$$

oder $1\le\rho_{yy'}$, im Widerspruch zu $\rho_{yy'}\le 1$. Damit folgt dann (3.2.32). □

Die allgemeine Spearman-Brown-Formel für n parallele Tests: Ein Test habe n Komponenten mit den Scores $Y_1,\ldots,Y_n$ und dem Gesamtmesswert X, dem wahren Wert τ und dem Fehler ε, wobei gilt

$$X=\sum_{i=1}^n Y_i,\ \tau=\sum_{i=1}^n\tau_i,\ \varepsilon=\sum_{i=1}^n\varepsilon_i \qquad (3.2.35)$$

Für die Erwartungswerte (über die Personen aus der Population $\mathcal{P}$) ergeben sich die Ausdrücke

$$\mathbb{E}(X)=\sum_{i=1}^n\mathbb{E}(Y_i),\ \mathbb{E}(\tau)=\sum_{i=1}^n\mathbb{E}(\tau_i),\ \mathbb{E}(\varepsilon)=\sum_{i=1}^n\mathbb{E}(\varepsilon_i)=0, \qquad (3.2.36)$$

wegen $E(\varepsilon_i)=0$ für alle i. Für die Varianz von X und τ erhält man[41]

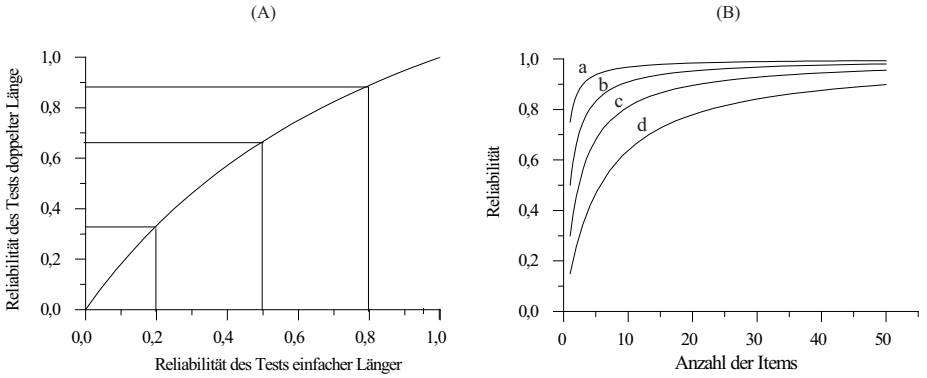
$$\mathbb{V}(X) = \sum_{i=1}^n\mathbb{V}(Y_i)+2\sum_{i>j}\mathbb{K}(Y_i,Y_j) \qquad (3.2.37)$$

$$\mathbb{V}(\tau) = \sum_{i=1}^n\mathbb{V}(\tau_i)+2\sum_{i>j}\mathbb{K}(\tau_i,\tau_j) \qquad (3.2.38)$$

$$\mathbb{V}(\varepsilon) = \sum_{i=1}^n\mathbb{V}(\varepsilon_i) \qquad (3.2.39)$$

[41] Die Summe $2\sum_{i>j}\mathbb{K}(Y_i,Y_j)$ ergibt sich aus der Summe $\sum_i\sum_j\mathbb{K}(Y_i,Y_j)$, wenn man bedenkt, dass ja $\mathbb{K}(Y_i,Y_j)=\mathbb{K}(Y_j,Y_i)$ gilt.

Abbildung 19: (A) Reliabilität bei Verdoppelung der Testlänge. Für $\rho_{yy'} = .2$ ist $\rho_{xx'} = .333$, für $\rho_{yy'} = .5$ ergibt sich $\rho_{xx'} = .666$, für $\rho_{yy'} = .8$ ist $\rho_{xx'} = .888$. (B) Reliabilität bei n-facher Testverlängerung. (a) $\rho_{yy'} = .75$, (b) $\rho_{yy'} = .5$, (c) $\rho_{yy'} = .30$, (d) $\rho_{yy'} = .15$.



wobei in (3.2.39) berücksichtigt wurde, dass die Fehler unkorreliert sind (Satz **??**).

Die $Y_1, \ldots, Y_n$ seien Scores von parallelen Tests, d.h. τ-äquivalenten Tests mit gleicher Fehlervarianz. Damit gelten

$$\tau_1 = \cdots = \tau_n, \;\; \mathbb{V}(\varepsilon_1) = \cdots = \mathbb{V}(\varepsilon_n)$$

und aus (3.2.37), (3.2.38) und (3.2.39) erhält man die Ausdrücke

$$\mathbb{V}(X) = n\mathbb{V}(Y)(1 + (n-1)\rho_{yy'}) \tag{3.2.40}$$

$$\mathbb{V}(\tau) = n^2\mathbb{V}(\tau_1) \tag{3.2.41}$$

$$\mathbb{V}(\varepsilon) = n\mathbb{V}(\varepsilon_1). \tag{3.2.42}$$

Während die Varianz $\mathbb{V}(\varepsilon)$ proportional zu n wächst, erhöht sich die Varianz der wahren Werte τ proportional zu n^2. Da aber $\mathbb{V}(\tau)$ die Varianz der τ_a-Werte in der Population $\mathcal{P}$ ist, bedeutet dies Ergebnis, dass die wahren Unterschiede zwischen den Personen offenbar um so besser erfasst werden, je größer die Anzahl n der Testkomponenten ist.

Aus (3.2.40) und (3.2.41) ergibt sich sofort die Reliabilität $\rho_{xx'}$ des zusammengesetzten Tests:

$$\rho_{xx'} = \frac{\mathbb{V}(\tau)}{\mathbb{V}(X)} = \frac{n^2\mathbb{V}(\tau_1)}{n\mathbb{V}(Y)(1 + (n-1)\rho_{yy'})},$$

und da $\mathbb{V}(\tau_1)/\mathbb{V}(Y) = \rho_{yy'}$ erhält man hieraus sofort die allgemeine Spearman-Brown-Formel

$$\rho_{xx'}^{(n)} = \alpha = \frac{n\rho_{yy'}}{1 + (n-1)\rho_{yy'}}, \tag{3.2.43}$$

wobei etwa $Y = Y_1$, und wo $\rho_{xx'}^{(n)}$ statt einfach $\rho_{xx'}$ geschrieben wurde, um die Beziehung zum Wert n hervorzuheben. Es gilt

$$\lim_{n\to\infty} \rho_{xx'}^{(n)} = \frac{\rho_{yy'}}{\rho_{yy'}} = 1. \tag{3.2.44}$$

Dieser Grenzwert ergibt sich aus (3.2.43), indem man dort den Zähler und den Nenner durch n dividiert:

$$\rho_{xx'}^{(n)} = \frac{\rho_{yy'}}{1/n - (1-1/n)\rho_{yy'}},$$

woraus (3.2.44) sofort folgt.

> **Interpretation:** (3.2.44) bedeutet, dass die Reliabilität eines Tests mit der Anzahl der (parallelen) Komponenten wächst und sogar beliebig groß werden, also gegen 1 streben kann, unabhängig von der Reliabilität $\rho_{yy'}$ der Komponenten.

Für $n = 2$ reduziert sich (3.2.43) auf den Fall der Reliabilität eines Tests mit doppelter Länge, also (3.2.30), Seite 157.

3.2.5 Interne Konsistenz und Cronbachs α

Gesucht ist Abschätzung für den Mindestwert der Reliabilität $\rho_{xx'}$ für den Fall von n Komponenten (üblicherweise Items). Der folgende Satz wurde zuerst von Cronbach (1951) vorgestellt.

Satz 3.2.1 *Es sei*

$$X = Y_1 + \cdots + Y_n, \quad Y_j = \tau_i + \varepsilon_j, \ \forall j \tag{3.2.45}$$

Dan gilt

$$\rho_{xx'} \geq \frac{n}{n-1}\left(1 - \frac{\sum_j \mathbb{V}(Y_j)}{\mathbb{V}(X)}\right). \tag{3.2.46}$$

Die rechte Seite heißt auch *Cronbachs* α, d.h.

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_j \mathbb{V}(Y_j)}{\mathbb{V}(X)}\right). \tag{3.2.47}$$

Ein Beweis für (3.2.46) (nicht Cronbachs (1971) Herleitung, sondern basierend auf Lord & Novick (1968), p. 88-89) wird im Anhang, Abschnitt A.4, Seite 287, gegeben, da die Details für die Interpretation von Cronbachs α nicht von Bedeutung sind.

Um die Interpretation von α zu erleichtern ist es nützlich, (3.2.46) in etwas anderer Form anzuschreiben. Dazu wird angemerkt, dass aus $X = Y_1 + \cdots + Y_n$ folgt, dass

$$\mathbb{V}(X) = \sum_j \mathbb{V}(Y_j) + \sum_{i \neq j} \sigma_{ij} = \sum_j \mathbb{V}(Y_j) + \sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j, \tag{3.2.48}$$

wobei ρ_{ij} die Korrelation zwischen den Scores Y_i und Y_j ist. Dann folgt

$$\sum_j \mathbb{V}(Y_j) = \mathbb{V}(X) - \sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j, \tag{3.2.49}$$

und man kann α in den Formen

$$\alpha = \frac{n}{n-1} \frac{\sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j}{\mathbb{V}(X)} \tag{3.2.50}$$

$$= \frac{n}{n-1} \frac{\sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j}{\mathbb{V}(X)} = \frac{n}{n-1} \left(\frac{\sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j}{\sum_j \sigma_j^2 + \sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j} \right) \tag{3.2.51}$$

anschreiben. Es sind verschiedene Fälle zu betrachten:

1. **Unabhängige Komponenten (perfekte Heterogenität):** Sind die Y_j unabhängig voneinander, d.h. sind die $\rho_{ij} = 0$ für alle $i \neq j$, so ist $\alpha = 0$, und die Ungleichung (3.2.46) ist trivial, denn $\rho_{xx'} \geq 0$ gilt in jedem Fall. D.h. ein *perfekt heterogener Test* hat eine Reliabilität gleich Null, für alle Werte von n (Anzahl der Items oder Komponenten):

   $$\text{Rel}(X) = 0, \quad \rho_{ij} = 0, \quad \text{für alle } i \neq j. \tag{3.2.52}$$

   Verallgemeinernd kann man sagen: je weniger die Komponenten (z.B. die Items) untereinander korrelieren, d.h. je heterogener der Gesamttest (mit dem Score X) ist, desto kleiner wird der Wert von α sein. Die Umkehrung gilt allerdings nicht: von einem kleinen Wert von α kann nicht zwingend auf niedrige Korrelationen und damit auf geringe interne Konsistenz geschlossen werden, wie noch elaboriert werden wird.

2. **Perfekte Homogenität:** In (3.2.51) kann man den Zähler und den Nenner durch die Summe $\sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j$ teilen; für $\rho_{ij} = 1$ für alle $i \neq j$ folgt nun

   $$\alpha = \frac{n}{n-1} \left(\frac{1}{1 + \sum_j \sigma_j^2 / \sum_{i \neq j} \sigma_i\sigma_j} \right). \tag{3.2.53}$$

   Der genaue Wert von α hängt nun von den Werten von σ_j^2, $j = 1, \ldots, n$ ab; ohne weitere Spezifikation der σ_j^2 läßt sich nicht viel sagen. Interessant ist aber der Spezialfall $\sigma_j = \sigma^2$ für alle j, also der Fall gleich großer Varianzen. Dann gilt ja

   $$\sum_j \sigma_j^2 = n\sigma^2, \quad \sum_{i \neq j} \sigma_i\sigma_j = n(n-1)\sigma^2,$$

und

$$\frac{\sum_j \sigma_j^2}{\sum_{i\neq j} \sigma_i\sigma_j} = \frac{n\sigma^2}{n(n-1)\sigma^2} = \frac{1}{n-1}.$$

Damit erhält man für α

$$\alpha = \frac{n}{n-1}\left(\frac{1}{1+1/(n-1)}\right) = \frac{n}{n-1}\frac{n-1}{n} = 1.$$

Im Falle perfekt korrelierender, also perfekt homogener Items mit gleichen Varianzen erhält man einen Test mit der Reliabilität 1.

3. **Der allgemeine Fall:** Nun gelte $\rho_{ij} \neq 0$ für einige, wenn nicht alle $i \neq j$. Es zeigt sich, dass auch bei niedrigen Korrelationen der Wert von α bemerkenswert hoch sein kann. Dies gilt insbesondere dann, wenn der Wert von n, der Anzahl von Komponenten bzw. Items, nicht allzu klein ist (man erinnere sich: ein möglichst großer Wert erhöht die Reliabilität eines Tests!).

   Um dies zu sehen, kann man den Wert von α in Abhängigkeit von n, der Anzahl der Komponenten (Items) betrachten. Es gibt $n(n-1)$ Paare $\rho_{ij}\sigma_i\sigma_j$, $i \neq j$. Der durchschnittliche Wert S der $\rho_{ij}\sigma_i\sigma_j$ und die durchschnittliche Varianz sind dann

$$S = \frac{1}{n(n-1)}\sum_{i\neq j}\rho_{ij}\sigma_i\sigma_j \tag{3.2.54}$$

$$\bar{\sigma}^2 = \frac{1}{n}\sum_j \sigma_j^2 \tag{3.2.55}$$

   Es sei

$$q = \frac{\bar{\sigma}^2}{S}. \tag{3.2.56}$$

   Die Gleichung (3.2.51) für α kann dann in der Form

$$\alpha = \frac{1}{(q-1)/n+1}, \quad 1 \leq q \leq \infty \tag{3.2.57}$$

   geschrieben werden (s. Beweis weiter unten). Der Wert von Cronbachs α hängt also ab (i) von n, der Anzahl der Komponenten bzw. Items, und (ii) vom Quotienten der durschnittlichen Varianz der Items und der durchschnittlichen Kovarianz zwischen den Items. Dieser Sachverhalt wird in den Gleichungen (3.2.50) und (3.2.51) vermutlich noch nicht sofort deutlich. α wird offenbar maximal (strebt gegen 1), wenn dieser Quotient gegen 1 strebt, wenn also die durchschnittliche Varianz und die durchschnittliche Kovarianz gleich groß sind, – es wird gezeigt werden, dass die durchschnittliche Kovarianz nie größer als die durchschnittliche Varianz sein kann.

Beweis: Aus den Definitionen (3.2.54) von S und (3.2.55) von $\bar{\sigma}^2$ folgt

$$n\bar{\sigma}^2 = \frac{1}{n}\sum_j \sigma_j^2, \quad n(n-1)S = \frac{1}{n(n-1)}\sum_{i\neq j}\rho_{ij}\sigma_i\sigma_j,$$

und aus (3.2.51) folgt

$$\begin{aligned}\alpha &= \frac{n}{n-1}\left(\frac{n(n-1)S}{(n\bar{\sigma}^2 - n(n-1)S)}\right)\\ &= \frac{nS}{\bar{\sigma}^2 + (n-1)S)} = \frac{1}{\bar{\sigma}^2/(nS) + (n-1)/n)}\\ &= \frac{1}{(\bar{\sigma}^2/S - 1)/n + 1} = \frac{1}{(q-1)/n + 1}.\end{aligned}$$

Es bleibt zu zeigen, dass $1 \leq q \leq \infty$. Die Ungleichung $q \leq \infty$ wird zuerst gezeigt. Es ist $q = \bar{\sigma}^2/S$, und $q \to \infty$ für $S \to 0$.

Eine *hinreichende*, wenn auch nicht *notwendige* Bedingung für $S \to 0$ ist $\rho_{ij} \to 0$ für alle $i \neq j$, wenn also die Items unkorreliert sind. Andere Bedingungen für $S \to 0$ werden weiter unten diskutiert.

Nun muß gezeigt werden, dass $q \geq 1$. Diese Ungleichung folgt sofort, wenn man bedenkt, dass $\alpha \leq \rho_{xx'} \leq 1$ für alle n sein muß. Man hat also wegen (3.2.57) $1/((q-1)/n + 1) \leq 1$, d.h. $1 \leq (q-1)/n + 1$ bzw. $0 \leq (q-1)/n$, so dass $q \geq 1$ und damit $\bar{\sigma}_n^2 \geq S$ für alle n.

Der Schluß von $\alpha \leq 1$ auf $q \geq 1$ und damit auf $\bar{\sigma}^2 \geq S$ ist zwar korrekt, macht aber nicht explizit, wie sich die Ungleichung $\bar{\sigma}^2 \geq S$ aus den Definitionen von $\bar{\sigma}^2$ und S ergibt.

Generell gilt für irgend zwei reelle Zahlen a und b die Ungleichung

$$0 \leq (a-b)^2 = a^2 + b^2 - 2ab,$$

also $2ab \leq a^2 + b^2$, d.h. $2\sigma_i\sigma_j \leq \sigma_i^2 + \sigma_j^2$. Dann gilt sicherlich

$$2\sum_{i,j}\sigma_i\sigma_j \leq \sum_{i,j}\sigma_i^2 + \sum_{i,j}\sigma_j^2 = 2n\sum_j \sigma_j^2.$$

Nun ist $\sum_{i,j}\sigma_i\sigma_j = \sum_{i\neq j}\sigma_i\sigma_j + \sum_j \sigma_j^2$, so dass

$$\sum_{i\neq j}\sigma_i\sigma_j \leq n\sum_j \sigma_j^2 - \sum_j \sigma_j^2 = (n-1)\sum_j \sigma_j^2 \tag{3.2.58}$$

folgt. Wegen $\sum_{i\neq j}\sigma_i\sigma_j = n(n-1)S$ und $\sum_j \sigma_j^2 = n\bar{\sigma}_n^2$ ist aber (3.2.58) äquivalent mit der Ungleichung

$$n(n-1)S \leq n(n-1)\bar{\sigma}_n^2,$$

so dass $S \leq \bar{\sigma}_n^2$ und damit $\bar{\sigma}_n^2/S = q \geq 1$ folgt. Dass diese Ungleichung $\alpha \leq 1$ impliziert, sieht man sofort aus (3.2.57). □

S und $\bar{\sigma}^2$ sind als Durchschnitte (arithmetische Mittel) für eine gegebene Stichprobe von n Komponenten definiert worden. Genau genommen müßte

also S_n und $\bar{\sigma}_n^2$ statt einfach S und $\bar{\sigma}^2$ geschrieben werden, denn diese Mittelwerte werden sich für verschiedene Werte von n unterscheiden, zumal es sich bei den n Items jeweils um Stichproben aus einem Universum von Items handelt. Für einen fixen Wert von n muß darauf nicht besonders abgehoben werden, es kommt nur darauf an, zu sehen, dass α außer vom Wert von n vom Verhältnis $q = \bar{\sigma}^2/S$ der Mittelwerte abhängt, und das ist für eine Diskussion der internen Konsistenz schon hinreichend. Will man aber speziell die Abhängigkeit von α von n untersuchen, muß man fragen, wie sich die Variation der Stichprobenmittelwerte S_n und $\bar{\sigma}_n^2$ auswirkt. Es sei nun

$$S = \mathbb{E}(\mathbb{K}(Y_i, Y_j)), \quad \bar{\sigma}^2 = \mathbb{E}(\mathbb{V}(Y_i)). \tag{3.2.59}$$

S und $\bar{\sigma}^2$ sind damit Mittelwerte über alle möglichen Kovarianzen bzw. Varianzen in der Population. Für eine gegebene Stichprobe von n Items existieren nun Größen ΔS_n und $\Delta\bar{\sigma}_n^2$ derart, dass für die Mittelwerte S_n und $\bar{\sigma}_n^2$ für diese Stichprobe

$$S_n = S + \Delta S_n, \quad \bar{\sigma}_n^2 = \bar{\sigma}^2 + \Delta\bar{\sigma}_n^2. \tag{3.2.60}$$

geschrieben werden kann. ΔS_n und $\Delta\bar{\sigma}_n^2$ sind einfach die Differenzen zwischen S und S_n bzw. zwischen $\bar{\sigma}^2$ und $\bar{\sigma}_n^2$, und es sind zufällige Variablen insofern, als die Auswahl der Items als zufällige Auswahl aus dem Universum der möglichen Items betrachtet werden kann. Nun weiß man aus der Statistik, dass unter sehr allgemeinen Bedingungen ein Mittelwert gegen den Erwartungswert strebt, wenn der Stichprobenumfang größer wird[42], also

$$\lim_{n\to\infty} (\bar{x}_n - \mathbb{E}(X)) = 0.$$

Das heißt aber, dass die ΔS_n und $\Delta\bar{\sigma}_n^2$ gegen Null gehen für wachsendes n. Dann folgt aus (3.2.57)

$$\lim_{n\to\infty} \alpha = 1. \tag{3.2.61}$$

Denn man kann nun

$$\alpha_n = \frac{1}{\frac{1}{n}(q_n - 1) + 1}, \quad q_n = \frac{1}{n}\left(\frac{S + \Delta S_n}{\bar{\sigma}^2 + \Delta\bar{\sigma}_n^2} - 1\right)$$

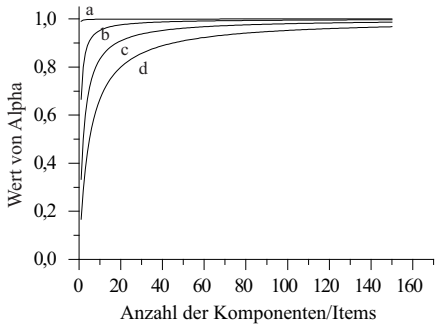
schreiben, wobei α_n statt α geschrieben wurde, um die Abhängigkeit von n auszudrücken. Nun gilt sicherlich

$$\lim_{n\to\infty} \frac{1}{n}\left(\frac{S + \Delta S_n}{\bar{\sigma}^2 + \Delta\bar{\sigma}_n^2} - 1\right) = 0,$$

denn $\lim_{n\to\infty} 1/n = 0$, und da der Ausdruck in den beiden Klammern

[42] Bekanntlich ist die Varianz eines arithmetischen Mittels durch $\mathbb{V}(\bar{x}_n) = \mathbb{V}(X)/n$, $\mathbb{V}(X)$ die Varianz der zufälligen Veränderlichen X. Für $\mathbb{V}(X) < \infty$ folgt dann $\lim_{n\to\infty} \mathbb{V}(\bar{x}_n) = 0$, da $1/n$ gegen Null strebt. Das heißt aber, dass $\bar{x}_n$ für wachsenden Stichprobenumfang immer dichter an $\mathbb{E}(X)$ liegt.

Abbildung 20: Abhängigkeit von α von der Anzahl n der Komponenten bzw. Items. a: $q = 1.01$, b: $q = 1.5$, c: $q = 3.0$, c: $q = 6$.



nicht mit n wächst, folgt die Konvergenz gegen 1. Dass ΔS_n und $\Delta\bar{\sigma}_n^2$ ebenfalls gegen Null streben, ist oben bereits diskutiert worden. Also kann $\lim_{n\to\infty} \alpha_n = 1$ gefolgert werden. Dies bedeutet, dass mit größer werdendem Wert von n der Wert von α zumindest im Mittel größer wird. Wollte man eine Kurve α_n versus n zeichnen, so würde sie für kleinere Werte von n wegen der ΔS_n und $\Delta\bar{\sigma}_n^2$ ein wenig verruckelt aussehen, – da diese Größen aber für größer werdendes n verschwinden, wird die Kurve mit wachsendem n immer glatter gegen 1 streben. Abbildung 20 zeigt solche Kurven für verschiedene Werte von q, wobei aber die zufälligen Größen ΔS_n und $\Delta\bar{\sigma}_n^2$ vernachlässigt wurden, da es der Gesamteindruck der Kurvenverläufe ist, der vermittelt werden soll.

Der Wert von q bzw. q_n bestimmt *die Rate*, mit der α mit wachsendem n gegen 1 strebt. Je kleiner der Wert von S, desto größer ist der Wert von q (gegeben $\bar{\sigma}^2$), und desto langsamer strebt α gegen 1. Wichtig für die Interpretation eines α-Wertes ist allerdings, dass auch für relativ große Werte von q (also relativ kleine Werte von S) ein relativ großer Wert für α, etwa $\alpha > .8$, resultiert. Der Rückschluß von solchen α-Werten auf eine gute interne Konsistenz und damit auf eine gute Homogenität der Items ist also keineswegs zwingend, denn kleine Werte von S *können*, müssen aber nicht kleine Korrelationen zwischen den Items und damit Heterogenität des Tests bedeuten; dieser Punkt wird weiter unten bei der Diskussion von α als Maß für die interne Konsistenz noch weiter elaboriert. In Abbildung 20 ist im Fall (a) $q = 1.01$, und α ist für alle Werte von n dicht an 1. Im Fall (c) ist $q = 6$, d.h. S hat nur 1/6 des Wertes der durchschnittlichen Varianz. Gleichwohl liegt α schon bei $n = 40$ über dem Wert .8.

4. **Parallele Komponenten:** Für den Fall, dass alle Komponenten parallel sind, gilt $\sigma_1^2 = \cdots = \sigma_n^2 = \sigma_y^2$ und $\mathbb{V}(X) = n\sigma_y^2[1 + (n-1)\rho_{yy'}$, $\rho_{yy'}$ die

Reliabilität der einzelnen Komponenten (vergl. (3.2.40)). Dann ist

$$\alpha = \frac{n}{n-1}\left(1 - \frac{n\sigma_y^2}{n\sigma_y^2[1+(n-1)\rho_{yy'}]}\right) = \frac{n\rho_{yy'}}{1+(n-1)\rho_{yy'}} \qquad (3.2.62)$$

Die rechte Seite ist aber gerade die Reliabilität für einen Test mit n-facher Länge, vergl. (3.2.43).

Cronbachs α als Maß für die interne Konsistenz: Das Resultat $\alpha = 0$ für den Fall $\rho_{ij} = 0$ für alle $i \neq j$ läßt zunächst die Interpretation von α als Maß für die interne Konsistenz sinnvoll *erscheinen*, denn je kleiner die Korrelationen zwischen den Items sind, d.h. je heterogener der Test ist, desto kleiner wird der Wert von α. Heterogenität bedeutet, das verschiedene Merkmale gemessen werden, was sich faktorenanalytisch durch eine Anzahl $r > 1$ latenter Dimensionen ausdrückt.

Aber kleine Korrelationen ρ_{ij} oder gar $\rho_{ij} = 0$ für alle $i \neq j$ sind nur eine *hinreichende*, leider aber keine *notwendige* Bedingung für $\alpha = 0$ bzw. für einen kleinen α-Wert. Es ist ja

$$S = \sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j,$$

Der Fall $S \approx 0$ kann auch etwa dann eintreten, wenn die ρ_{ij} alternierende Vorzeichen haben, wenn also im Extremfall $\rho_{ij} = 1$ für einige Items aus einer Teilmenge M_1 und $\rho_{ij} = -1$ für andere Items aus einer Teilmenge M_2 und für geeignete Werte der σ_i, σ_j

$$\sum_{i,j \in M_1, i \neq j} \rho_{ij}\sigma_i\sigma_j \approx -\sum_{k,l \in M_2, k \neq l} \rho_{kl}\sigma_k\sigma_l$$

gilt. Die Existenz von Teilmengen mit entweder hoch positiv oder hoch negativ korrelierenden Items kann eine durchschnittliche Kovarianz nahe oder gleich Null und damit einen kleinen α-Wert erzeugen. Man hat den Fall eines eigentlich sehr homogenen Tests – die negativen Korrelationen können u. U. durch Umformulierung der Items sogar in positive verwandelt werden –, der aber gleichwohl einen Wert $\alpha \approx 0$ implizieren kann. In einem Leistungstest wird man diesen Fall allerdings kaum beobachten, da der Fall, dass eine Aufgabe gelöst wird genau dann, wenn eine bestimmte andere nicht gelöst wird, sicherlich sehr selten sein wird. Bei Fragen zu politischen oder sonstigen Einstellungen ist dieser Fall aber vorstellbar, weil die Beantwortung u. A. von der Formulierung der Items abhängt.

Umgekehrt läßt aber ein großer Wert von α noch nicht auf Homogenität schließen, wie oben wegen der allgemeinen Konvergenz $\alpha \to 1$ mit wachsendem n gezeigt wurde. Denn α wächst mit dem Wert von n schon dann, wenn nicht alle Korrelationen zwischen den Items gleich Null sind – die Korrelationen müssen noch nicht einmal groß sein. Da sie streng gleich Null in der Praxis nie sein werden kann man bei hinreichend großem Wert von n immer einen relativ hohen

α-Wert erwarten. Gleichzeitig ist ein eher größerer Wert von n wünschenswert, da er eine höhere Reliabilität des Tests impliziert.

Insgesamt erscheint Cronbachs α, wenn es für sich betrachtet wird, kein gutes Maß für die interne Konsistenz zu sein. Eine Inspektion der Korrelationen zwischen den Komponenten bzw. Items wird in jedem Fall fruchtbarer sein als die alleinige Fokussierung auf den Wert von α.

Es läßt sich zeigen, dass eine hinreichende und notwendige Bedingung für die Beziehung $\rho_{xx'} = \alpha$, die Beziehung

$$\tau_i = a_{ij} = \tau_j, \text{ für alle } i, j \tag{3.2.63}$$

ist (Lord & Novick (1968), p. 90).

Dichotome Items und die Kuder-Richardson Formeln: Es wurde bereits den Spezialfall von α diskutiert, der entsteht, wenn die Items dichotom beantwortet werden: $Y_i = 1$ mit Wahrscheinlichkeit p_i und $Y_i = 0$ mit der Wahrscheinlichkeit $1 - p_i$. Man erhält dann die

Kuder-Richardson- Formel 20

$$\alpha = \alpha_{20} = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} p_i(1-p_i)}{\mathbb{V}(X)}\right]. \tag{3.2.64}$$

Gilt darüber hinaus $p_i = p$ für alle i, so vereinfacht sich die Formel zur

Kuder-Richardson-Formel 21

$$\alpha = \alpha_{21} = \frac{n}{n-1}\left[1 - \frac{np(1-p)}{\mathbb{V}(X)}\right]. \tag{3.2.65}$$

Guttmans λ_3: Perfekte Parallelität etwa von Items $Y_1, \ldots, Y_n$ wird es selten geben. Berechnet man die Split-Half-Reliabilität, wo wird sie von der Art, in der die Items in zwei Hälften geteilt werden, abhängen, ebenso wie der Wert des zu dieser Aufteilung gehörenden α-Wertes. Cronbach (1951) betrachtete den Fall eines Tests mit $2n$ Items. Es gibt dann

$$N = \frac{1}{2}\frac{(2n)!}{(n!)^2}$$

verschiedene Weisen, den Test in Split-Half-Teile aufzuteilen. Für jede Aufteilung kann das entsprechende α_j. $j = 1, \ldots, N$ berechnet werden. Für den gesamten Test mit $2n$ Items sei $\alpha = \alpha_{2n}$. Cronbach hat dann gezeigt, dass

$$\mathbb{E}(\alpha_j) = \alpha_{2n}, \tag{3.2.66}$$

wobei $\mathbb{E}(\alpha_j)$ die Erwartung über alle N Split-Half-αs meint.

Der Wert von α hängt von den Korrelationen zwischen den Items ab. Für den Fall negativer Korrelationen kann der Fall eintreten, dass α negativ wird; als untere Abschätzung für die Reliabilität ist dieser Wert aber nutzlos, da $\rho_{xx'}$ stets größer oder gleich Null ist. Guttmann (1945) hat eine robustere untere Abschätzung der Reliabilität hergeleitet, nämlich

$$\lambda_3 = 1 - \frac{1}{\mathbb{V}(X)} \left[\sum_{i=1}^{n} \mathbb{V}(Y_i) + \sqrt{\frac{2n}{n-1} \sum_{i>j} \mathbb{K}^2(Y_i.Y_j)} \right]. \tag{3.2.67}$$

Guttman hat mehrere untere Grenzen hergeleitet, λ_3 ist die dritte, daher der Index 3.

3.2.6 Gewichtung der Testwerte und die Maximierung von α

Es ist im vorangegangenen Abschnitt angenommen worden, dass der Gesamttestwert einfach die Summe der Scores der einzelen Items ist. Es ist aber möglich, dass es besser ist, die einzelnen Items zu gewichten, so dass der Gesamtwert eine gewichtete Summe ist. So kann es nützlich erscheinen, schwierigeren Items ein höheres Gewicht zu geben als leichteren, denn eine schwierigere Aufgabe lösen zu können bedeutet, eine größere Fähigkeit zu haben, als wenn man nur eine leichtere lösen kann. Das Ziel ist, die Gewichtung so vorzunehmen, dass die Qualtität des Tests verbessert wird. Die soll insbesondere bedeuten, dass der Wert von Cronbachs α so groß wie möglich wird.

Es sei X_a der Gesamttestwert für eine Person $a \in \mathcal{P}$, und w_j sei das Gewicht, mit der Score oder Punktwert der j-ten Aufgabe in X_a eingeht. Es ist dann

$$X_a = \sum_{j=1}^{n} w_j Y_{aj}. \tag{3.2.68}$$

Der Erwartungswert von X_a ist dann

$$\mathbb{E}(X_a) = \sum_{j=1}^{n} w_j \mathbb{E}(Y_{aj}), \tag{3.2.69}$$

und die Varianz ist

$$\mathbb{V}(X_a) = \mathbb{V}\left[\sum_{j=1}^{n} w_j Y_{aj} \right] = \sum_{j=1}^{n} w_j^2 \mathbb{V}(Y_{aj}), \tag{3.2.70}$$

wobei angenommen wird, dass die Y_{aj} stochastisch unabhängig sind und mithin alle Kovarianzen verschwinden.

Man kann X_a als die Realisierung einer zufälligen Veränderlichen X auffassen. Dazu stelle man sich vor, dass man in die Population (= Urne) greift und zufällig

eine Person a auswählt. Damit nimmt X den X_a an. X soll jetzt also einen möglichen Testwert bezeichnen, X_a bezeichnet einen bestimmten Wert, nämlich den der Person a. Natürlich kann man jetzt den Erwartungswert und die Varianz von X bestimmen: man hat zunächst

$$X = \sum_{j=1}^{n} w_j Y_j, \tag{3.2.71}$$

wobei Y_j die zufällige Veränderliche ist, die einen möglichen Testwert in der j-ten Aufgabe repräsentiert. Erwartungswert und Varianz von X sind dann

$$\mathbb{E}(X) = \sum_{j=1}^{n} w_j \mathbb{E}(Y_j), \tag{3.2.72}$$

$$\mathbb{V}(X) = \sum_{j=1}^{n} w_j^2 \mathbb{V}(Y_j) + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \mathbb{K}(Y_i, Y_j). \tag{3.2.73}$$

Man bemerke, dass hier die Kovarianzen nicht verschwinden, denn über die Population gemittelt kann es ja Korrelationen zwischen den Variablen Y_i und Y_j geben (Korrelationen zwischen den Items!). Die Ausdrücke für $\mathbb{E}(X)$ und $\mathbb{V}(X)$ tauchen auch – mit Gewichten $w_j = 1$ für alle j – bei der Herleitung von (**??**) auf. Es läßt sich nun zeigen, dass α maximiert wird, wenn für die Gewichte die Ladungen der Items auf der ersten Dimension einer Faktorenanalyse der Items, also der Matrix der Korrelationen $\mathbb{K}(Y_i, Y_j)/\sqrt{\mathbb{V}(Y_i)\mathbb{V}(Y_j)}$, gewählt werden (Lord, 1958).

3.2.7 Schätzung der Reliabilität

Maßzahl für die Reliabilität ist die Korrelation $\rho_{xx'}$, wobei X und X' Testwerte in Paralleltests sind. Man kann $\rho_{xx'}$ einfach schätzen, indem man, für eine hinreichend große Stichprobe, die Stichprobenkorrelation $r_{xx'}$ berechnet.

Retest-Reliabilität: Eine einfache Möglichkeit, einen τ-äquivalenten Test zu erhalten, ist, den gleichen Test zu zwei verschiedenen Zeitpunkten an den gleichen Personen zu administrieren; man spricht dann von *Retest-Reliabilität*. Man setzt dabei voraus, dass das erste Testen keinen Einfluß auf das zweite Testen hat, die Erwartungswerte und Varianzen also gleich bleiben.

Split-Half-Reliabilität: Die zweite Möglichkeit ist, den zu evaluierenden Test in zwei Hälften zu teilen; man spricht von *Split-Half-Reliabilität*. Es gibt verschiedene Möglichkeiten der Aufteilung: (i) *Even-Odd:* hier werden die Items durchnummeriert, die Items mit einer geraden Nummer kommen in Test 1, die mit der ungeraden in Test 2; (ii) *zufällige Aufteilung:* man verteilt die Items nach dem Zufall auf Test 1 bzw. Test 2, (iii) *Itemzwillinge:* man bildet Paare von Items mit gleicher Schwierigkeit und Trennschärfe und sortiert dann jeweils ein Item

eines Paars in Test 1, das andere in Test 2. Tatsächlich berechnet man nun aber die Reliabilität eines Tests von nur halber Länger, so dass der Reliabilitätswert korrigiert werden muß. Spearman-Brown.

Parallelität: Bei der Berechnung der Retest- oder Split-Half-Reliabilität muß man festlegen, was unter "parallel" verstanden werden soll: τ-Äquivalenz, d.h. gleiche τ-Werte und gleiche Fehlervarianzen $\mathbb{V}(\varepsilon)$, oder essentielle τ-Äquivalenz, oder kongenerische Tests. Nach Gleichung (3.2.4), Seite 149, gilt für die Kovarianz der X- und X'-Werte $\mathbb{K}(X, X') = \mathbb{V}(\tau)$ und wegen $\sigma_x^2 = \sigma_{x'}^2$ hat man $\sigma_x \sigma_{x'} = \sigma_x^2$, so dass

$$\mathbb{V}(\tau) = \mathbb{K}(X, X'), \tag{3.2.74}$$

und weiter

$$\mathbb{V}(\varepsilon) = \sigma_x^2 = \mathbb{V}(X) - \mathbb{K}(X, X'). \tag{3.2.75}$$

Gilt also essentielle τ-Äquivalenz und Unkorreliertheit der Fehler, so erhält man eine Schätzung der Reliabilität nach

$$\mathrm{Rel}(X) = \frac{\mathbb{K}(X, X')}{\mathbb{V}(X)}, \tag{3.2.76}$$

wobei für $\mathbb{K}(X, X')$ und $\mathbb{V}(X)$ die jeweiligen Stichprobenschätzungen einzusetzen sind.

Interne Konsistenz: Hier wird Cronbachs α oder, im Falle dichotomer Items, die Kuder-Richardson-Formel (vergl. Seite 167) angewendet. Voraussetzung ist die Eindimensionalität des gemessenen Merkmals. Eine konfirmatorische Faktorenanalyse kann dazu dienen, diese Homogenität festzustellen bzw. Items zu finden, die sich auf einer Skala anordnen lassen.

3.2.8 Reliabilität und Testlänge

Ein Test bestehe aus n parallelen Subtests (Items) mit den Scores X_i, und der Gesamtscore sei

$$X(n) = \sum_{i=1}^{n} X_i. \tag{3.2.77}$$

Man spricht von einem *Test der Länge n*. Die Frage ist, wie die Testparameter von n abhängen.

Der Ausdruck für $X(n)$ kann in der Form

$$X(n) = \sum_{i=1}^{n} (\tau_i + \varepsilon_i) = n\tau + \sum_{i=1}^{n} \varepsilon_i, \quad n = 0, 1, 2, \ldots \tag{3.2.78}$$

geschrieben werden, wenn $\tau_i = \tau$ für alle i. Der Parameter n kann nun durch den kontinuierlichen Zeitparameter t ersetzt werden. Dann ergibt sich

$$X(t) = nt + \sum_{i=1}^{t} \varepsilon_i = t\tau + \varepsilon(t), \tag{3.2.79}$$

(Lord & Novick 81968), p. 104 – der Ausdruck $\sum_{i=1}^{t} \varepsilon_i$ ist mathematisch ein wenig ungenau, weil die diskreten i-Werte ja gerade durch das kontinuierliche t ersetzt wurden, aber es kommt hier nur auf die Idee an). Für eine fest gewählte Person ist der Erwartungswert von $X(t)$ gerade $\mathbb{E}(X) = t\tau$, und $\varepsilon(t) = X(t) - t\tau$. Ein wichtiger Spezialfall ist dann gegeben, wenn $X(t)$ die Anzahl der bis zum Zeitpunkt t gelösten Aufgaben oder beantworteten Fragen ist. Der Fall eines Tests doppelter Länge korrespondiert zu $t = 2$, etc. Es sollen nun die folgenden Annahmen gelten:

$$\begin{aligned} \mathbb{E}(\varepsilon(t)|\tau) &= 0 && (3.2.80)\\ \mathbb{E}[\varepsilon(t_2) - \varepsilon(t_1)]^2 &= a^2|t_2 - t_1| && (3.2.81)\\ \mathbb{E}[(\varepsilon(t_2) - \varepsilon(t_1)(\varepsilon(_4) - \varepsilon(t_3))] &= 0, && (3.2.82) \end{aligned}$$

wobei die Erwartungwerte über die Personen genommen werden. a ist ein Skalenfaktor. Die Intervalle werden als *halb offen* definiert, d.h. etwa $(t_1, t_2] = \{t|t_1 < t \leq t_s\}$. (3.2.82) insbesondere bedeutet, dass die Fehler in nicht überlappenden Zeitintervallen nicht miteinander korrelieren sollen.

(3.2.79) impliziert $\mathbb{E}(X(t)) = t\tau$, so dass

$$\mathbb{E}[X(t)/t] = \tau, \quad \mathbb{V}[X(t)/t] = a^2/t. \tag{3.2.83}$$

Dann folgt

$$\lim_{t\to\infty} \mathbb{E}[X(t)/t] = \tau, \quad \lim_{t\to\infty} \mathbb{V}(X(t)/t) = 0, \tag{3.2.84}$$

denn

$$\mathbb{E}[X(t)/t] = \mathbb{E}[(t\tau + \varepsilon(t)/t] = \tau + \mathbb{E}(\varepsilon(t)/t) = \tau.$$

(3.2.84) bedeutet, dass die Wahrscheinlichkeit, dass $\mathbb{E}[X(t)/t] \neq \tau$ ist, für größer werdenden t-Wert gegen Null geht. Dieser Sachverhalt steht hinter der Aussage, der wahre Wert τ eines Probanden sei der Erwartungswert für einen Test unendlicher Länge.

Bezüglich der Fehler $\varepsilon(t)$ zu verschiedenen Zeitpunkten $t_1, t_2, \ldots$ läßt sich nun eine grundlegende Aussage beweisen:

Satz 3.2.2 *Es gilt*

$$\mathbb{E}[(\varepsilon(t_2) - \varepsilon(t_1))(\varepsilon(t_4) - \varepsilon(t_3))] = a^2 t^*, \tag{3.2.85}$$

wobei $t^* = (t_1, t_2] \cap (t_3, t_4]$.

Anmerkung: $(t_1, t_2] \cap (t_3, t_4]$ ist der Durchschnitt der Zeitintervalle $(t_1, t_2]$ und $(t_3, t_4]$. t^* ist die Länge dieses Durchscnitts. Ist $t_2 < t_3$, so ist der Durchschnitt leer und folglich ist $t^* = 0$.

Beweis: Für den Spezialfall $t_2 = t_4$ reduziert sich (3.2.85) zu (3.2.81). Der schon angemerkte Fall $t^* = 0$ liefert (3.2.82).

Es sei nun der allgemeine Fall $t_1 < t_3 < t_2 < t_4$ gegeben. Der Ausdruck auf der linken Seite von (3.2.85) bleibt richtig, wenn man $\varepsilon(t_3) - \varepsilon(t_3)$ und $\varepsilon(t_4) - \varepsilon(t_4)$ einfügt. Zur Abkürzung werde $\Delta_{ij} = \varepsilon(t_i) - \varepsilon(t_j)$ eingeführt. Man erhält dann für die linke Seite von (3.2.85)

$$\begin{aligned} \mathbb{E}[\Delta_{21} - \Delta_{43}] &= \mathbb{E}[\Delta_{23} + \Delta_{31}][\Delta_{42} + \Delta_{23}] \\ &= \mathbb{E}[\Delta_{23}\Delta_{42}] + \mathbb{E}[\Delta_{31}\Delta_{42}] \\ &\quad + \mathbb{E}[\Delta_{23}^2] + \mathbb{E}[\Delta_{31}\Delta_{23}] \\ &= a^2 t^*. \end{aligned}$$

Die Details des Ausmultiplizierens und Vereinfachens sind hier übergangen worden, weil sie einerseits zu Übungszwecken selbst durchgeführt werden können und sie andererseits im Folgenden nicht benötigt werden. □

Die folgenden Sätze werden nicht im Einzelnen bewiesen; das Schema der Herleitung gleicht dem von Satz 3.2.2.

Lord & Novick (1968), p. 107, führen die Notation

$$X_g(t_g) = X(t_2) - X(t_1), \quad t_g = t_2 - t_1 \tag{3.2.86}$$

ein; von ihr wird im Folgenden Gebrauch gemacht.

Satz 3.2.3 *Es sei $t_2 > t_1$. Dann gilt*

$$\mathbb{E}\left[\frac{X(t_2) - X(t_1)}{t_2 - t_1} = \mathbb{E}\left(\frac{X_g}{t_g}\right)\right] = {}_-\tau, \tag{3.2.87}$$

mit $\mu_\tau = \mathbb{E}(\tau)$. Weiter sei

$$\mathbb{V}\left(\frac{\varepsilon(t_2) - \varepsilon(t_1)}{t_2 - t_1}\right) = \mathbb{V}\left(\frac{\varepsilon_g}{t_g}\right) = a^2/(t_2 - t_1) \tag{3.2.88}$$

$$\mathbb{V}\left(\frac{X(t_2) - X(t_1)}{t_2 - t_1}\right) = \mathbb{V}\left(\frac{X_g}{t_g}\right) = \mathbb{V}(\tau) - \mathbb{V}\left(\frac{\varepsilon_g}{t_g}\right). \tag{3.2.89}$$

Mit $t_2 = t$ und $t_1 = 0$ folgt dann

$$\mathbb{V}(\varepsilon(t)) = a^2 t \tag{3.2.90}$$

$$\mathbb{V}(\varepsilon(t)/\varepsilon) = a^2/t \tag{3.2.91}$$

$$\mathbb{V}(X(t)) = t^2\mathbb{V}(\tau) + \mathbb{V}(\varepsilon(t)) \tag{3.2.92}$$

$$\mathbb{V}(X(t)/t) = \mathbb{V}(\tau) + \mathbb{V}(\varepsilon(t)/t). \tag{3.2.93}$$

Der folgende Satz enthält einige nützliche Aussagen über Kovarianzen:

Satz 3.2.4 *Es sei $t_1 < t_2$ und $t_3 < t_4$. Dann gilt*

$$\mathbb{K}\left(\frac{\varepsilon(t_2)}{t_2}, \frac{\varepsilon(t_1)}{t_1}\right) = a^2/t_2 \tag{3.2.94}$$

$$\mathbb{K}\left(\frac{X(t)}{t}, \tau\right) = \mathbb{V}(\tau) \tag{3.2.95}$$

$$\mathbb{K}\left(\frac{X(t)}{t}, \frac{\varepsilon(t)}{t}\right) = \mathbb{V}(\varepsilon(t)/t) \tag{3.2.96}$$

$$\mathbb{K}\left(\frac{X(t_2) - X(t_1)}{t_2 - t_1}, \frac{X(t_4) - X(t_3)}{t_4 - t_3}\right) = \mathbb{V}(\tau) + \frac{a^2 t^*}{(t_2 - t_1)(t_4 - t_3)} \tag{3.2.97}$$

Für die Korrelationen zwischen beobachteten, wahren und Fehler-Scores läßt sich die Aussagen des folgenden Satzes herleiten:

Satz 3.2.5 *Es sei wieder $t_1 < t_2$ und $t_3 < t_4$. Dann gelten die Aussagen*

$$\rho^2\left(\frac{X(t)}{t}, \tau\right) = \frac{\mathbb{V}^4(\tau)}{\mathbb{V}(X(t)/t)\mathbb{V}(\tau)} \tag{3.2.98}$$

$$\rho^2\left(\frac{X(t_2) - X(t_1)}{t_2 - t_1}, \frac{X(t_4) - X(t_3)}{t_4 - t_3}\right) = \frac{\mathbb{V}(\tau) - \frac{a^2 t^*}{(t_2 - t_1)(t_4 - t_3)}}{(\mathbb{V}(\tau) + \frac{a^2}{t_2 - t_1})(\mathbb{V}(\tau) + \frac{a^2}{t_4 - t_3})} \tag{3.2.99}$$

Satz 3.2.6 *Es gelten die Aussagen*

$$\lim_{t\to\infty} \rho^2\left(\frac{X(t)}{t}, \tau\right) = 1 = \lim_{t\to 0} \rho^2\left(\frac{X(t)}{t}, \tau\right) = 0 \tag{3.2.100}$$

$$\lim_{t\to\infty} \left(\frac{X(t)}{t}, \frac{\varepsilon(t)}{t}\right) = 0 = \lim_{t\to 0} \left(\frac{X(t)}{t}, \frac{\varepsilon(t)}{t}\right) = 1 \tag{3.2.101}$$

$$\lim_{\substack{t_g\to\infty \\ t_{g'}\to\infty}} \left(\frac{X_g}{t_g}, \frac{X'_{g'}}{t_{g'}}\right) = 1 = \lim_{\substack{t_g\to 0 \\ t_{g'}\to 0}} \left(\frac{X_g}{t_g}, \frac{X'_{g'}}{t_{g'}}\right) = 0 \tag{3.2.102}$$

Diese Aussagen lassen sich verbalisieren:

1. Die Reliabilität eines Tests unendlicher Länge ist 1. Die Reliabilität eines Tests der Länge 0 ist Null.

2. Die Korrelation zwischen dem beobachteten Wert und dem Fehler-Score ist gleich Null. Die Korrelation zwischen dem beobachteten Score und dem Fehler-Score bei einem Test der Länge 0 ist gleich 1.

3. Die Korrelation zwischen parallelen Formen der Länge unendlich eines Tests ist gleich 1. Die Korrelationen zwischen parallelen Formen eines Tests der Länge Null ist wiederum Null.

Diese Aussagen erscheinen auf den ersten Blick extrem, da es ja keine Tests mit unendlicher Länge gibt, und Tests der Länge Null scheinen trivial zu sein. Der Punkt ist, dass man 'unendlich lang' mit 'lang' oder 'länger' übersetzen muß, und die Länge Null auf mit 'kurzer Test'. Die Korrelationen sind dann 'groß' oder 'klein'.

Die vorangehenden Resultate erweisen sich als nützlich, wenn man Fragen nach dem Wert einer Testverlängerung stellt. Um die Formeln zu adaptieren, wird zunächst $t = 1$ gesetzt. Weiter wird $X(t)/t = X$ gesetzt. Ein Test der Länge kt ist dann einfach ein Test der Länge k; er wird mit $X(k)$ bezeichnet.

Satz 3.2.7 *Es gilt*

$$\mathbb{E}\left(\frac{X(k)}{k}\right) = \mathbb{E}\left(\frac{k\tau + \varepsilon(k)}{k}\right) \tag{3.2.103}$$

$$\mathbb{V}\left(\frac{X(k)}{k}\right) = \frac{\mathbb{V}(X)(1 + (k-1)\rho_{xx'})}{k} \tag{3.2.104}$$

$$\mathbb{K}\left(\frac{X(k)}{k}\right) = \frac{\mathbb{V}(X)(1 + (k-1)\rho_{xx'})}{k}, \tag{3.2.105}$$

mit $\rho_{xx'} = \rho(X(1), X'(1))$, $X(1)$ eine Komponente von $X(k)$ und $\mathbb{V}(X)$ ist die Varianz der beobachteten Scores für $t = 1$, (d.h. $k = 1$).

Es ergibt sich nun die Spearman-Brown-Formel für die Reliabilität eines verlängerten Tests. Dazu seien X und X' parallele Messungen mit Einheitslänge, und $X(k)$, $X'(l)$ seien die dazu korrespondierenden Tests der Längen k bzw. l.

Satz 3.2.8 *Ein Test werde um k parallele Einheiten verlängert. Dann gilt*

$$\rho\left(\frac{X(k)}{k}, \frac{X'(k)}{k}\right) = \frac{k\rho_{xx'}}{1 + (k-1)\rho_{xx'}}. \tag{3.2.106}$$

Dieses Ergebniss kann benützt werden, um vorherzusagen, um welchen Betrag sich die Reliabilität verbessert, wenn ein Test um k Einheiten verlängert wird. Die Voraussetzung ist aber – wie in Satz 3.2.8 schon gesagt wurde, dass die Einheiten parallel sind. Kann Parallelität nicht vorausgesetzt werden, sollte man sich auf Cronbachs α (3.2.46), Seite 160, (mit gebotener Vorsicht) beziehen.

3.2.9 Die Schätzung des wahren Wertes eines Probanden

Gegeben sei ein Proband $a \in \mathcal{P}$. Gesucht wird eine Abschätzung seines wahren Wertes τ_a. In Abschnitt 3.2.1 ist darauf hingewiesen worden, dass die Gleichung $X_a = \tau_a + \varepsilon_a$ als Regressionsgleichung aufgefasst werden kann. Für eine *beliebige* Regressionsbeziehung $Y = \alpha + \beta X + e$ hat man bekanntlich $\mathbb{E}(Y) = \alpha + \beta\mathbb{E}(X)$, so dass $\alpha = \mathbb{E}(Y) - \beta\mathbb{E}(X)$, $\beta = \rho_{xy}\sigma_x/\sigma_y$, $\sigma_x^2 = \mathbb{V}(X)$, $\sigma_y^2 = \mathbb{V}(Y)$.

Bei der Beziehung $X = \tau + \varepsilon$ gilt insbesondere $\beta_{x\tau} = 1$, und da generell für den Korrelationskoeffizienten $\rho_{x\tau} = \beta_{x\tau}\sigma_\tau/\sigma_x$ gilt, hat man insbesondere

$$\rho_{x\tau} = \frac{\sigma_\tau}{\sigma_x}.$$

Andererseits will man τ vorhersagen, d.h. man betrachtet die Regressionsgleichung

$$\tau = \beta_{\tau x}X + \alpha - \varepsilon, \tag{3.2.107}$$

und es folgt

$$\rho_{\tau x} = \beta_{\tau x}\frac{\sigma_x}{\sigma_\tau} = \frac{\sigma_\tau}{\sigma_x},$$

so dass

$$\beta_{\tau x} = \frac{\sigma_\tau^2}{\sigma_x^2} = \rho_{xx'}, \tag{3.2.108}$$

d.h. der Regressionskoeffizient für die Regression von τ auf den Messwert X ist gleich dem Reliabilitätskoeffizienten. Also hat man, mit

$$\alpha = \mathbb{E}(\tau) - \beta\tau x\mathbb{E}(X) = \mathbb{E}(X) - \rho_{xx'}\mathbb{E}(X) = \mathbb{E}(X)(1 - \rho_{xx'}),$$

$$\hat{\tau} = \rho_{xx'}X + (1 - \rho_{xx'})\mathbb{E}(X). \tag{3.2.109}$$

Der wahre Wert τ wird also vorausgesagt einerseits durch den Messwert X, und andererseits durch den Erwartungswert $\mathbb{E}(X)$ der Messwerte in der Population, wobei X mit dem Reliabilitätskoeffizienten $\rho_{xx'}$ und $\mathbb{E}(X)$ mit dessen Komplement $1 - \rho_{xx'}$ gewichtet wird. Bei großer Reliabilität, $\rho_{xx'} \approx 1$, bekommt der Messwert X ein starkes Gewicht, während wegen$1 - \rho_{xx'} \approx 0$ der Erwartungswert ein geringes Gewicht bekommt. Bei geringer Reliabilität, $\rho_{xx'} \approx 0$, ist es gerade umgekehrt.

StandardMessfehler und Standardschätzfehler Aus $X = \tau + \varepsilon$ folgt

$$\mathbb{V}(X) = \mathbb{V}(\tau) + \mathbb{V}(\varepsilon) + 2\mathbb{K}(\tau, \varepsilon),$$

und da $\mathbb{K}(\tau, \varepsilon) = 0$, folgt

$$\mathbb{V}(X) = \mathbb{V}(\tau) + \mathbb{V}(\varepsilon), \tag{3.2.110}$$

so dass

$$1 = \frac{\mathbb{V}(\tau)}{\mathbb{V}(X)} + \frac{\mathbb{V}(\varepsilon)}{\mathbb{V}(X)} = \rho_{xx'} + \frac{\mathbb{V}(\varepsilon)}{\mathbb{V}(X)}. \tag{3.2.111}$$

Hieraus folgt sofort

$$\sqrt{\mathbb{V}(\varepsilon)} = \sigma_\varepsilon = \sigma_x\sqrt{1 - \rho_{xx'}}, \tag{3.2.112}$$

mit $\sigma_x = \sqrt{\mathbb{V}(X)}$. σ_ε ist der *StandardMessfehler*.

Aus $X = \tau + \varepsilon$ folgt

$$\tau = X - \varepsilon, \tag{3.2.113}$$

und

$$\mathbb{V}(\tau) = \mathbb{V}(X) - \mathbb{V}(\varepsilon), \tag{3.2.114}$$

$$\frac{\mathbb{V}(\tau)}{\mathbb{V}(X)} = 1 - \frac{\mathbb{V}(\varepsilon)}{\mathbb{V}(X)}. \tag{3.2.115}$$

In (3.2.5) wurde die Beziehung $\rho_{xx'} = \rho^2_{x\tau} = \mathbb{V}(\tau)/\mathbb{V}(X)$ genannt. Dann kann (3.2.115) in der Form

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

geschrieben werden.

Versucht man, den wahren Wert einer Person aufgrund seines Scores vorherzusagen, vergl. (3.2.113), so wird man diesen Wert nicht fehlerfrei bestimmen können: da der Score X fehlerbehaftet ist, wird auch die Vorhersage des τ-Wertes einen Fehler haben. Den individuellen Fehler wird man nicht angeben können, – könnte man ihn angeben, so könnte man seine Vorhersage ja korrigieren und hätte damit eine fehlerfreie Vorhersage des wahren Wertes, aber gerade diese kann man ja gerade nicht haben. Man kann aber den "durchschnittlichen" Fehler angeben. In der einfachen Regression gilt ja

$$\sigma_e = \sqrt{\mathbb{V}(e)} = \sigma_y \sqrt{1 - r^2_{xy}}.$$

Ersetzt man hier y durch τ, so erhält man

$$\sigma_\epsilon = \sigma_\tau \sqrt{1 - \rho^2_{\tau x}}, \tag{3.2.116}$$

wobei σ_ϵ für σ_e geschrieben wurde. σ_ϵ heißt *Standardschätzfehler*. Definiert man den *Standardvorhersagefehler* gemäß

$$\sigma_\Delta = \sigma_x \sqrt{1 - \rho^2_{xx'}}, \tag{3.2.117}$$

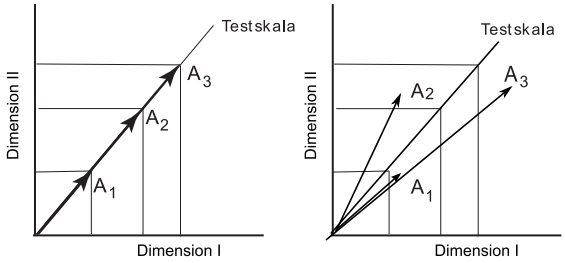
so hat man die Ungleichung

$$\sigma_\Delta \geq \sigma_\varepsilon \geq \sigma_\epsilon. \tag{3.2.118}$$

Die Gleichheitszeichen gelten nur dann, wenn entweder $\rho_{xx'} = 0$ oder $\rho_{xx'} = 1$ (Lord & Novick, p. 68).

Die Vorhersage $\hat{\tau}$ ist wegen des unvermeidlichen Fehlers nicht genau; man wird also ein Konfidenzintervall berechnen wollen. Üblicherweise wird die Normalverteilung für $\hat{\tau}$ angenommen, so dass man $\hat{\tau} \pm \sigma_\epsilon z_{\alpha/2}$ berechnet. Die Annahme der Normalverteilung muß natürlich nicht korrekt sein. Interessanterweise wird die Problematik dieser Annahme selten diskutiert. Eine Ausnahme ist Lumsden (1976), der eine Reihe von Problemen in Zusammenhang mit dem Begriff des wahren Wertes diskutiert.

Abbildung 21: Homogener (eindimensionaler) Test: links idealer Test, rechts mit zufälligen Abweichungen von der Testskala



3.3 Itemcharakteristika

Im Folgenden werden die Begriffe der Schwierigkeit und der Trennschärfe, die ja schon in Abschnitt 2.6 im Zusammenhang mit Itemfunktionen vorgestellt wurden, noch einmal im Rahmen der KTT definiert.

3.3.1 Die Homogenität von Items

Der Begriff der Homogenität von Items ist bereits charakterisiert worden und wird hier wegen der Systematik noch einmal erwähnt: Item sind homogen, wenn sie dasselbe Merkmal messen. Dieses Merkmal kann eine Kombination von p Merkmalen sein, so dass ein Item als p-dimensionaler Vektor $\mathbf{x}(x_1, \ldots, x_p)'$ darstellbar ist; die Komponenten x_i, $i = 1, \ldots, p$ sind die Ausmaße, in denen diese Merkmale benötigt werden, damit das Item "positiv" beantwortet werden kann. k Items sind *homogen*, wenn die entsprechenden Vektoren $\mathbf{x}_j$, $j = 1, \ldots, k$ alle dieselbe Richtung haben und sich nur hinsichtlich ihrer Länge unterscheiden, so dass etwa $\mathbf{x}_{j'} = \alpha \mathbf{x}_j$, $\alpha > 0$. Dann gilt auch $x_{ij'} = \alpha x_{ij}$ für alle i, d.h. die Zusammensetzung der Items ist bis auf Proportionalitätsfaktoren, die für alle Komponenten gelten, gleich. Abbildung 21 (links) zeigt den Idealfall, rechts wird der fehlerbehaftete Fall illustriert. Homogene Items sind durch hohe Korrelationen charakterisiert. So seien $\mathbf{x}_j$ und $\mathbf{x}_{j'}$ zwei Vektoren, die homogene Items repräsentieren, so dass $\mathbf{x}_{j'} = \alpha \mathbf{x}_j$. Die Korrelation zwischen ihnen ist dann durch

$$r_{jj'} = \cos \theta_{jj'} = \frac{\mathbf{x}_j' \mathbf{x}_{j'}}{\|\mathbf{x}_j h\| \|\mathbf{x}_{j'}} = \frac{\alpha \|\mathbf{x}\|^2}{\alpha \|\mathbf{x}_j\|^2} = 1, \tag{3.3.1}$$

was auch schon daraus folgt, dass die Vektoren eine identische Orientierung haben müssen, so dass $\theta = 0$, und $\cos 0 = 1$. Dies gilt für alle Items, deren Vektoren dieselbe Orientierung haben, unabhängig von der Schwierigkeit der Items, die man durch die Länge $\|\mathbf{x}_j\|$ der Vektoren repräsentieren kann: die α kürzen sich ja stets heraus.

Perfekte Homogenität wird man höchst selten finden; nimmt man also an, dass sich die $\mathbf{x}_j$ nur wegen kleiner "Fehler" in der Orientierung unterscheiden, so werden die Korrelationen kleiner als 1 sein. Sie sollten allerdings einigermaßen hoch sein, um von homogenen Items sprechen zu können.

In Punkt 2, Seite 161 wird die Homogenität im Zusammenhang mit Cronbachs α diskutiert. Man denke auch an den Begriff der lokalen Homogenität, wie er von Ellis & Wollenberg (1993) eingeführt wurde (vergl. Seite 69).

3.3.2 Die Schwierigkeit von Items

Es sei Y_{ij} die (numerische) Kodierung für die i-te Person beim g-ten Item, $i = 1, \ldots, m$ und $g = 1, \ldots, n$, dh es gebe m getestete Personen und n Items. Im vorangegangenen Abschnitt 3.2.6 ist die Gewichtung von Items eingeführt worden. Eine solche Gewichtung kann im Folgenden mitgedacht werden, indem man sich die Gewichte in die Score-Variablen y_{ig} aborbiert vorstellt. Dazu definiert man einfach $y'_{ig} = w_j y_{ig}$ und benennt dann die y'_{ig} wieder in y_{ig} um. Der Testscore der i-ten Person ist dann die Summe

$$x_i = \sum_{g=1}^{n} y_{ig}. \tag{3.3.2}$$

Nimmt man den Erwartungswert auf beiden Seiten, so erhält man

$$\mathbb{E}(x_i) = \tau_i = \sum_{g=1}^{n} \mathbb{E}(y_{ig}). \tag{3.3.3}$$

Es werden nun speziell dichotome Items betrachtet, so dass $Y_{ig} \in \{0, 1\}$, dh Y_{ig} kann entweder den Wert 0 oder den Wert 1 annehmen. Dann erhält man für den Erwartungswert τ_{x_g} für das g-te Item

$$\tau_{x_g} = \sum_{i=1}^{m} \pi_{ig} = \pi_g, \tag{3.3.4}$$

wobei π_g die Wahrscheinlichkeit ist, dass ein Item "positiv" beantwortet wird; tatsächlich ist ja jetzt

$$\mathbb{E}(y_{ig}) = \pi_g \cdot 1 + (1 - \pi_g) \cdot 0 = \pi_g,$$

wobei die Erwartung über alle Personen aus $\mathcal{P}$ genommen wird.

Für die Klassische Testtheorie (KTT) gilt nun

Definition 3.3.1 *Die Wahrscheinlichkeit π_g ist die* Schwierigkeit *des g-ten Items.*

In Abschnitt 2.6.3 wurde die Schwierigkeit eines Items I_g als der Parameter κ_g eingeführt, wobei κ_g eine Art Schwellenwert ist: die Merkmalsausprägung zum Zeitpunkt des Testens muß mindestens gleich κ_g sein, damit es "gelöst" wird, d.h.

$$F_g(\theta) = \int_{a_g(\theta-\kappa_g)}^{\infty} \varphi(z)dz,$$

wobei φ die Standardnormalverteilung ist. κ_g entspricht aber nicht dem Index π_g in der KTT. π_g ist ein Mittelwert von Antworten über eine Population. $F_g(\theta)$ gibt die Wahrscheinlichkeit einer Antwort (Lösung) als Funktion des Fähigkeitsparamters[43] θ an. Eine Antwort – also der Wert $Y_g = 1$ der entsprechenden Indikatorvariable – erlaubt aber keinen eindeutigen Rückschluß auf den Wert von θ. Eine Person, die das Item I_g beantwortet hat, hat einen Fähigkeitsparameter θ, oder θ', oder θ'', etc. Skaliert man die θ so, dass in der Population $\mathbb{E}(\theta) = 0$ und $\mathbb{V}(\theta) = 1$ ist nimmt man an, dass die θ-Werte *in der Population* standardnormalverteilt, also $N(0,1)$-verteilt sind, so ist die Wahrscheinlichkeit, dass bei einer Person aus der Population $Y_g = 1$ gilt, durch

$$P(Y_g = 1) = F_g(\theta)\varphi(\theta) + F_g(\theta')\varphi(\theta') + \cdots,$$

wobei φ die Dichte der $N(0,1)$-Verteilung ist. Dies ist eine sehr ungenaue Schreibweise, die nur verdeutlichen soll, was gemeint ist, dass nämlich *über alle* θ, die in der Population auftreten können, summiert wird. Dies führt zu dem Ausdruck

$$P(Y_g = 1) = \pi_g = \int_{-\infty}^{\infty} F_g(\theta)\varphi(\theta)d\theta. \tag{3.3.5}$$

Es läßt sich zeigen, dass dieses Integral auf eine einfache Form gebracht werden kann (Lord & Novick (1968), p. 377), nämlich

$$\pi_g = \int_{\gamma_g}^{\infty} \varphi(z)dz = \Phi(-\gamma_g), \tag{3.3.6}$$

mit

$$\gamma_g = \frac{a_g b_g}{\sqrt{1 + a_g^2}}. \tag{3.3.7}$$

Abschnitt 2.6.3, Gleichung (2.6.15) wurde $\alpha_g = 1/\sigma_g$, $b_g = \kappa_g$ angegeben, wobei σ_g^2 die Varianz der individuellen Verteilung des Merkmals η ist (die für alle Personen gleich sein soll). Setzt man diese Ausdrücke für α_g und b_g in (3.3.7) ein, erhält man

$$\pi_g = \frac{\kappa_g}{\sqrt{1 + \sigma_g^2}}. \tag{3.3.8}$$

[43] Der auch als Erwartungswert einer individuell fluktuierenden Merkmalsausprägung gesehen werden kann.

Damit erhält man eine Beziehung zwischen dem Parameter κ_g, der die Schwierigkeit *an sich* des Items I_g angibt, und dem populationsabhängigen Schwierigkeitsparameter π_g, der den Anteil der $(Y_g = 1)$-Antworten in der Population angibt. Für den Spezialfall σ_g^2 =konstant für alle Items I_g ist π_g proportional zum Schwellenparameter κ_g; dieser Spezialfall wird beim Rasch-Modell angenommen. Generell wird die Population durch die Varianz σ_g^2 charakterisiert. Je größer σ_g^2 ist, desto kleiner wird π_g. Das ist plausibel, denn mit größer werdender intraindividueller Varianz des Merkmals η wird die Wahrscheinlichkeit größer, dass $\eta > \kappa_g$ ist. Also werden mehr Personen in der betrachteten Population das Item beantworten oder lösen.

Dass bei dieser Interpretation Populationen nur in Bezug auf die intraindividuelle Varianz unterschieden werden, mag unbefriedigend erscheinen. Die Interpretation ergibt sich aus der Standardisierung der θ-Verteilung, die Lord & Novick vorgenommen haben, um das Integral auf die einfache Form (3.3.6) bringen zu können. Läßt man populationsspezifische $\mathbb{E}(\theta)$-Werte zu, ergibt sich eine auch von $\mathbb{E}(\theta)$ abhängende Interpretation des π_g-Wertes.

Schätzung der Schwierigkeit von Items: Es wird zunächst der Fall dichotomer Items betrachtet. Man hat zwei mögliche Fälle, nämlich den Fall, das Probanden raten können; dies kann der Fall sein, wenn Aufgaben gelöst werden, oder den Fall, dass Raten vernachlässigt werden kann, wie etwa im Fall von Persönlichkeitstests.

Für eine Stichprobe erhält man als Schätzung für τ_g, wenn kein Raten berücksichtigt werden muß,

$$\hat{p}_g = \bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_{ig} \tag{3.3.9}$$

und $\hat{p}_g$ ist eine Schätzung für π_g. Muß die Möglichkeit in Betracht gezogen werden, dass die Probanden auch raten, so wird im Falle von k Antwortkategorien die Schätzung

$$\hat{p}_g = \frac{1}{N} \left(N_r - \frac{N_f}{k-1} \right) \tag{3.3.10}$$

berechnet, wobei N der Stichprobenumfang, N_r die Anzahl der korrekten und N_f die Anzahl der falschen Beantwortungen ist. Für den Spezialfall $k = 2$, d.h. für dichotome Items ergibt sich daraus

$$\hat{p}_g = \frac{N_r - N_f}{N}. \tag{3.3.11}$$

Polynome Items: Der Schwierigkeitsindex bei nicht-dichotomen Items wird üblicherweise als das arithmetische Mittel der Einschätzungen definiert. Gegebenfalls repräsentiert allerdings der Median die Schwierigkeit.

3.3.3 Die Trennschärfe

Definition 3.3.2 *Die Trennschärfe*[44] *eines Items ist durch die Korrelation* $\rho_{Xg} = \rho(y_g, X)$ *zwischen den Werten* Y_g *und dem Gesamttestwert* X *gegeben.* ρ_{Xg} *heißt auch* Trennschärfeindex *für das* g*-te Item.*

Der Gesamtwert der i-ten Person im Test ist $X_i = \sum_g y_{ig}$. Die Korrelation zwischen irgendzwei Items – genauer gesagt: zwischen den Repräsentationen Y_g und Y_h der Antworten – ist

$$\rho_{gh} = \frac{\mathbb{K}(Y_g, Y_h)}{\sigma_g \sigma_h}, \tag{3.3.12}$$

mit $\sigma_g = \sqrt{\mathbb{V}(Y_g)}$, $\sigma_h = \sqrt{\mathbb{V}((Y_h)}$ und $\mathbb{V}(Y_g) = \mathbb{E}[(Y_g - \mathbb{E}(Y_g))^2]$; $\mathbb{V}(Y_h)$ ist analog definiert. Man erinnere sich, dass $\rho_{jj} = \sigma_j^2 \mathbb{K}(Y_j, Y_j)$. Damit kann man die Varianz der X_i-Werte in der Form

$$\mathbb{V}(X) = \sigma_x^2 = \sum_{g=1}^{n} \sum_{h=1}^{n} \sigma_g \sigma_h \rho(Y_g, Y_h) = \sum_{g=1}^{n} \sum_{h=1}^{n} \sigma_g \sigma_h \rho_{gh} \tag{3.3.13}$$

schreiben. Sind die Items insbesondere dichotom, so ist

$$\mathbb{V}(Y_g) = \pi_g(1 - \pi_g), \tag{3.3.14}$$

und π_g ist die Schwierigkeit des g-ten Items. Es gilt der

Satz 3.3.1 *Die Streuung* σ_x *der* X*-Werte ist eine gewogene Summe der Trennschärfen der Items, d.h. es gilt*

$$\sigma_x = \sum_g \sigma_g \rho_{Xg}. \tag{3.3.15}$$

Beweis: Es ist

$$\mathbb{V}(X) = \sigma_x^2 = \mathbb{K}(X, X) = \mathbb{K}(X, \sum_g Y_g) = \mathbb{E}(X \sum_g Y_g) - \mathbb{E}(X)\mathbb{E}(\sum_g Y_g),$$

und diese Gleichung kann in der Form

$$\sigma_x^2 = \sum_g \mathbb{E}(XY_g) - \sum_g \mathbb{E}(X)\mathbb{E}(Y_g) = \sum_g [\mathbb{E}(XY_g) - \mathbb{E}(X)\mathbb{E}(Y_g)]$$

geschrieben werden, d.h. es gilt

$$\sigma_x^2 = \sum_g \mathbb{K}(X, Y_g) = \sum_g \sigma_x \sigma_g \rho_{Xg} = \sigma_x \sum_g \sigma_g \rho_{Xg},$$

so dass nach Division durch σ_x die Gleichung (3.3.15) resultiert. □

[44] Discriminating power

Für die Trennschärfe selbst gilt nun

Satz 3.3.2 *Der Trennschärfeindex für das g-te Item ist durch*

$$\rho_{Xg} = \frac{1}{\sigma_x} \sum_{h=1}^{n} \sigma_h \rho_{gh} \tag{3.3.16}$$

gegeben.

Beweis:

$$\rho_{Xg} = \rho(X, Y_g) = \frac{\mathbb{K}(\sum_h Y_h, Y_g)}{\sigma_x \sigma_g} = \frac{\sum_h \mathbb{E}(Y_h Y_g) - \sum_h \mathbb{E}(Y_h)\mathbb{E}(Y_g)}{\sigma_x \sigma_g},$$

d.h.

$$\rho_{Xg} = \frac{\sum_h \mathbb{K}(Y_h Y_g)}{\sigma_x \sigma_g} = \frac{\sigma_g \sum_h \sigma_h \rho_{gh}}{\sigma_x \sigma_g},$$

woraus sofort (3.3.16) folgt.

Der Ausdruck (3.3.16) zeigt, dass der Trennschärfeindex eine gewogene Summe der Korrelationen zwischen dem g-ten und den übrigen Items ist. Ist ρ_{Xg} groß, so bedeutet dies, dass man von der Antwort auf das g-te Item auf den Gesamttestwert schließen kann. Konkret heißt dies, dass etwa die Beantwortung des g-ten Items auf einen hohen X-Wert, die Nichtbeantwortung auf einen niedrigen X-Wert schließen läßt. Insofern trennt das g-te Item Personen mit hohen X-Werten von solchen mit niedrigen X-Werten. Ist andererseits $\rho_{Xg} \approx 0$, so ist dieser Schluß nicht zulässig; das Item trennt dann Personen mit hohem X-Wert nicht von solchen mit niedrigem X-Wert.

Auch bei der Trennschärfe kann die Beziehung zur Itemfunktion (Ogive-Modell) hergestellt werden. Wird wieder angenommen, dass θ in der Population $N(0,1)$-verteilt ist (also $\mathbb{E}(\theta) = 0$ und $\mathbb{V}(\theta) = 1$), so findet man (Lord & Novick (1968), p. 378)

$$\rho_{\theta g} = \frac{a_g}{\sqrt{1 + a_g^2}} \frac{\varphi(\gamma_g)}{\sqrt{\pi_g(1 - \pi_g)}} \tag{3.3.17}$$

wobei $\rho_{\theta g}$ die punkt-biseriale Korrelation zwischen dem Item I_g und θ ist. Die Größe

$$\rho'_g = \frac{a_g}{\sqrt{1 + a_g^2}} \tag{3.3.18}$$

ist gleich der biserialen Korrelation $\rho'_{\theta g}$ zwischen θ und der Indikatorvariablen Y_g, und diese Korrelation ist ein Maß für die Trennschärfe des Items I_g. Für den Parameter α_g ergibt sich

$$\alpha_g = \frac{\rho'_g}{\sqrt{1 - \rho'^2_g}}. \tag{3.3.19}$$

a_g ist demnach eine monoton wachsende Funktion der biserialen Korrelation zwischen dem Item I_g und der latenten Variable η.

Schätzung der Trennschärfe: Es werden zunächst dichotome Items betrachtet. Der Trennschärfeindex ist als Korrelation der Itembeantwortung mit dem Gesamttestwert definiert worden. Der Produkt-Moment-Korrelationskoeffizient wird, wenn eine der Variablen eine $(0,1)$-Variable ist, zum punktbiserialen Korrelationskoeffizient.

Allerdings ist hier zu beachten, dass die Korrelation der Antworten auf I_g mit dem Gesamtwert einen verfälschten Wert geben kann, denn die Antwort auf I_g ist ja im Gesamtwert enthalten. Man korrigiert dies, indem man die Antworten auf I_g mit der von der Antwort auf I_g *bereinigten* Gesamtantwort korreliert. Diese Korrektur ist unter dem Namen *part-whole-correction* bekannt:

$$\rho(y_g, X - y_g) = \frac{\rho_{y_g x}(\sigma_x - \sigma_g)}{\sqrt{\sigma_g^2 + \sigma_x^2 - 2\rho_{xg}\sigma_x\sigma_g}} \tag{3.3.20}$$

3.3.4 Itemvalidität

Die Validität eines Items ist durch die Korrelation $\rho(Y_g, \xi) = \rho_{g\xi}$ von Y_g mit einem externen Kriterium[45] ξ gegeben, also

$$\rho_{g\xi} = \frac{\mathbb{K}(Y_g, \xi)}{\sigma_g \sigma_\xi}. \tag{3.3.21}$$

3.4 Validität

3.4.1 Die allgemeine Definition

Definition 3.4.1 *Es sei $X = \sum_g Y_g$ der Testscore für einen Test, und es sei ξ den Wert eines externen Kriteriums, also einer Variable, die das zu messende Merkmal repräsentiere. Die Gültigkeit des Tests ist die Korrelation $\rho_{X\xi}$ zwischen den X-Werten und den korrespondierenden ξ-Werten.*

Es gilt der

Satz 3.4.1 *Die Korrelation $\rho_{X\xi}$ ist durch*

$$\rho_{X\xi} = \frac{\mathbb{K}(\xi, X)}{\sigma_x \sigma_\xi} = \frac{\sum_g \sigma_g \rho_{g\xi}}{\sum_g \sigma_g \rho_{gX}}. \tag{3.4.1}$$

gegeben, wobei $\rho_{g\xi}$ die Validität des g-ten Items ist, und ρ_{gX} die Trennschärfe dieses Items.

[45] Ursprünglich wurde η statt ξ ("xi") geschrieben. Die Umbenennung erfolgte, weil der Buchstabe η bereits in einem anderen Zusammenhang eingeführt wurde.

Beweis: Die Kovarianz zwischen X und ξ kann in der Form

$$\mathbb{K}(\xi, X) = \mathbb{K}(\xi, \sum_g Y_g) = \mathbb{E}(\xi \sum_g Y_g) - \mathbb{E}(\xi)\mathbb{E}(\sum_g Y_g)$$

angeschrieben werden, d.h.

$$\mathbb{K}(\xi, X) = \sum_g \mathbb{E}(\xi Y_g) - \sum_g \mathbb{E}(\xi)\mathbb{E}(Y_g) = \sum_g [\mathbb{E}(\xi Y_g) - \mathbb{E}(\xi)\mathbb{E}(Y_g)],$$

so dass

$$\mathbb{K}(\xi, X) = \sum_g \mathbb{K}(\xi, Y_g) = \sigma_\xi \sum_g \sigma_g \rho_{g\xi}$$

folgt. Dann ist

$$\rho_{X\xi} = \frac{\mathbb{K}(\xi, X)}{\sigma_\xi \sigma_x} = \frac{\sigma_\xi \sum_g \sigma_g \rho_{g\xi}}{\sigma_\xi \sigma_x} = \frac{\sum_g \sigma_g \rho_{g\xi}}{\sigma_x}, \tag{3.4.2}$$

und wegen Satz 3.3.1, Gleichung (3.3.15) folgt (3.4.1). □

Die Gesamtvalidität eines Tests ist also proportional zur gewogenen Summe der Itemvaliditäten und umgekehrt proportional zur gewogenen Summe der Trennschärfeindices.

$\rho_{X\xi}$ wird offenbar maximal, d.h. gleich 1, wenn nach (3.4.1)

$$\sum_g \sigma_g \rho_{g\xi} = \sum_g \sigma_g \rho_{gX}.$$

Dies ist sicher der Fall, wenn $\rho_{g\xi} = \rho_{gX}$ für alle Items I_g, wenn also alle Itemgültigkeiten gleich den Itemtrennschärfen sind. Da $\rho_{X\xi} \leq 1$ sein muß, kann man davon ausgehen, dass im Allgemeinen $\rho_{g\xi} \leq \rho_{gX}$, d.h. die Itemgültigkeiten sind i. A. kleiner, höchstens gleich den korrespondierenden Trennschärfen.

Setzt man den Ausdruck für σ_x aus (3.3.13) in (3.4.2) ein, so erhält man den zu (3.4.1) äquivalenten Ausdruck

$$\rho_{X\xi} = \frac{\sum_g \sigma_g \rho_{g\xi}}{\sqrt{\sum_g \sum_h \sigma_g \sigma_h \rho_{gh}}}. \tag{3.4.3}$$

Dies heißt, dass die Gültigkeit eines Tests, wie schon bemerkt, proportional zur gewogenen Summe der Itemvaliditäten ist, und dass sie umgekehrt proportional zur Wurzel der gewogenen Summe der Iteminterkorrelationen ist: je kleiner diese Korrelationen sind, desto gültiger wird der Test. Kleine Iteminterkorrelationen können wiederum bedeuten, dass die Items verschiedene Dimensionen (Faktoren) messen. Es ist plausibel, dass insbesondere komplexe Merkmale um so valider gemessen werden, je mehr Aspekte (Dimensionen) des Merkmals erfasst werden.

Die Definition der Validität eines Tests kann verallgemeienert werden zum Begriff der Validität in Bezug auf einen anderen Test mit den Scores Y:

Definition 3.4.2 *Die zufällige Veränderliche X repräsentiere die Messungen eines Tests, und die zufällige Veränderliche Y repräsentiere die Messungen eines anderen Tests. Die* Validität *der Messungen X in Bezug auf die Messungen Y ist durch den Korrelationskoeffizienten*

$$\rho_{xy} = \frac{\mathbb{K}(X, Y)}{\sigma_x \sigma_y} \tag{3.4.4}$$

gegeben, wobei $\sigma_x = \sqrt{V(X)}$ und $\sigma_y = \sqrt{\mathbb{V}(Y)}$.

Die Idee hinter dieser Definition ist natürlich, dass man für die zu messende Variable, für die ein Test die Werte X liefert, letztlich stets einen zweiten Test derselben Variablen benötigt; dieser liefere die Messwerte Y. So können die Y-Werte Messungen anhand eines bereits etablierten Tests sein, über dessen Gültigkeit bereits Informationen vorliegen. Im Idealfall weiß man bereits, dass dieser Test das interessierende Merkmal sehr gut erfasst. Der Test, dessen Gültigkeit erfasst werden soll, könnte etwa eine Kurzform des bereits vorliegenden Tests sein. Sind X die Werte eines Berufseignungstests, könnten Y die Noten für die Abschlußprüfung der Berufsausbildung sein.

3.4.2 Parallele und nichtparallele Messungen

Sind die Messungen X und Y nicht parallel, so folgt

$$\rho_{xy} = \frac{\mathbb{K}(\tau_x + \varepsilon_x, \tau_y + \varepsilon_y)}{\sigma_x \sigma_y},$$

und

$$\mathbb{K}(\tau_x + \varepsilon_x, \tau_y + \varepsilon_y) = \mathbb{E}[(\tau_x + \varepsilon_x)(\tau_y + \varepsilon_y)] - \mathbb{E}(\tau_x + \varepsilon_x)\mathbb{E}(\tau_y + \varepsilon_y).$$

Multipliziert man die Ausdrücke aus und berücksichtigt die Befunde (3.1.2) bis (**??**), so ergibt sich hieraus

$$\mathbb{K}(X, Y) = \mathbb{K}(\tau_x, \tau_y), \tag{3.4.5}$$

die Kovarianz zwischen den X- und Y-Werten ist gleich der Kovarianz zwischen den wahren Werten τ_x und τ_y. Der *Gültigkeitskoeffizient* ρ_{xy} ist also im Falle nicht paralleler Messungen durch

$$\rho_{xy} = \frac{\mathbb{K}(\tau_x, \tau_y)}{\sigma_x \sigma_y} \tag{3.4.6}$$

gegeben. Für den Spezialfall $\tau_x = \tau_y$, $\sigma_x = \sigma_y$ erhält man daraus wieder den Fall paralleler Messungen und es ist

$$\rho_{xy} = \frac{\mathbb{V}(\tau)}{\mathbb{V}(X)} = \rho_{x\tau}^2; \tag{3.4.7}$$

die Validität ist dann gleich der Reliabilität des Tests. Im Falle nicht-paralleler Tests wird die Validität ρ_{xy} natürlich kleiner als dieser Maximalwert sein, so dass man allgemein

$$\rho_{xy} \leq \rho_{x\tau}^2 = \mathrm{Rel}(X) \tag{3.4.8}$$

gelten wird, d.h. die Reliabilität eines Tests ist eine obere Grenze für die Validität des Tests. Dieser Punkt wird im folgenden Abschnitt noch weiter elaboriert: es wird gezeigt, dass die Validität i. A. kleiner oder höchstens gleich der Wurzel aus der Reliabilität ist. Dies ist kein Widerspruch zu (3.4.8), denn wenn die Reliabilität gleich 1 ist, gilt notwendig auch (3.4.8), da eine Korrelation nie größer als 1 sein kann.

3.4.3 Attenuierung, Reliabilität und Validität

Häufig ist man nicht an den Messungen für sich interessiert, sondern an den durch sie (hoffentlich) erfassten *latenten Variablen* und an den Beziehungen zwischen den verschiedenen latenten Variablen. Es seien also X und Y die Messungen zweier Tests, die die latenten Variablen τ_x bzw. τ_y erfassen mögen. Da die X- und Y-Werte fehlerbehaftet sind, wird die Korrelation ρ_{xy} kleiner sein als die eigentlich interessierende Korrelation $\rho(\tau_x, \tau_y)$.

Attennuierung: Die Gleichung (3.4.6) liefert bereits eine Beziehung zwischen $\mathbb{K}(\tau_x, \tau_y)$ und ρ_{xy}. Aus ihr läßt sich aber noch nicht allgemein ablesen, ob stets $\rho_{xy} \geq \rho(\tau_x, \tau_y)$ oder $\rho_{xy} \leq \rho(\tau_x, \tau_y)$ gilt.

Satz 3.4.2 *Es ist*

$$\rho(\tau_x, \tau_y) = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}\rho_{yy'}}}, \tag{3.4.9}$$

sowie

$$\rho(\tau_x, \tau_y) \geq \rho_{xy}. \tag{3.4.10}$$

Beweis: Aus der Definition der Reliabilität in (3.2.3), Seite 148 folgt

$$\mathbb{V}(\tau_x) = \rho_{xx'}\mathbb{V}(X), \;\; \mathbb{V}(\tau_y) = \rho_{yy'}\mathbb{V}(Y).$$

Dann kann man $\rho(\tau_x, \tau_y)$ in der Form

$$\rho(\tau_x, \tau_y) = \frac{\mathbb{K}(\tau_x, \tau_y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)\rho_{xx'}\rho_{yy'}}}$$

schreiben, und wegen (3.4.6) erhält man daraus (3.4.9).

Würde $\rho(\tau_x, \tau_y) = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}\rho_{yy'}}} \leq \rho_{xy}$ gelten, so folgte $1 \leq \sqrt{\rho_{xx'}\rho_{yy'}}$, was aber nicht sein kann, da $\rho_{xx'} \leq 1$ und $\rho_{yy'} \leq 1$ und mithin $\sqrt{\rho_{xx'}\rho_{yy'}} \leq 1$ gelten muß. Also folgt (3.4.10). □

Die Gleichung (3.4.9) zeigt, wie eine Beziehung zwischen nicht direkt beobachtbaren Größen anhand bekannter Größen berechnet werden kann. Die Gleichung gibt ebenfalls eine Beziehung zwischen der Gültigkeit von X bezüglich Y (und umgekehrt) und den Reliabilitäten $\rho_{xx'}$ und $\rho_{yy'}$ an:

Satz 3.4.3 *Die Gültigkeit ist stets kleiner, höchstens gleich der Wurzel aus der Reliabilität des Tests, also*

$$\rho_{xy} \leq \sqrt{\rho_{xx'}}. \tag{3.4.11}$$

Beweis: Aus (3.4.9) folgt

$$\rho_{xy} = \rho(\tau_x, \tau_y)\sqrt{\rho_{xx'}\rho_{yy'}}. \tag{3.4.12}$$

Nun gilt sicherlich $\rho_{xx'} \leq 1$ und $\rho_{yy'} \leq 1$, mithin folgt $\sqrt{\rho_{xx'}\rho_{yy'}} \leq 1$, so dass einerseits

$$\rho_{xy} \leq \rho(\tau_x, \tau_y), \tag{3.4.13}$$

und wegen $\rho(\tau_x, \tau_y) \leq 1$ folgt ebenso

$$\rho_{xy} \leq \sqrt{\rho_{xx'}\rho_{yy'}}, \tag{3.4.14}$$

d.h. die Validität ist höchstens gleich dem geometrischen Mittel der Reliabilitäten der X- und der Y-Werte. Wegen $\rho_{xx'} \leq 1$ und $\rho_{yy'} \leq 1$ folgt weiter

$$\sqrt{\rho_{xx'}\rho_{yy'}} \leq \sqrt{\rho_{xx'}},$$

so dass schließlich (3.4.11) folgt. □

Die bis hierher herleiteten Beziehungen zwischen Reliabilität und Validität erweisen sich nun als komplex; Lord & Novick (1968), p. 332 sprechen sogar von einem (Gültigkeits-) Paradox. Denn um die Gültigkeit zu erhöhen, wird man – als notwendige, wenn auch noch nicht hinreichende Bedingung – den oberen Wert $\sqrt{\rho_{xx'}}$, also die Wurzel aus der Reliabilität erhöhen. Der untere Wert für die Reliabilität ist aber Cronbachs α, und wie der Gleichung (3.2.47), Seite 160, zu entnehmen ist, wird α um so größer, je größer die Varianz $\mathbb{V}(X) = \sigma_x^2$ ist. In (3.2.57), Seite 162, wurde α in der Form

$$\alpha = \frac{1}{(q-1)/n + 1}, \quad q = \frac{\bar{\sigma}^2}{S}$$

angegeben, mit

$$S = \frac{1}{n(n-1)} \sum_{g \neq h} \rho_{gh}\sigma_g\sigma_h \quad \bar{\sigma}^2 = \frac{1}{n}\sum_g \sigma_g^2.$$

Für festen Wert von n wird α groß, wenn q klein ist, und q wird klein, wenn S relativ zu $\bar{\sigma}^2$ groß wird. Das ist der Fall, wenn die Korrelationen ρ_{gh} zwischen den Items groß sind. Dies korrespondiert zu (3.3.13), und (3.3.15) (Seite 181) bedeutet, dass die Trennschärfeindices möglichst groß sein sollen. Die Gleichung (3.4.1),

Seite 183, zeigt aber an, dass hohe Trennschärfen (relativ zu den Iteminterkorrelationen) die Gültigkeit eher reduzieren (man kann η durch Y ersetzen). Die Gleichung (3.4.3), Seite 184, impliziert aber, dass die ρ_{gh} eher klein sein sollten, damit die Gültigkeit groß wird, eine Forderung, die mit der Annahme kompatibel ist, dass ein Test um so gültiger sein wird, je mehr verschiedene Aspekte eines Merkmals erfaßt werden.

Eine praktische Methode, die Gültigkeit und die Reliabilität zu erhöhen, besteht darin, die ρ_{Xg}-Werte der Items gegen die korrespondierenden ρ_{gY}-Werte aufzutragen ("plotten"). Dies wird auf Grund der Beziehung (3.4.2) motiviert. Die Punktwolke sollte durch eine Regressionsgerade "erklärt" werden. Es läßt sich zeigen, dass die Steigung der Regressionsgeraden durch den Wert von $\rho_{X\eta}$, wie in (3.4.1) definiert, gegeben ist. Items, die zu weit von der Regressionsgeraden entfernt sind, insbesondere die am unteren und am oberen Ende der Punktwolke, sollten entfernt werden, – dadurch wird die Validität erhöht.

3.4.4 Arten von Validitäten.

Es werden verschiedene Arten von Validität diskutiert (Cronbach & Meehl (1955))

1. **Inhaltsvalidität:** Dieser Begriff entspricht am ehesten dem intuitiven Begriff von Gültigkeit: Inhaltsvalidität reflektiert das Ausmaß, mit dem ein Test das zu messende Merkmal in einem möglichst weiten Sinne erfasst und ist deshalb mit dem Begriff der Inhaltsvalidität ist deshalb mit der *logischen Validität* eng verwandt. Die Testitems sollen eine Stichprobe aus dem Universum aller möglicher Items, die das Merkmal erfassen, sein. Eine Quantifizierung ist oft schlecht möglich. Inhaltsvalidität soll deduktiv erfasst werden.

2. **Kriteriumsvalidität:** Gelegentlich lassen sich bestimmte Kriterien definieren, die das Merkmal charakterisieren. Die Testwerte können dann mit Kriteriumswerten korreliert werden. Die Kriterien sind z.B: Leistungen, die vorhergesagt werden, etwa ein erfolgreicher Studienabschluß, erfolgreiche feindliche Übernahmen konkurrierender Firmen, etc. Wird das Kriterium nach der Testgabe überprüft, wird auch von *prädiktiver Validität* gesprochen. Wird das Kriterium zur gleichen Zeit erhoben, zu der auch der Test gegeben wird, so ist die Rede von *kurrenter Validität* (*concurrent validity*)

3. **Konstruktvalidität:** Diese Form der Validität wird betrachtet, wenn es um den Test eines Merkmals geht, das zunächst nicht operational definiert ist. Der Testkonstrukteur muß dann fragen, welche "Konstrukte" – also letztlich auch wieder Merkmale – die unterschiedlichen Messwerte im Test erzeugen, also die Varianz der Testwerte. Der Konstrukteur muß dazu bestimmte Regeln – "Operationen" – festlegen, die das Merkmal eben als Konstrukt definieren.

Ein Beispiel ist der MMPI[46]. Das ursprüngliche Ziel war, zwischen bestimmten Gruppen von Patienten und "Normalen" zu unterscheiden. In der weiteren Forschung sollte dann versucht werden, die Persönlichkeit, die mit den verschiedenen Testwertprofilen einhergehen, zu charakterisieren mit dem Ziel, Verhaltensweisen vorhergesagen zu können, deren Charakteristika bisher noch nicht in empirischen Validierungsstudien untersucht werden konnten.

Ein weiterer Validitätsbegriff ist der der *faktoriellen Validität*. Diese bezieht sich auf die Korrelation der Testwerte mit einem Faktor, wie er in faktorenanalytischen Untersuchungen des Merkmals gewonnen wurde. Ein Beispiel ist der Generalfaktor der Intelligenz, wie er von Spearman 1904 vorgeschlagen wurde. Dieser Faktor soll eine gute Inhaltsvalidität haben und die Testwerte sollen gut mit ihm korrelieren, d.h. Tests sollen hoch auf diesem g-Faktor laden.

3.4.5 Validität bei Testverlängerung

Es ist gezeigt worden, wie sich die Reliabilität bei Verlängerung des Tests verbessert. Eine analoge Aussage kann für die Validität hergeleitet werden (Lord & Novick (1968), p. 114).

Satz 3.4.4 *Es seien X und Y verschiedene Tests (*generated by distinct processes*) mit Einheitslänge, und es seien $X(k)$ und $Y(l)$ die gleichen Tests, aber mit den Längen k bzw. l. Dann gilt*

$$\rho^2\left[\frac{X(k)}{k}, \frac{Y(l)}{l}\right] = \frac{kl\rho^2(X,Y)}{[1+(k-1)\rho_{xx'}][1+(l-1)\rho_{yy'}]}. \qquad (3.4.15)$$

$\rho^2[X(k)/k, Y(l)/l]$ ist die Gültigkeit (Validität) eines verlängerten oder verkürzten Tests in Bezug auf einen zweiten verlängerten oder verkürzten Test und ist eine streng monoton wachsende (fallende) Funktion von k und l.

Der Satz erlaubt es, die Gültigkeit eines Tests, der auf k Einheiten verlängert wird, in Bezug auf einen zweiten Test mit Einheitslänge ($l = 1$) zu berechnen. Dazu setzt man in (3.4.15) einfach $l = 1$ und erhält

$$\rho^2\left(\frac{X(k)}{k}, Y\right) = \frac{k\rho^2(X,Y)}{1+(k-1)\rho_{xx'}}. \qquad (3.4.16)$$

Um zu sehen, wie $\rho^2(X(k)/k, Y)$ sich für größer werdenden k-Wert verhält, teilt man Zähler und Nenner durch k und erhält

$$\rho^2\left(\frac{X(k)}{k}, Y\right) = \frac{\rho^2(X,Y)}{1/k+(1-1/k)\rho_{xx'}},$$

[46] Minnesota Multiphasic Personality Inventory

so dass

$$\lim_{k\to\infty} \rho^2\left(\frac{X(k)}{k}, Y\right) = \frac{\rho^2(X,Y)}{\rho_{xx'}}.$$

Aus (3.4.15) ergeben sich weitere *Attenuierungsformeln.* Es ist

$$\begin{aligned}\rho^2(\tau_x, Y) &= \lim_{k\to\infty} \rho^2\left(\frac{X(k)}{k}, k\right) \\ &= \lim_{k\to\infty} \frac{\rho^2(X,Y)}{1/k + (1-1/k)\rho_{xx'}} = \frac{\rho^2(X,Y)}{\rho_{xx'}}. \qquad (3.4.17)\end{aligned}$$

Analog dazu findet man

$$\rho^2(X, \tau_y) = \lim_{l\to\infty} \rho^2\left(X, \frac{Y(l)}{l}\right) = \frac{\rho^2(X,Y)}{\rho_{yy'}}. \qquad (3.4.18)$$

Schließlich findet man

$$\rho^2(\tau_x, \tau_y) = \lim_{\substack{k\to\infty \\ l\to\infty}} \rho^2\left(\frac{X(k)}{k}\frac{Y(l)}{l}\right) = \frac{\rho^2(X,Y)}{\rho_{xx'}\rho_{yy'}}, \qquad (3.4.19)$$

(verl. Gleichung (3.4.9)).

3.5 Die Abschätzung von Veränderungen (Gain)

Vielfach möchte man das Ausmaß der Veränderung einer Merkmalsausprägung bestimmen, – etwa in Therapieerfolgsstudien, Trainings- und anderen Interventionsmaßnahmen. Man hat dann Messungen X_1 vor der Maßnahme (die natürlich auch darin bestehen kann, dass man einfach Zeit vergehen läßt), und Messungen X_2 nach der Maßnahme. $\tau_1 = \mathbb{E}(X_1)$ und $\tau_2 = \mathbb{E}(X_2)$ seien die wahren Werte im Prätest bzw. im Posttest. Die Messungen X_1 und X_2 werden sich wegen der unvermeidlichen Messfehler in jedem Fall unterscheiden; man ist also an der Differenz $\Delta\tau = \tau_2 - \tau_1$ der wahren Werte interessiert. Diese "wahren" Differenzen $\tau_2 - \tau_1$ heißen in der englischsprachigen Literatur auch *gain*, was für Steigerung, Gewinn etc steht.

Die wahre Differenz $\tau_2 - \tau_1$ kann vom wahren Wert τ_1 im Prätest abhängen. Man ist deshalb an der Korrelation zwischen den $\tau_2 - \tau_1$- und den τ_1-Werten interessiert. Es sei $\sigma_1^2 = \mathbb{V}(X_1)$, $\sigma_2^2 = \mathbb{V}(X_2)$, $\rho_{12} = \rho(X_1, X_2)$, $\rho_{11'} = \rho(X_1, X_1')$, $\rho_{22'} = \rho(X_2, X_2')$. Dann gilt[47]

$$\rho(\tau_2 - \tau_1, \tau_1) = \frac{\sigma_2\rho_{12} - \sigma_1\rho_{11'}}{\sqrt{(\sigma_1^2\rho_{11'} + \sigma_2^2\rho_{22'} - \rho_{12}\sigma_1\sigma_2)\rho_{11'}}} \qquad (3.5.1)$$

[47]Die Formel, die in Lord & Novick (1968) auf Seite 73 gegeben wird, kann etwas vereinfacht werden und erweist sich dann als der Formel (3.5.1) äquivalent.

Anmerkung: $\rho(\tau_2-\tau_1,\tau_1)$ ist eine Korrelation zwischen nicht direkt beobachtbaren Größen, die aber durch beobachtbare Größen ausgedrückt werden kann. Die Korrelation $\rho(\tau_2-\tau_1,\tau_1)$ hängt nicht nur ab von der Korrelation $\rho_{12}=\rho(X_1,X_2)$ zwischen Prä- und Posttest und den Varianzen σ_1^2 und σ_2^2, sondern darüber hinaus von den Reliabilitäten $\rho_{11'}$ und $\rho_{22'}$. □

Beweis: Zur Erinnerung: Die Größen τ_1, τ_2 sind zwar wahre Werte und insofern konstante, aber sie sind auch zufällige Veränderliche, als sie zufällig mit den Personen variieren: testet man eine zufällig gewählte Person $a\in\mathcal{P}$, so nimmt τ_1, bzw. τ_2 den Wert T_{a1} bzw. T_{a2} an. Es sei $\mu_1=\mathbb{E}(\tau_1)$, $\mu_2=\mathbb{E}(\tau_2)$.

Es ist

$$\rho(\tau_2-\tau_1,\tau_1)=\frac{\mathbb{K}(\tau_2-\tau_1,\tau_1)}{\sigma(\tau_2-\tau_1)\sigma(\tau_1)}$$

$$\begin{aligned}\mathbb{K}(\tau_2-\tau_1,\tau_1) &= \mathbb{E}[(\tau_2-\tau_1)\tau_1]-\mathbb{E}(\tau_2-\tau_1)\mathbb{E}(\tau_1)\\ &= \mathbb{E}(\tau_1\tau_2)-\mathbb{E}(\tau_1^2)-\mathbb{E}(\tau_1\tau_2)+\mathbb{E}(\tau_1^2)\\ &= \mathbb{K}(\tau_1,\tau_2)-\mathbb{V}(\tau_1)\end{aligned}$$

Weiter ist nach (3.4.6), Seite 185,

$$\mathbb{K}(\tau_1,\tau_2)=\rho(X_1,X_2)\mathbb{V}(X_1)\mathbb{V}(X_2)=\rho_{12}\sigma_1\sigma_2$$

und nach (3.2.5), Seite 149

$$\mathbb{V}(\tau_1)=\mathbb{V}(X_1)\rho(X_1,X_1')=\sigma_1^2\rho_{11'},$$

so dass

$$\mathbb{K}(\tau_2-\tau_1,\tau_1)=\rho_{12}\sigma_1\sigma_2-\sigma_1^2 rho_{11'}=\sigma_1(\rho_{12}\sigma_2-\sigma_1 rho_{11'}).$$

Weiter ist

$$\mathbb{V}(\tau_2-\tau_1)=\mathbb{V}(\tau_1)+\mathbb{V}(\tau_2)-2\mathbb{K}(\tau_1,\tau_2)=\sigma_1^2\rho_{11'}+\sigma_2^2\rho_{22'}-\rho_{12}\sigma_1\sigma_2,$$

so dass

$$\begin{aligned}\sigma(\tau_2-\tau_1)\sigma(\tau_1) &= \sqrt{(\sigma_1^2\rho_{11'}+\sigma_2^2\rho_{22'}-\rho_{12}\sigma_1\sigma_2)\rho_{11'}\sigma_1^2}\\ &= \sigma_1\sqrt{(\sigma_1^2\rho_{11'}+\sigma_2^2\rho_{22'}-\rho_{12}\sigma_1\sigma_2)\rho_{11'}}\end{aligned}$$

und man erhält

$$\rho(\tau_2-\tau_1),\tau_1)=\frac{\rho_{12}\sigma_2-\sigma_1\rho_{11'}}{\sqrt{(\sigma_1^2\rho_{11'}+\sigma_2^2\rho_{22'}-\rho_{12}\sigma_1\sigma_2)\rho_{11'}}},$$

und dies ist (3.5.1). □

Die Vorhersage der wahren Veränderung: Es ist oft von Interesse, die wahre Veränderung – den Gain – anhand der gemessenen X_1- und X_2-Werte abzuschätzen. Generell gilt für eine beliebige Regressionsgleichung $y=bX+a+e=\hat{y}+e$.

Lord & Novick schreiben $R(y|x)$ für $\hat{y} = bx+a$. Setzt man nun mit Lord & Novick $G = \tau_2 - \tau_1$, so hängt die" Vorhersage" von G von zwei Variablen, X_1 und X_2, ab, man muß also eine multiple Regression mit den zwei Prädiktoren X_1 und X_2 betrachten. Man erhält die Gleichung

$$R(G|X_1, X_2) = B_1X_1 + B_2X_2 + C, \tag{3.5.2}$$

wobei B_1, B_2 und C noch zu bestimmende Regressionskoeffizienten sind. Es sei

$$\mu_1 = \mathbb{E}(X_1) = \mathbb{E}(\tau_1), \quad \mu_2 = \mathbb{E}(X_2) = \mathbb{E}(\tau_2).$$

Dann ist

$$\mathbb{E}(R(G|X_1, X_1)) = \mu_1 - \mu_2 = B_1\mu_1 + B_2\mu_2 + C$$

und

$$R(G|X_1, X_2) - (\mu_1 - \mu_2) = B_1(X_1 - \mu_1) + B_2(X_2 - \mu_2) + (\mu_1 - \mu_2).$$

Geht man zu standardisierten Werte über, erhält man

$$\tilde{R}(G|X_1, X_2) = \beta_1 Z_1 + \beta_2 Z_2, \tag{3.5.3}$$

mit[48]

$$\beta_1 = \frac{\rho_{G_1} - \rho_{G_2}\rho_{12}}{1 - \rho_{12}^2}, \quad \beta_2 = \frac{\rho_{G_2} - \rho_{G_1}\rho_{12}}{1 - \rho_{12}^2}, \tag{3.5.4}$$

wobei

$$\rho_{G_1} = \frac{\sigma_1\rho_{11'} - \sigma_2\rho_{12}}{\sigma_G}, \quad \rho_{G_2} = \frac{\sigma_2\rho_{22'} - \sigma_1\rho_{12}}{\sigma_G}. \tag{3.5.5}$$

3.6 Die Schätzung der Item- und Testparameter

In der KTT sind die Summen- und die Itemscores die Ausgangsgrößen der Testanalyse und der Charakterisierung der Items. Die Frage ist allerdings, ob z.B. der Summenscore in der Tat die gesamte Information über die Ausprägung des gemessenen Merkmals bei einer Person repräsentiert, – es könnte ja sein, dass es das Antwort*muster* einer Person ist, die relevante Information enthält, von der ein Teil bei der Bildung des Summenscores verloren geht. Das Antwortmuster der i-ten Person läßt sich durch einen Vektor

$$\vec{x}_v = \begin{pmatrix} x_{v1} \\ x_{v2} \\ \vdots \\ x_{vn} \end{pmatrix} \tag{3.6.1}$$

[48]Eine Herleitung der folgenden Ausdrücke findet man in Mortensen, U.: Multiple Regression und Korrelation, Katogoriale Regression, p. 13, http://www.uwe-mortensen.de/multregress1.pdf

darstellen. Die Komponenten des Vektors entsprechen den Werten x_{vj} in der a-ten Zeile der Datenstruktur. Zusätzlich zur Verteilung der Summen- und Itemscores kann man nun die Verteilung der *möglichen* Antwortvektoren betrachten. Die Menge der möglichen Antwortvektoren ist gleich der Menge der möglichen Verteilungen von Nullen und Einsen für eine Person. Für n Items existieren 2^n mögliche Vektoren. Diese Anzahl wächst sehr schnell mit dem Wert von n an: für $n = 1$ gibt es zwei Vektoren, für $n = 2$ hat man $2^2 = 4$ Vektoren, für $n = 3$ hat man $2^3 = 8$, für $n = 4$ erhält man $2^4 = 16$, und für $n = 5$ existieren $2^5 = 32$ Vektoren. Für $n = 10$ gibt es schon 1024 mögliche Vektoren bzw. Antwortmuster, und für $n = 20$ hat man $2^{20} = 1048576$ mögliche Antwortmuster. Es ist klar, dass man bei vielen Tests mit $n \geq 10$ Items so gut wie niemals Stichprobenumfänge haben wird, die es gestatten, eine Häufigkeitsverteilung über alle möglichen Antwortvektoren zu erstellen. Gleichwohl ist es möglich, Informationen aus Antwortvektoren durch Anwendung bestimmter Testmodelle zu gewinnen, indem man die Häufigkeiten der tatsächlich in der Stichprobe vorkommenden Antwortmuster betrachtet.

Die Betrachtung von Summenscores einerseits und Antwortmustern (-vektoren) andererseits legt eine Unterscheidung von Tests nahe:

1. *quantitativ* variierende latente – also i. A. nicht direkt beobachtbare – Persönlichkeitsmerkmale, und

2. *qualitative* Merkmale bzw. latente Variablen.

Summenscores, etwa in Leistungstests, reflektieren dann die quantitative Ausprägung eines Merkmals. Fokussiert man mehr auf die Antwortmuster, so zielt man damit eher auf qualitative Unterscheidungen von Personen. Eine dritte Klasse von Tests besteht aus einer Kombination von quantitativen und qualitativen Merkmalen.

3.7 Gruppenheterogenität und andere Faktoren

3.7.1 Gruppenheterogenität und Reliabilität

Die Reliabilität ist durch $\mathbb{V}(\tau)/\mathbb{V}(X) = \rho^2_{x\tau}$ definiert. Für $\rho_{x\tau} \neq 0$ wird man also eine Regression der Scores X auf die wahren Werte τ erwarten können: ein eher kleiner τ-Wert wird mit einem eher kleinen X-Wert einhergehen, ein größerer τ-Wert mit einem eher größeren X-Wert. Je geringer die Fehlervarianz σ^2_ε, desto enger werden die (X, τ)-Paare um die Regressionsgerade herum liegen (eine lineare Regression stillschweigend vorausgesetzt). Der Zusammenhang zwischen X und τ wird um so deutlicher erscheinen, je größer dass Intervall von τ-Werten ist. Beschränkt man den Bereich der τ-Werte auf einen relativ kleinen Bereich $\tau_0 \leq \tau \leq \tau_1$ (relativ zum Wert von σ^2_ε, so kann die Wolke der (X, τ)-Werte so aussehen, als gäbe einen Zusammenhang zwischen X und τ. In anderen Wor-

ten, die Probandengruppe, anhand der die Reliabilität abgeschätzt wird, sollte hinreichend heterogen bezüglich der τ-Werte sein.

Es werde zunächst der Begriff der Homoskedastizität formal definiert:

Definition 3.7.1 *Es seien $\mathcal{P}$ und $\mathcal{P}'$ zwei Populationen von Probanden mit $\mathcal{P}' \subset \mathcal{P}$. Die Messungen in der Population $\mathcal{P}$ sind* homoskedastisch*, wenn*

$$\mathbb{V}(\varepsilon|\tau) = \mathbb{V}(\varepsilon) = \sigma_\varepsilon^2 \tag{3.7.1}$$

gilt, d.h. wenn die Varianz des "Fehlers" für alle τ-Werte identisch ist.

Der Begriff der Homoskedastizität ist aus der Regressionstheorie bekannt. Die Formulierung der Definition 3.7.1 fokussiert auf die Forderung, dass $\mathbb{V}(\varepsilon) = \sigma_\varepsilon^2$ für alle τ-Werte den gleichen Wert haben soll. Man betrachte nun eine Population $\mathcal{P}'$, die aus $\mathcal{P}$ hervorgeht, indem man Personen mit τ-Werten aus einen bestimmten Bereich von τ-Werten zufällig aus $\mathcal{P}$ herausnimmt. Die Messungen aus $\mathcal{P}'$ werden mit $\tilde{X}$ bezeichnet. Gesucht ist die Reliabilität der $\tilde{X}$-Messungen. Es gilt

$$\rho_{\tilde{X}\tau}^2 = 1 - \frac{\sigma_X^2}{\sigma_{\tilde{X}}^2}(1 - \rho_{X\tau}^2). \tag{3.7.2}$$

Beweis: Wegen $X = \tau + \varepsilon$ gilt $\sigma_X^2 = \sigma_\tau^2 + \sigma_\varepsilon^2$, d.h. $\sigma_\varepsilon^2 = \sigma_X^2 - \sigma_\tau^2$. Wegen $\rho_{xx'} = \sigma_{x\tau}^2 = \mathbb{V}(\tau)/\mathbb{V}(X) = \sigma_\tau^2/\sigma_X^2$ folgt $\sigma_\tau^2 = \rho_{xx'}\sigma_X^2$. Mithin hat man

$$\sigma_\varepsilon^2 = \sigma_X^2 - \rho_{xx'}\sigma_X^2 = \sigma_X^2(1 - \rho_{xx'}) = \sigma_X^2(1 - \rho_{x\tau}^2).$$

Nach Voraussetzung ist aber $\mathbb{V}(\varepsilon|\tau) = \sigma_\varepsilon^2$ für alle τ und damit für alle Probanden aus $\mathcal{P}'$; d.h. aber

$$\sigma_\varepsilon^2 = \sigma_X^2(1 - \rho_{x\tau}^2) = \sigma_{\tilde{X}}^2(1 - \rho_{\tilde{x}\tau}^2).$$

Löst man diese Gleichung nach $\rho_{\tilde{x}\tau}^2$ auf, so erhält man (3.7.2). □

Die Gleichung (3.7.2) impliziert, dass einerseits $\rho_{\tilde{X}\tau}^2$ eine monoton wachsende Funktion des Quotienten $\sigma_X^2/\sigma_{\tilde{X}}^2$ ist, und dass andererseits $\rho_{\tilde{x}\tau}^2 = \rho_{x\tau}^2$ genau dann, wenn $\sigma_X^2/\sigma_{\tilde{X}}^2 = 1$, wenn also $\sigma_X^2 = \sigma_{\tilde{X}}^2$ gilt. Allgemein gilt für $\sigma_X^2 \neq \sigma_{\tilde{X}}^2$

$$\rho_{\tilde{x}\tau}^2 \begin{cases} \leq \rho_{xx'}, & \text{wenn } \sigma_{\tilde{x}}^2 \leq \sigma_x^2 \\ \\ \geq \rho_{xx'}, & \text{wenn } \sigma_{\tilde{x}}^2 \geq \sigma_x^2. \end{cases} \tag{3.7.3}$$

Dies bedeutet, dass die Einschränkung der Varianz der τ-Werte (durch Reduktion von $\mathcal{P}$ auf $\mathcal{P}'$) die Reliabilität des Tests *reduziert*. Man muß hier sehen, dass diese Reduktion nur eingeführt wurde, um die Abhängigkeit der Reliabilität von der Varianz der τ-Werte zu zeigen. Was in der Tat gezeigt wurde, dass für einen gegebenen Test die Reliabilität um so höher ist, je höher die Varianz der τ-Werte in der betrachteten Population ist. Allerdings geht in diese Betrachtungen die Annahme $\sigma_\varepsilon^2 = \sigma_{\tilde{\varepsilon}}^2$ ein. In empirischen Studien wird diese Annahme aber nicht notwendig bestätigt, vergl. Boldt (1966).

3.7.2 Speed- und Power-Komponenten

Es gibt zwei Weisen, ein Item eines Tests nicht korrekt zu beantworten: entweder man erzeugt eine inkorrekte Antwort, oder man versucht erst gar nicht, die Aufgabe zu lösen. Es sei I die Anzahl der inkorrekt oder gar nicht versuchten Lösungen, W sei die Anzahl der inkorrekten Antworten (W für 'wrong'), und U die Anzahl der gar nicht erst versuchten Lösungen. Dann gilt

$$I = W + U. \tag{3.7.4}$$

Es gibt nun zwei extreme Fälle:

1. **Power:** Die Probanden haben unbeschränkt Zeit, die Fragen (Aufgaben) zu beantworten, so dass jede Aufgabe zumindest versucht wird, so dass nicht nur $U = 0$, sondern auch der Erwartungswert gleich Null ist, $E(U) = 0$, und $\mathbb{V}(U) = 0$, da eben angenommen wird, dass zumindest der Versuch gemacht wird, jede Aufgabe zu lösen. Nur die Anzahl der inkorrekten Antworten kann ungleich Null sein, so dass $I = W$. Die Anzahl der gelösten Aufgaben hängt nur von der *power*, also der Fähigkeit der Probanden, ab.

2. **Speed:** Andererseits kann man Aufgaben betrachten, die im Prinzip von jedem Probanden gelöst werden können. Man kann aber davon ausgehen, dass die Geschwindigkeit, mit der eine Aufgabe beantwortet wird, für die einzelnen Probanden verschieden ist. Deswegen werden sich die Probanden nach einer bestimmten Zeit hinsichtlich der beantworteten Aufgaben oder Fragen voneinander unterscheiden. Es wird also nicht nur $W = 0$ sein, sondern grundsätzlich wird $\mathbb{E}(W) = 0$ und $\mathbb{V}(W) = 0$ sein, so dass $I = U$.

Reine Power- oder Speed-Tests werden eher die Ausnahme sein; im Allgemeinen sind Tests Mischungen aus Speed- und Power-Komponenten. Eine erste Diskussion der Mischung von Speed- und Power-Komponenten geht auf Gullicksen (1950) zurück. Ist E die Anzahl der Fehler insgesamt, so hat man $E = W + U$ und

$$\mathbb{V}(E) = \mathbb{V}(W) + \mathbb{V}(U) + 2\rho(U, W)\sqrt{\mathbb{V}(U)\mathbb{V}(W)}. \tag{3.7.5}$$

Für einen reinen Power-Test ist $\mathbb{V}(E) = \mathbb{V}(W)$, für einen reinen Speed-Test ist $\mathbb{V}(E) = \mathbb{V}(U)$. Für Mischungen kann man die Quotienten

$$\frac{\mathbb{V}(W)}{\mathbb{V}(E)} \text{ bzw. } \frac{\mathbb{V}(U)}{\mathbb{V}(E)} \tag{3.7.6}$$

betrachten. Ist $\mathbb{V}(W)/\mathbb{V}(E)$ groß (nahe bei 1), hat man eher einen Power-Test (und dann ist $\mathbb{V}(U)/\mathbb{V}(E)$ klein), und umgekehrt hat man eher einen Speed-Test, wenn $\mathbb{V}(U)/\mathbb{V}(E)$ groß und $\mathbb{V}(W)/\mathbb{V}(E)$ klein ist. Rindler (1979) hat aber darauf hingewiesen, dass es bei dieser Interpretation Probleme geben kann, denn die Korrelationen zwischen U und W werden nicht notwendig adäquat in Rechnung

gestellt, – wer versucht, alle Aufgaben zu beantworten, macht u. U. mehr Fehler als jemand, der dies nicht versucht.

Stafford (1971) schlug einen *Speededness-Quotienten* vor, um die Speed-Power-Anteile zu schätzen. Es sei, für den i-ten Probanden, U_i die Anzahl der Aufgaben, bei denen gar nicht erst versucht wurde, sie zu lösen. O_i sei die Anzahl derAufgaben, bei denen die Lösung versucht wurde, bei denen aber der Versuch nicht zuende gebracht wurde, und W_i sei die Anzahl der Aufgaben, für die eine falsche Antwort gegeben wurde. Für die i-te Person wird der Quotient

$$SQ_i = \frac{U_i}{W_i + O_i + U_i} \tag{3.7.7}$$

eingeführt. Für den gesamten Test wird der Quotient

$$SQ = \frac{\sum_i U_i}{\sum_i W_i + \sum_i O_i + \sum_i U_i} \tag{3.7.8}$$

definiert (es wird also *nicht* die Summe $\sum_i SQ_i$ gebildet!). Analog zu diesem Quotienten kann ein Speededness-Quotient für ein einzelnes, etwa das k-te, Item eingeführt werden:

$$SQ_k = \frac{U_k}{W_k + O_k + U_k}. \tag{3.7.9}$$

Summiert man analog zu (3.7.8) über die Items, erhält man wieder den Quotienten SQ. Für einen reinen Speed-Test gilt dann wieder $\sum_i W_i = \sum_i O_i = 0$, und für einen reinen Power-Test hat man $\sum_i U_i = 0$ und damit $SQ = 0$. Stafford äußert seine Überzeugung, dass der SQ-Wert für verschiedene Probandengruppen bestimmt werden sollte, um den Testern anzuzeigen, ob der Test für eine gegebene Gruppe geeignet ist, – der Betrag an *Speededness* könnte ja mit den pädagogischen Zielen interferieren.

3.7.3 Messbedingungen und Reliabilität

Parallele Tests: Die Abschätzung der Reliabilität durch Berechnung von $\rho_{xx'}$ basiert auf der Annahme, dass die Messungen parallel sind, d.h. dass die τ-Werte bei Messwiederholungen konstant bleiben. In der Praxis muß diese Bedingung nicht notwendig erfüllt sein, da Müdigkeit, psychische Sättigung, Gedächtnis- und Praxiseffekte eine eine effektive Veränderung der τ-Werte bedeuten können. Für tatsächlich parallele Messungen hat man

$$\rho^2_{x\tau} = \frac{\mathbb{V}(\tau)}{\mathbb{V}(X)} = 1 - \frac{\sigma^2_\varepsilon}{\sigma^2_x}, \tag{3.7.10}$$

der deshalb auch *Präzisionskoeffizient* (*coefficient of precision*, Lord & Novick (1968), p. 134) heißt. Dieser Koeffizient zeigt an, in welchem Ausmaß die Reliabilität durch den Messfehler σ^2_ε beeinträchtigt wird. In der Praxis werden im

Wesentlichen drei Methoden zur Bestimmung der Reliabilität verwendet: (i) die Retestmethode, (ii) die interne Analyse, und die (iii) Paralleltestmethode. Bei der Retestmethode wird der Test bei jedem Probanden zweimal appliziert. Das Problem bei dieser Methode sind Gedächtnis-, Übungs-, Müdigkeitseffekte, die systematisch wirken und daher in die τ-Werte eingehen. Wählt man das Zeitintervall zwischen den Messungen so groß, dass Gedächtnis- und Übungseffekte vernachlässigbar werden, kann es trotzdem zu systematischen Veränderungen der τ-Werte gekommen sein. In diesem Fall kommt es zu Unterschätzungen der Reliabilität. Sind Gedächtnis- und Übungseffekte vorhanden, kann es zu Überschätzungen der Reliabilität kommen.

Split-Half-Methode: Bei der Split-Half-Methode werden diese Effekte vermieden, aber die Gesamtreliabilität muß über eine Spearman-Brown-Formel geschätzt werden (*stepped-up reliability*). Es kann aber gezeigt werden, dass im Falle nicht paralleler Testhälften die Reliabilität unterschätzt wird. Lord & Novick (p.135) empfehlen hier, Cronbachs α als untere Schätzung der Reliabilität zu berechnen. Die Probleme mit Cronbachs α sind aber bereits diskutiert worden. Ein anderes Problem, das mit der Split-half-Methode verknüpft ist, ergibt sich aus der Notwendigkeit, den Test in zwei Hälften zu teilen. Denn oft existiert einerseits eine Zeitgrenze für das Lösen der Aufgaben, andererseits sind die Aufgaben der Schwierigkeit nach geordnet. Viele Probanden, die die erste Hälfte der Aufgaben lösen, scheitern an der Aufgaben der zweiten Hälfte. Dadurch wird die erste Hälfte ein Power-Test, während die zweite ein Speed-Test wird. In einem solchen Fall sollte man nicht die ersten $n/2$ Aufgaben als ersten Test und die zweiten $n/2$ als zweiten Test verwenden.

Die Items seien homogen. Man kann nun so vorgehen, dass man den Test nach Maßgabe der Zeit, die für das Lösen der Aufgaben benötigt wird, aufteilt. Man betrachte einen Speed-Test, der insgesamt eine Stunde dauere und der homogen wäre, würden keine zeitlichen Beschränkungen eingeführt. Der Test kann nun in vier verschieden zeitlich begrenzte Teile aufgeteilt werden, jeder von 15 Minuten Dauer. Diese vier Teiltests werden als Komponenten betrachtet, die in sich jeweils homogen sind. Deswegen sind nun Split-Halves, Drittel etc in sich homogen und untereinander parallel, – vorausgesetzt, die Items sind insgesamt homogen und Müdigkeits- und Praxiseffekte können vernachlässigt werden.

Nummeriert man die Aufgaben durch und teilt die Items nach dem Gerade-ungerade-Prinzip auf (die Items 1, 3, 5, etc kommen in die eine Hälfte, die Items 2, 4, 6, etc in die andere), so erhält man im Allgemeinen eine gute Aufteilung. Allerdings können Probleme auftreten, wenn der Test ein Speeded-Test oder teilweise Speeded-Test ist. Jeder Proband, der die letzten $2r$ Items nicht vollständig beantworten kann, wird einen Score gleich Null für diese Items haben und die entsprechenden odd- und even-Scores werden perfekt miteinander korrelieren. Werden diese Items nun der Gesamtmenge der beantworteten Items beigefügt, kann sich eine Reliabilitätschätzung ergeben, die von der tatsächlichen Reliabili-

tät erheblich abweicht.

Für die Methode der parallelen Tests müssen zwei Parallelformen des Tests entwickelt werden. Kleine Variationen in der Testadministration sowie Fluktuationen des gemessenen Merkmals innerhalb der Probanden können sich auf die Korrelation zwischen den Testformen auswirken. Die resultierende Reliabilitätsschätzung wird dann auch als *Stabilitätschätzung* bezeichnet. Der Stabilitätskoeffizient wird im Allgemeinen kleiner als die Größe (3.7.10) ausfallen, ist also eine untere Schätzung der Reliabiltät.

3.7.4 Attenuationskorrekturen

In Satz 3.4.2, Seite 186, wurde die Attenuierungsformel

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

mit $\rho(\tau_x, \tau_y) \geq \rho_{xy}$ angegeben. Hierin ist $\rho_{xx'}$ die Reliabilität des Tests mit den X-Scores, und $\rho_{yy'}$ ist die Reliabilität des Tests mit den Y-Scores. Auf Seite 190 wurde die Gleichung

$$\rho^2(\tau_x, \tau_y) = \lim_{\substack{k\to\infty \\ l\to\infty}} \rho^2\left(\frac{X(k)}{k}\frac{Y(l)}{l}\right) = \frac{\rho^2(X,Y)}{\rho_{xx'}\rho_{yy'}},$$

angegeben. Die rechte Seite von $\rho(\tau_x, \tau_y) = \rho_{xy}/\sqrt{\rho_{xx'}\rho_{yy'}}$ kann also als Grenzwert der Korrelation $\rho(X(k)/k, Y(l)/l)$ angesehen werden. Eine Möglichkeit, $\rho(\tau_x, \tau_y)$ abzuschätzen, besteht darin, für eine hinreichend große Stichprobe ρ_{xy}, $\rho_{xx'}$ und $\rho_{yy'}$ abzuschätzen und daraus wiederum $\rho(\tau_x, \tau_y)$ zu schätzen. Die Praxis hat gezeigt, dass für diese Schätzungen $\hat{\rho}(\tau_x, \tau_y) > 1$ resultieren kann, was nicht sein darf. Die Ursache hierfür ist, dass $\rho_{xx'}$ und $\rho_{yy'}$ gelegentlich unterschätzt werden, was insbesondere dann geschieht, wenn die Schätzungen nur auf *einer* Testdarbietung beruhen. Die Unterschätzungen resultieren daraus, dass die Tests zur Bestimmung von $\rho_{xx'}$ und $\rho_{yy'}$ wiederholt dargeboten werden müssen und dabei Fehler in die Messungen eingehen (die üblichen Faktoren, die bei Testwiederholungen störend auftreten), die bei der Berechnung von rho_{xy} nicht auftreten. Berechnet man die Reliablitäten anhand einer Analyse der internen Konsistenz, so resultiert ein Mangel an Homogenität ebenfalls in einer Unterschätzung der Reliabilitäten. Dies verweist auf die Notwendigkeit, die Reliabilität so genau wie nur irgend möglich zu schätzen.

3.8 Allgemeine Kommentare zur Klassischen Testtheorie

3.8.1 Zur Grundannahme

Die Grundannahme der KTT ist, dass ein Testscore X additiv zerlegt werden kann in einen wahren Wert τ und einen Fehler ε, also $X = \tau + \varepsilon$, wobei den

Axiomen der KTT zufolge τ und ε stochastisch unabhängig voneinander sind; es soll also

$$X = \tau + \varepsilon$$

gelten. $\tau = \mathbb{E}(X)$ ist einfach der Erwartungswert der zufälligen Veränderlichen X.

Wie Lumsden (1976) ausführt, geht die Idee, Fehler, die unabhängig vom "wahren" Wert einzuführen, auf den Statistiker Undy Yule[49] zurück, der sie Spearman[50] nahelegte (Spearman, 1910). Yule führte die Fehler als *irrelevant factors* ein, die *innumerable and ubiquitous* seien; sie ergeben sich aus der Tatsache, dass

> "the scientific conquest of nature is essentially achieved by artificially simplifying her process; the factor that we happen to be investigating is allowed to vary, while the remaining factors are kept as constant as possible, there effect being regarded as irrelevant for the purpose in hand." (Spearman (1910), p. 278)

Das ist die Idee eines Messfehlers, wie sie allgemein dem Allgemeinen Linearen Modell zugrunde liegt. Sie wirkt plausibel, ist wohl deswegen zu einer der Basisannahmen der KTT avanciert und vereinfacht vor allem viele Formeln. Die Frage ist allerdings, ob sie bei vielen Anwendungen der KTT auch gerechtfertigt ist. Denn die Messwerte sind ja i. a. nicht zu vergleichen mit Messungen, wie etwa Längen- oder Gewichtsmessungen oder Messungen, wie man sie etwa in physiologischen Untersuchungen gewinnt. Wie Lord & Novick (1968, p. 97) anmerken, beziehen sich die meisten Anwendungen der KTT auf dichotome Items; "Messungen" nehmen dann nur die Werte 1 oder 0 an, oder man hat Werte auf Ratingskalen, etwa - 1, 0, + 1, oder 1 bis 7. Die Unabhängigkeit des Fehlers ε vom wahren Wert τ ist bei solchen Messwerten nicht mehr gegeben. Dies ist leicht einzusehen: hat eine Person einen wahren Wert τ etwa nahe bei -1 oder +1, so können Fehler keinen Wert kleiner als -1 bzw. +1 erzeugen, es gibt nur Abweichungen Richtung 0 oder +1 bzw. 0 und -1. Damit sind die Fehler vom wahren Wert abhängig. Handelt es sich im Test um Multiple-Choice-Items, so sind wahre Werte und Fehler negativ korreliert (Bock & Wood (1971), dort weitere Literaturangaben).

Lumsden argumentiert, dass ein wesentlicher Grund für Lord & Novick, die Annahme $X = \tau + \varepsilon$ nachgerade kritiklos zu akzeptieren, die Möglichkeit ist, die Reliabilität über eine Regressionsgleichung zu definieren. Samajima (1977a) formuliert in diesem Zusammenhang eine herbe Kritik an der KTT: die Reliabilitätsschätzung hängt von der speziellen Gruppe von Testpersonen ab, deren

[49] George Undy Yule (1871 – 1951), Britischer Statistiker; nannte sich gewöhnlich nur Undy Yule

[50] Charles Spearman (1863-1945), britischer Psychologe, der bei W. Wundt in Leipzig und später beiOswald Külpe in Würzburg studierte; später wurde er Professor für Psychologie am University College in London. Spearman hatte wesentlichen Anteil an der Entwicklung der Testtheorie.

Werte der Reliabilitätsschätzung zugrunde liegen. Das gleiche gilt für den Standardschätzfehler. Samejima schreibt:

> "Thus it is clear that the reliability coefficient in classical test theory is at the mercy of the heterogeneity of the group of examinees, which has nothing to do with the test itself. We can easily make an erroneous test look good by using a heterogeneous group of subjects. ... A fatal deficiency of classical test theory is that it cannot specify the standard error of estimation independently from the reliability coefficient, and thus independently from a specific group of examinees." (Samejima (1977a), Abschnitt 4).

Die Vorhersage $\hat{\tau}$ eines wahren Werts geschieht in der KTT allgemein über die Gleichung (3.2.109), Seite 175, also

$$\hat{\tau} = \rho_{xx'} X + (1 - \rho_{xx'})\mathbb{E}(X).$$

Ebenso wie $X = \tau + \varepsilon$ eine lineare Regressionsgleichung ist, ist dieser Ausdruck für $\hat{\tau}$ eine lineare Gleichung. Die Frage ist allerdings, ob diese Gleichung immer gilt; gerade bei dichotomen Items können sich nämlich auch nichtlineare Beziehungen ergeben. Es soll deshalb auf den wichtigen Spezialfall dichotomer Items gesondert eingegangen werden.

3.8.2 Dichotome Items

Es sei I_g ein dichotomes Item und X_g sei der mögliche Score; die zusätzliche Indizierung mit a für eine bestimmte Person $a \in \mathcal{P}$ kann im Moment noch vernachlässigt werden, um die Ausdrücke einfach zu halten. Es gelte insbesondere $X_g = \{0, 1\}$, d.h. X_g soll nur die Werte 0 oder 1 annehmen können. Insbesondere soll $X_g = 1$ gesetzt werden, wenn das Item "richtig" beantwortet wurde, und $X_g = 0$, wenn es "falsch" beantwortet wurde. Der Score entspricht also dem Wert einer Indikatorfunktion für dieses Item. Es werde die Itemfunktion für ein solches Item und ihre Beziehung zur Definition des "wahren Wertes" τ_g betrachtet. Für den Fall, dass X_g eine Indikatorvariable ist, gilt mit $p_g = P(X_g = 1)$

$$\mathbb{E}(X_g) = 1 \cdot p_g + 0 \cdot (1 - p_g) = p_g = \tau_g. \tag{3.8.1}$$

Der wahre Wert ist jetzt also gleich der Wahrscheinlichkeit, mit der das Item "positiv" beantwortet oder gelöst wird; diese Wahrscheinlichkeit kann natürlich von Person zu Person verschieden sein, d.h. p_g und damit τ_g ist eine Funktion der Ausprägung θ des gemessenen Merkmals, so dass man $\tau_g(\theta)$ schreiben kann. Nach Definition 2.6.1, Seite 31, ist diese Funktion aber gerade durch die Itemfunktion $F_g(\theta)$ gegeben, so dass man also die Beziehung

$$\tau_g(\theta) = F_g(\theta) \tag{3.8.2}$$

hat.

Es ergibt sich nun die Frage nach den möglichen Werten von ε_g. Da X_g nur die Werte 0 oder 1 annehmen kann, muß offenbar

$$\varepsilon_g = X_g - \tau_g = \begin{cases} -\tau_g, & \text{wenn } X_g = 0 \\ 1 - \tau_g, & \text{wenn } X_g = 1 \end{cases} \tag{3.8.3}$$

gelten, denn wenn $\varepsilon_g = -\tau_g$, folgt $X_g = \tau_g - \tau_g = 0$, und wenn $\varepsilon_g = 1 - \tau_g$, folgt $X_g = \tau_g + 1 - \tau_g = 1$. Wegen (3.8.2) werden die Werte der Fehlervariablen also durch die Wahrscheinlichkeit, mit der das Item beantwortet wird, definiert. Ganz sicher ist ε_g nicht normalverteilt. Vielmehr hat ε_g offenbar die Verteilung

$$P(\varepsilon_g = -\tau_g) = 1 - \tau_g, \quad P(\varepsilon_g = 1 - \tau_g) = \tau_g, \quad \tau_g = \tau_g(\theta).$$

Durch (3.8.3) wird schon deutlich, dass die Unabhängigkeit von ε_g von τ_g nicht mehr gegeben ist, da die Werte von ε_g ja Funktionen von τ_g sind, und da überdies $\tau_g = \tau_g(\theta) = F_g(\theta)$, muß $\mathbb{V}(\varepsilon_g)$ ebenfalls eine Funktion von θ sein, d.h. bei dichotomen Items variiert die Varianz der Fehler von Person zu Person, sofern sich die Personen hinsichtlich ihres θ-Wertes unterscheiden. Andererseits war in (3.1.3), Seite 141, die Behauptung $\mathbb{K}(\varepsilon_{*g}, \tau_{*g}) = 0$ aufgestellt worden. Beim Beweis dieser Aussage wurden die Gleichungen (**??**) und (**??**), Seite **??**, aufgestellt, und wegen $\mathbb{E}(\varepsilon_{*g}|\tau_{*g}) = 0$ folgt dann die Behauptung, dass ε_{*g} und τ_{*g} unabhängig voneinander sind. In der Tat folgt auch hier, dass die bedingte Erwartung des Fehlers gleich Null ist, denn man hat

$$\mathbb{E}(\varepsilon_g|\tau_g) = -\tau_g(1 - \tau_g) + \tau_g(1 - \tau_g) = 0,$$

und damit $\mathbb{K}(\varepsilon_g, \tau_g) = 0$, folgt man dem Ansatz Holland et al.s auf Seite **??**. Der Hintergrund für diese Konfusion ist natürlich, dass für dichotome Items und damit für Scores $X = \{0, 1\}$ die Messfehlertheorie nicht mehr sinnvoll angewendet werden kann, weil ein von τ_g unabhängiger Fehler ε_g im Sinne dieser Theorie eben gar nicht existiert.

Bevor weitere Folgerungen bezüglich des Ansatzes der KTT für den Fall dichotomer Items gezogen werden, sollen noch die Itemfunktionen $F_g(\theta)$ für dichtome Items betrachtet werden. Nach (3.8.1) entspricht dem der Plot von $\tau_g(\theta)$ versus θ. Beliebt ist die Annahme eines *kongenerischen* Tests, wie sie von Jöreskog (1970, 1971) eingeführt wurde (in Abschnitt 3.2.2 wird weiter darauf eingegangen). Für kongenerische Items gilt

$$\tau_g = \nu_g + \lambda_g \theta, \tag{3.8.4}$$

d.h. τ_g ist eine lineare[51] Funktion der Merkmalsausprägung θ. Die Werte für ν_g und λ_g ergeben sich durch die Konstruktion der Messvorrichtung bzw. des Tests. Da $\tau_g = F_g(\theta)$, ist die Itemfunktion durch

$$F_g(\theta) = \tau_g = \nu_g + \lambda_g \theta \tag{3.8.5}$$

[51] eigentlich: *affine* Funktion

gegeben. Dies ist der in (2.6.3), Seite 32, betrachtete Fall linearer Itemfunktionen. Offenbar kann τ_g nun nur zwischen $0 = \nu_g + \lambda_g\theta_0$ und $1 = \nu_g + \lambda_g\theta_1$ variieren; für $\theta < \theta_0$ würden sich negative τ_g-Werte und damit negative Wahrscheinlichkeiten ergeben, also muß $F_g(\theta) = \tau_g = 0$ für $\theta < \theta_0$ gesetzt werden. Da es keine Wahrscheinlichkeiten größer als 1 gibt, wird $F_g(\theta) = \tau_g = 1$ für $\theta > \theta_1$ gesetzt. Für $\theta < \theta_0$ kann das Item nie gelöst werden, für $\theta > \theta_1$ wird es stets gelöst. Man kann nun Indices wie Trennschärfe und Schwierigkeit in Abhängigkeit von den Parametern ν und λ diskutieren, aber über den Nutzen algebraischer Fingerübungen hinaus ergeben solche Betrachtungen keinen Sinn, es sei denn eine *psychologische Theorie* impliziert lineare Itemfunktionen und man mochte diese Theorie testen. Nur existiert eine solche Theorie nicht. Bekanntlich wird der Parameter θ bei vielen Fragestellungen wiederum als lineare Funktion weiterer Prädiktoren X_j konzipiert, etwa $\theta = \sum_j b_j X_j$, die experimentelle Bedingungen, sozio-ökonomische Variablen etc repräsentieren und deren (Regressions-) Gewichte b_j geschätzt werden müssen. Bei einer Beschränkung der abhängigen Variablen θ und damit $p_g = \tau_g(\theta)$ auf ein beschränktes Intervall derart, dass (3.8.5) gilt, kommt es in der Regel zu Verschätzungen der Parameterwerte, die den Sinn eines linearen Ansatzes von vornherein in Frage stellen.

Dichotome Items implizieren aber nicht notwendig lineare Itemfunktionen, denn dass τ und θ durch eine lineare Beziehung, $\tau = \mu + \lambda\theta$, miteinander verbunden sind, ist keine Annahme, die zum Kern der KTT gehört, sondern die sich nur als nützlich im Zusammenhang mit der Diskussion der Reliabilität von Tests erweist. Es ist keinesfalls *a priori* klar, wie die Beziehung zwischen der Merkmalsausprägung θ und dem "wahren Wert" τ_g und damit der Wahrscheinlichkeit $F_g(\theta)$ aussieht, denn die Beziehung zwischen τ_g und θ hängt davon ab, in welcher Weise das effektiv gemessene Merkmal das eigentlich interessierende, durch θ repräsentierte Merkmal erfasst, bzw. in welcher Weise die Messapparatur, die ja nicht notwendig aus Items im Sinne von Fragen bestehen muß, auf verschiedene Ausprägungen von θ reagiert. Man denke etwa an die Messung von psychologischem Stress (θ) durch Messung der Konzentration bestimmter Hormone im Speichel (X bzw. τ). Denkbar wäre z.B. die Beziehung

$$\tau_g(\theta) = 1 - \exp\left(-\nu_g\theta^{\lambda_g}\right), \quad \theta \geq 0,\ \nu_g > 0,\ \lambda_g > 0. \tag{3.8.6}$$

Dies ist die Weibull-Funktion, die einen S-förmigen Verlauf mit $0 \leq \tau_g(\theta) \leq 1$ hat (vergl. Abbildung 2, Seite 32). θ repräsentiert dabei das "wahre" Ausmaß an Stress, und $\tau_g(\theta)$ den "wahren" Anteil des Hormons am Speichel. Je größer der Wert von ν_g, desto weiter links auf der θ-Skala liegt die Funktion, d.h. ν_g hat den Effekt eines Leichtigkeitsparameters, und je größer λ_g, desto steiler ist die Funktion, so dass man λ_g als Trennschärfeparameter betrachten kann. Die Weibull-Funktion wird oft in Zusammenhang mit der Reliabilität (Bruch- und Reissfestigkeit, $X = 1$ wenn das Material bricht) von Materialien diskutiert, und in der Psychophysik wird sie oft zur Definition psychometrischer Funktionen herangezogen, wenn also die Wahrscheinlichkeit des Entdeckens von Stimuli oder

von Unterschieden zwischen Stimuli modelliert werden soll ($X = 1$, wenn der Stimulus bzw. die Stimulusdifferenz entdeckt wird). Die Frage, ob ein Stimulus präsentiert wurde oder nicht, repräsentiert ein dichotomes Item. Aber auch der eben als Beispiel genannte Stress kann über dichotomes Item abgefragt werden: ist bei einer gegebenen Person die Konzentration eines bestimmten Hormons größer als ein bestimmter, vorher festgelegter Wert, so wird $X_g = 1$ notiert, anderfalls $X_g = 0$. Der Erwartungswert für die Person ist $\tau_g(\theta)$, und $\tau_g(\theta)$ könnte durch (3.8.6) gegeben sein. Das Problem, dass man allerdings nach wie vor hat, wenn man beim Postulat $X_g = \tau_g(\theta) + \varepsilon_g$ der KTT bleibt, ist die Interpretation von ε_g, denn man wird sofort auf (3.8.3) zurückgeführt. Dies ist der Preis, den man zahlt, wenn beim Ansatz des ALM bleibt.

Es wird nun deutlich, dass es im Grunde gar keinen Sinn macht, sich bei dichotomen Items an das ALM und damit an die Annahme $X_g = \tau_g(\theta) + \varepsilon_g$ zu halten. Wenn $\tau_g(\theta)$ schon die Wahrscheinlichkeit ist, dass $X_g = 1$, so ist die Einführung einer weiteren zufälligen Veränderlichen ε_g redundant, sie trägt zur Erklärung der Stochastizität von X_g nichts bei, was sich eben darin äußert, dass ε_g gemäß (3.8.3) durch $\tau_g(\theta)$ definiert wird. Es genügt völlig, sich auf die Wahrscheinlichkeit $\tau_g(\theta)$ zu beschränken. Damit hat man automatisch die Klasse der Modelle des ALM verlassen und zu einer allgemeineren Klasse, der der *Generalisierten Linearen Modelle* (GLM, Nelder & Wedderburn (1972)), gewechselt. In Abschnitt 4 wird diese Klasse kurz vorgestellt.

Man kann fragen, wieso dichotome Items überhaupt im Rahmen der KTT und damit im Rahmen des Allgemeinen Linearen Modells der Statistik diskutiert wurden, da die vorangegangene Betrachtung ja direkt auf das GLM als allgemeinem Rahmen für diese Art Items führt. Dies mag daran liegen, dass die Definition (3.8.3) des Fehlers entfällt bzw. weniger deutlich wird, wenn X den Summenscore einer Person repräsentiert. Dann ist ja, für eine beliebige Person, $X = \sum_g X_g$. Sind alle Items dichotom, so ist X binomialverteilt, *falls* alle Items für die jeweils betrachtete Person gleich schwierig sind, falls also $p_g = p$ gilt. Für eine hinreichend große Anzahl n von Items nach dem Satz von deMoivre-Laplace approximativ normalverteilt, so dass

$$P(X = k) \approx \frac{1}{\sqrt{np(1-p)\pi}} \exp\left(-\frac{(k-np)^2}{2np(1-p)}\right). \tag{3.8.7}$$

Da $\tau = \mathbb{E}(X) = np$ haben die "Fehler" die Werte $\varepsilon = k - np$, für $k = 0, 1, \ldots, n$ und die Varianz $np(1-p)$, und diese Charakterisierung von ε wirkt weitaus weniger bizarr als die Charakterisierung (3.8.3). Im Falle ungleicher p_g-Werte ist X verallgemeinert binomialverteilt (Kendall & Stuart, 1969) bzw. es hilft wieder der zentrale Grenzwertsatz, demzufolge $X_a \sim N(\mu_a, \sigma_a^2)$ mit $\mu_a = \sum_g \mathbb{E}(X_{ag}) = \sum_g p_{ag}$, $\sigma_a^2 = \sum_g p_{ag} q_{ag}$ ist, und die Betrachtungen zu ε für den Fall gleicher p_g-Werte übertragen sich auf diesen Fall: $\varepsilon = k - \sum_g p_g$. Vielleicht sind diese Charakterisierungen der Fehler der Grund, wieso die KTT überhaupt auf dichotome Items angewendet wurde, zumal der Begriff der Itemfunktion, die ja die Beziehung zwi-

schen der Wahrscheinlichkeit einer Antwort auf ein bestimmtes Item in Abhängigkeit von der Ausprägung θ des Merkmals beschreibt, lange keine explizite Rolle in der KTT gespielt hat. Der Trennschärfeindex eines Items, der hier mit einem Parameter der Itemfunktion identifiziert wird, wurde einfach als punkt-biseriale Korrelation zwischen der Itembeantwortung und dem Gesamt-Score definiert, und der Widerspruch zum "Axiom" $\mathbb{K}(\varepsilon, \tau) = 0$ blieb verschleiert.

An einigen Stellen ist die Argumentation von Lord & Novick (1968) außerdentlich unklar. In Kapitel 11, in dem Methoden des *item sampling* diskutiert werden, wird die Häufigkeitsverteilung der Messfehler für binäre Items betrachtet. Es wird davon ausgegangen, dass ein Proband zufällig aus der Menge der möglichen Probanden gewählt wird. Der nun beobachtete (Summen-)Score x_a könne nun als die Anzahl der Erfolge in n voneinander unabhängigen Versuchen gewertet werden, – jeder Versuch entspricht der Bearbeitung einer der n Aufgaben des Tests. Unvermittelt kommt der Satz (Kursivsetzung von den Autoren): *Thus in repeated testing with different random samples of n items each, the frequency distribution of the observed score of examinee a will be*

$$f(x_a) = \binom{n}{x_a} \zeta_a^{x_a} (1 - \zeta_a)^{n - x_a}. \tag{3.8.8}$$

This is an ordinary binomial distribution with probability of success ζ_a (Lord & Novick (1968/2008), p. 250). Der Messfehler sei nun definiert gemäß

$$\eta_a \equiv x_a - n\zeta_a,$$

und die Häufigkeitsverteilung von η_a sei nun die gleiche wie die von x_a, bis auf eine Verschiebnung des "Ursprungs" (origin). Die Frage, die sich hier ergibt und die von Lord & Novick nicht beantwortet wird, ist die nach der Bedeutung von ζ_a: offenbar ist ζ_a eine Wahrscheinlichkeit, und wenn diese für die Person $a \in \mathcal{P}$ charakteristisch ist, so wird implizit angenommen, dass alle Aufgaben gleich schwierig sind. Die sich unnmittelbar daraus ergebende Frage ist, ob diese Annahme generell für die KTT gelten soll. Sicherlich nicht, denn sonst würden ja nicht verschiedene Schwierigkeitsindices π_g für die Items betrachtet werden müssen. Die Fähigkeit der Person a und die Schwierigkeit der Items werden bei dieser Charakterisierung des Messfehlers nicht auseinander gehalten.

In Kapitel 16 (Latent traits and item characteristic functions) wird noch einmal die bedingte Verteilung von Testscores betrachtet. Für ein gegebenes Item g wird die Itemfunktion $P_g(\theta)$ angenommen, also die Wahrscheinlichkeit, das Item zu lösen, gegeben die Fähigkeit θ. Der Testscore X_* ist wieder die Summe der X_{*g};

$$X_{*g} = \sum_{g=1}^{n} U_{*g},$$

wobei $U_{*g} = \{0, 1\}$ die Indikatorfunktion für das g-te Item ist. Die bedingte

Verteilung von X_* ist nun, so Lord & Novick, p. 384, durch

$$f(x|\theta) = \sum_{\sum u_g = x} \left(\prod_{g=1}^{n} P_g^{u_g} Q_g^{1-u_g} \right), \tag{3.8.9}$$

mit $Q_g = 1 - P_g$. Die Summe $\sum_{\sum u_g = x}$ soll bedeuten, dass über alle möglichen Antwortmuster summiert wird. Nur wenn die $P_g(\theta)$ identisch sind für alle g und θ ergibt sich hieraus eine Binomialverteilung, andernfalls hat man eine verallgemeinerte Binomialverteilung. Der wahre Wert ergibt sich in jedem Fall durch die Summe $\sum_g P_g$, und die Regression des Testscores auf die Fähigkeit ist durch

$$\mu_{X|\theta} = \sum_{g=1}^{n} P_g(\theta) = n\bar{P}(\theta) \tag{3.8.10}$$

gegeben (L& N, p. 385), mit der bedingten Varianz

$$\sigma^2_{X|\theta} = \sum_{g=1}^{n} P_g Q_g. \tag{3.8.11}$$

Sicherlich sind $\mu_{X|\theta}$ und $\sigma^2_{X|\theta}$ nicht unabhängig voneinander.

In Kapitel 23 (*Binomial Error Models*) präsentieren L & N noch einmal eine *strong true-score theory*. Die bedingte Verteilung des beobachteten Scores sei entweder binomial oder folge der verallgemeinerten Binomialverteilung (compound binomial), und es wird explizit zugelassen, dass die Verteilung der Fehler abhängig von den wahren Werten ist. Zunächst wird noch einmal das Modell (3.8.8) betrachtet; es werden u.a. Konfidenzintervalle für den wahren Wert hergeleitet. Ist man an der Verteilung der wahren Werte in der Population interessiert, so kann man die Verteilung der beobachteten Scores in der Population betrachten, die durch

$$\varphi(x) = \binom{n}{x} \int_0^1 g(\zeta)\zeta^x(1-\zeta)^{n-x}d\zeta, \quad 0 \le x \le n \tag{3.8.12}$$

gegeben sei, wobei $g(\zeta)$ die unbekannte Verteilung der wahren Werte in der Population sei. Eine Kandidatin für diese Verteilung ist die Beta-Verteilung

$$f(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}, \quad B(p,q) = \int_0^1 u^{p-1}(1-u)^{q-1}du. \tag{3.8.13}$$

Warum man nun all den komplizierten Betrachtungen folgen soll, die sich für die Verteilung der wahren Werte ergeben, ist allerdings fraglich angesichts des Arguments, dass die Annahme eines wahren Wertes ζ für alle Items kaum plausibel sein dürfte. In der Tat wird dann in Abschnitt 23.10 das *Compound binomial error model* eingeführt, denn das einfache Binomialmodell sei doch nur eine erste Approximation. Der wahre Wert ergibt sich nun durch

$$\zeta = \frac{1}{n} \sum_{g=1}^{n} P_g(\theta). \tag{3.8.14}$$

Die Frage bleibt offen, warum man den Umweg über die wahren Werte gehen soll, wenn doch die eigentliche Information über die Fähigkeit im Parameter θ liegt. Die probabilistischen Modelle fokussieren dann auch auf θ und die Schwierigkeit κ_g eines Items; eine Einführung liefert das folgende Kapitel.

4 Nichtklassische (Item Response-) Modelle

4.1 Das Generalisierte Lineare Modell

In Abschnitt 3.1 wurde in der Erläuterung 2 zum Ansatz $X = \tau + \varepsilon$ ausgeführt, dass er dem Allgemeinen Linearen Modell (ALM) entspricht. In Abschnitt 3.8.2 wurde gezeigt, dass dieser Ansatz im Falle dichotomer Items redundant ist, da τ bzw. τ_g bereits die Wahrscheinlichkeit für den Score $X_g = 1$ ist und diese nicht weiter über eine zusätzliche Verteilung für den "Fehler" ε_g definiert werden muß, – zumal sich zeigt, dass die Werte von ε_g durch die Werte von τ_g definiert werden (verl. (3.8.3), Seite 201), womit im Falle dichotomer Items die KTT inkonsistent wird, da diese Definition der Fehler der Forderung nach Unabhängigkeit von τ und ε widerspricht. In Abschnitt 3.8.2 wurde ebenfalls argumentiert, dass die KTT nicht impliziert, dass nur lineare Itemfunktionen betrachtet werden müssen, denn die Axiome der KTT beziehen sich nur auf die Beziehungen zwischen τ und ε, nicht aber auf die Beziehung zwischen τ und θ. Eine Beziehung zwischen τ und einer weiteren latenten Variablen θ ist auch ursprünglich in den Axiomen der KTT gar nicht explizit erklärt worden. Eine solche Variable wird aber gewissermaßen implizit mitgedacht: wird ein Merkmal verschieden operationalisiert, ergeben sich u. U. verschiedene Messvorschriften, die aber durch eine lineare Transformation aufeinander bezogen sein sollten. So kommt es zum Begriff der kongenerischen Tests, deren Erwartungswerte durch die lineare Transformation $\tau = \nu + \lambda\eta$ auf eine Variable η zurückgeführt werden, die gewissermaßen als kanonische Variable hinter all den gemessenen Variablen steht. Aber auch η entspricht einer Messvorschrift, und wer garantiert, dass η linear von der tatsächlichen Ausprägung des interessierenden Merkmals abhängt? Während man bei Meinungsaussagen noch vermuten kann, dass sie einigermaßen direkt die Intensität (oder was auch immer abgefragt wird) reflektieren, ist dies für psychophysiologische Messungen schon weit weniger selbstverständlich. Nimmt man aber eine nichtlineare Beziehung wie die Weibull-Funktion (3.8.6) an, so muß man nur noch die überflüssige Annahme $X_g = \tau_g + \varepsilon_g$ fallen lassen, um im Bereich des Generalisierten Linearen Modells (GLM) zu sein. Dieses Modell soll kurz in seiner allgemeinen Struktur vorgestellt werden, denn es enthält die Probabilistischen Testtheorien als Spezialfälle.

Das GLM geht auf Nelder & Wedderburn (1972) zurück. Nelder et al. gingen von dem schon in Abschnitt 3.8.2 beschriebenen Sachverhalt aus, dass die Anwendung des ALM

$$Y = b_0 + \sum_j b_j X_j + \varepsilon$$

auf abhängige Variablen Y, deren Wertebereich auf ein bestimmtes Intervall, insbesondere auf [0, 1], beschränkt ist, zu einer Verzerrung der Schätzungen für die Regressionsparameter führt, bzw. zu Wahrscheinlichkeitsschätzungen kleiner als Null oder größer als 1, wenn Y für eine Wahrscheinlichkeit steht. Zur Veranschaulichung betrachte man die folgenden Beispiele:

1. In einem Test zum Konzentrationsverlauf soll die Testperson in einem vorgegebenen Zeitintervall Δt möglichst viele einfache Aufgaben lösen, etwa Aufgaben aus dem Bereich des Kleinen Einmaleins. Der Score X ist die Anzahl k der gelösten Aufgaben. X ist eine zufällige Veränderliche, die nicht nur von Person zu Person, sondern auch für eine gegebene Person bei wiederholten Testdurchgängen variiert. Man möchte von X auf Parameter θ schließen, der die "Fähigkeit" der Person charakterisiert; eventuell möchte man darüber hinaus den Einfluß bestimmter unabhängiger Variablen, die vom Testleiter kontrolliert werden, erfassen.

2. Es wird ein Fragebogen mit n dichotomen Items konstruiert, mit dem im Rahmen der Evaluation von Therapiemaßnahmen das Ausmaß an Depressivität von Patienten in standardisierter Form erhoben werden soll. Es sollen die Parameter θ der Testpersonen bestimmt werden, die die tatsächliche Depressivität abbilden.

Im Beispiel zum Konzentrationsverlauf ist der Gesamtscore X einer Person gleích der Anzahl X der Aufgaben, die sie im Zeitintervall Δt löst, wobei die Frage nach den falschen Antworten hier der Einfachheit halber nicht weiter berücksichtigt wird. Geht man von der KTT und damit vom ALM aus, muß man weitere Annahmen über die Relation zwischen der Anzahl X und dem gesuchten Parameter θ und zu den genannten unabhängigen Variablen treffen. Es soll ja $X = \tau + \varepsilon$ gelten, wobei ε die zufälligen Komponenten in X repräsentieren soll. Man könnte versuchen, τ mit dem Erwartungswert einer approximierenden Normalverteilung in Verbindung zu bringen und dabei etwa $\tau = \theta$ setzen; der Fehler ε wäre dann wieder durch die Differenz $X - \mathbb{E}(X)$ definiert. Der Punkt hierbei ist aber, dass die Normalverteilung sicherlich nur eine Approximation sein kann, denn X kann nicht negativ werden. Wenn man diese Eigenschaft der X-Werte als nicht weiter wichtig ansieht, weil etwa die Mittelwerte $\bar{x}_a$ der Personen alle so groß sind, dass nach der Tchebyscheffschen Ungleichung die Wahrscheinlichkeit von X-Werten nahe bei Null vernachlässigbar erscheinen, bleibt aber mit der Frage konfrontiert, dass die Varianz der Fehler, also der Abweichungen $X - \mathbb{E}(X)$, von $\mathbb{E}(X)$ abhängt, die Unabhängigkeit von τ und ε also nicht mehr gegeben ist. Damit ergibt sich die Frage, wie X denn überhaupt verteilt sein kann. Mit der Beantwortung dieser Frage wird es aber womöglich überflüssig, überhaupt eine Normalverteilungsapproximation betrachten zu müssen, denn man kann dann ja gleich die angenommene Verteilung und deren Parameter in Bezug auf θ zur Interpretation der Daten heranziehen. Macht man die Annahme, dass die Aufgaben

unabhängig voneinander bearbeitet werden, so ist eine mögliche Verteilung für X die Poisson-Verteilung[52]

$$P(X = k) = e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^k}{k!}, \tag{4.1.1}$$

mit dem Erwartungswert und der Varianz

$$\mathbb{E}(X) = \lambda\Delta t, \quad \mathbb{V}(X) = \lambda\Delta t, \tag{4.1.2}$$

bei dieser Verteilung ist also die Varianz gleich dem Erwartungswert. Streng genommen kann auch diese Verteilung nur eine Approximation sein, da z.B. die Zeiten, die für das Lösen der Aufgaben benötigt werden, hier nicht berücksichtigt werden, aber die hierdurch erzeugten Verzerrungen sollen einmal als vernachlässigbar gelten. Für größer werdenden λ-Wert strebt die Poisson-Verteilung zwar auch gegen eine Normalverteilung, bei der aber Erwartungswert und Varianz nach wie vor gekoppelt sind, so dass es ist viel direkter und natürlicher ist, gleich mit der Poisson-Verteilung zu arbeiten. Wird Δt nicht variiert, kann man $\lambda^* = \lambda\Delta t$ setzen und dann wieder einfach λ statt λ^* schreiben; auf diese Weise ist Δt in λ "absorbiert" worden. Es sei

$$y = \sum_j b_j X_j, \tag{4.1.3}$$

wobei etwa speziell $b_1 = 1$ und $X_1 = \theta$ der Personparameter ist und $b_2 = 1$ und $X_2 = \kappa$ der Schwierigkeitsparameter, identisch für alle Aufgaben ist; die möglicherweise weiteren Prädiktoren können irgendwelche anderen, gerade interessierenden Variablen sein. Man kann nun

$$\lambda = y$$

setzen. Da aber $\lambda > 0$ ist, werden Einschränkungen bezüglich der b_j notwendig, was zu Verzerrungen der Gewichte b_j führen kann. Eine Alternative ist deshalb

$$\lambda = e^y \tag{4.1.4}$$

anzusetzen. Die b_j können jetzt positive oder negative Werte ohne weitere Restriktionen annehmen. Wesentlich bei diesen Betrachtungen ist, dass y zum Erwartungswert λ in Beziehung gesetzt wurde.

Im Falle der Meinungsbefragung kann man zunächst wieder annehmen, dass die Aufgaben unabhängig voneinander beantwortet werden. Die Wahrscheinlichkeit, dass ein Item I_g "positiv" beantwortet wird, sei $p_g = P(X_g = 1|y)$. Die Frage ist nun, wie p_g mit y aus (4.1.3) verknüpft werden kann. Geht man von der Vorstellung aus, dass die Einstellung bezüglich des betrachteten Merkmals

[52] Diese Verteilung wird gelegentlich als Verteilung seltener Ereignisse bezeichnet, was nicht inkorrekt ist, aber irreführend sein kann, denn die tatsächlich beobachteten X-Werte können durchaus "groß" sein, wenn nur λ hinreichend groß gewählt wird.

bei jeder Person zeit- und bedingungsabhängig variiert, kann man wieder eine latente ("unterliegende") Variable η einführen und postulieren, dass die Antworten von den Werten von η abhängen: je nach Formulierung der Frage wird sie "positiv" beantwortet, wenn entweder $\eta \leq \kappa_g$ oder $\eta > \kappa_g$ ist, wobei κ_g ein für die jeweilige Frage charakteristischer kritischer Wert ist. Nun muß man eine Verteilung für η annehmen. Alten Traditionen (Thurstone) folgend könnte man die Gauß-Verteilung adoptieren:

$$f_a(\eta) = \frac{1}{\sigma_a\sqrt{2\pi}} \exp\left(-\frac{(\eta-\mu_a)^2}{2\sigma_a^2}\right). \tag{4.1.5}$$

Hier ist der Index a hinzugefügt worden, um anzuzeigen, dass die Verteilung von η durch für die a-te Person charakteristische Parameter definiert sein kann. So hat η bei der a-ten Person den Erwartungswert $\mathbb{E}(\eta) = \mu_a$ und die Varianz $\mathbb{V}(\eta) = \sigma_a^2$. Die Wahrscheinlichkeit, dass die Person dem Item I_g zustimmt, könnte diesem Ansatz zufolge durch

$$P(\eta > \kappa_g|\mu_a, \sigma_a^2) = 1 - P(\eta \leq \kappa_g|\mu_a, \sigma_a^2) = 1 - \int_{-\infty}^{\kappa_g} f_a(\zeta)d\zeta. \tag{4.1.6}$$

gegeben sein. Um den Effekt der unabhängigen Variablen y aus (4.1.3) modellieren zu können, kann man nun

$$\mu_a = y_a = \sum_j b_j X_{aj} \tag{4.1.7}$$

setzen. Hier wurde auch y mit a indiziert, um anzuzeigen, dass die X_j für a spezifische Werte annehmen können. So könnte $b_1 = 1$ und $X_{a1} = \theta_a$ ein personspezifischer Parameter sein, der die Depressivität abbildet. Nun ist

$$P(\eta \leq \kappa_g) = P\left(\frac{\eta - y_a}{\sigma_a} \leq \frac{\kappa_g - y_a}{\sigma_a}\right) = P\left(Z \leq \frac{\kappa_g - y_a}{\sigma_a}\right),$$

so dass (4.1.6) über die Standardnormalverteilung ausgedrückt werden kann:

$$P(\eta > \kappa_g|y_a) = 1 - \int_{-\infty}^{(\kappa_g - y_a)/\sigma_a} f(z)dz, \tag{4.1.8}$$

mit $f(z) = \exp(-z^2/2)/\sqrt{2\pi}$, und das bedeutet

$$P(\eta > \kappa_g|y_a) = 1 - \Phi[(\kappa_g - y_a)/\sigma_a] = \Phi[(y_a - \kappa_g)/\sigma_a]. \tag{4.1.9}$$

Wieder wurde y_a nicht mit $p_{ag} = P(\eta > \kappa_g|y_a)$ gleichgesetzt, vielmehr wurde y_{ag} linear auf das *Probit* Φ^{-1} bezogen:

$$\Phi^{-1}(p_{ag}) = y_a, \quad \text{(Probit)} \tag{4.1.10}$$

Natürlich hätte man auch die logistische Verteilung für η annehmen können. Wie bereits in Abschnitt 2.6.3 gezeigt wurde, erhält man die Beziehung

$$P(\eta > \kappa_g | y_a) = p_{ag} = \frac{e^{y_a}}{1 + e^{y_{ag}}}, \tag{4.1.11}$$

und daraus nach kurzer Rechnung

$$\log\left(\frac{p_{ag}}{1 - p_{ag}}\right) = y_a. \quad \text{(Logit)} \tag{4.1.12}$$

Natürlich lassen sich noch andere Verteilungen diskutieren, etwa die in (3.8.6) (Seite 202) schon eingeführte Weibull-Verteilung. Für die Zwecke dieses Abschnitts genügt es aber, es bei den betrachteten Beispielen zu belassen.

Fasst man das Vorgehen zusammen, so lassen sich verschiedene Schritte kennzeichnen:

1. **Strukturannahme:** Es wird zunächst eine Variable $y_a = \sum_j b_j X_{aj}$ definiert. Die X_{aj} sind Prädiktoren für eine Indikatorvariable X_g bzw X_{ag}, die anzeigt, ob eine bestimmte Antwort oder allgemein Reaktion erfolgt ist. Ein Prädiktor kann u. a. ein Personenparameter θ sein.

2. **Verteilungsannahme:** Für X_g bzw. für $X = \sum_g X_g$ wird nicht mehr das ALM $X_g = \tau_g + \varepsilon_g$ bzw $X = \tau + \varepsilon$ angenommen, sondern es wird eine Wahrscheinlichkeitsverteilung für die Daten angenommen. Die hier betrachteten Verteilungen – die Poisson-, die Gauß- und die logistische Verteilung sind spezielle Beispiele aus einer Klasse von Verteilungen, die überhaupt in Frage kommen. Diese Klasse ist die *Exponentialfamilie*[53]: zu ihr gehören auch die Binomialverteilung, die Exponentialverteilung sowie die Verteilung der Summe unabhängiger, exponentialverteilter Variablen, die Weibull-Verteilung, Extremwertverteilungen, etc. Auf die allgemeine Diskussion der Exponentialfamilie soll hier verzichtet werden, da im Folgenden nicht weiter auf sie rekurriert werden wird. Wichtig ist aber, zu sehen, dass außer der Gauß- oder der logistischen Verteilung durchaus noch andere Verteilungen in Betracht gezogen werden können.

[53] Die Dichte ist, für irgendwelche Daten z, durch

$$f(z|\theta, \phi) = \exp\left[\alpha[\theta)(z\theta - g(\theta) + h(z)) + \beta(\phi, z)\right]$$

gegeben, mit $\alpha(\theta) > 0$. θ und ϕ sind irgendwelche Parameter, also nicht notwendig Personen- oder Itemparameter. ϕ ist ein "nuisance"-Parameter, d.h. ein Parameter, der in einer Verteilung auftritt, ohne dass man eigentlich an ihm interessiert ist. So ist man oft am Erwartungswert μ einer Normalverteilung interessiert, weil man annimmt, dass bestimmte experimentelle Variablen auf μ einwirken, und die Varianz σ^2 ist nicht weiter von Interesse, muß aber mit berücksichtigt werden. b, h und β sind geeignet zu wählende Funktionen. Bei geeigneter Wahl dieser Funktionen geht f in die Binomial-, Gauß-, logistische -, Weibull-Verteilung etc über. Ausgearbeitete Beispiele findet man etwa in Fahrmeir, Hamerle und Tutz (1996), Kapitel 2.

3. **Link-Funktion:** Die Größe $y = \sum_j b_j X_j$ wird über eine bestimmte Funktion H mit einem Parameter der betrachteten Wahrscheinlichkeitsverteilung, oft der Erwartungswert μ, in Beziehung gebracht:

$$H(\mu) = y. \tag{4.1.13}$$

H heißt *Link-Funktion* (linking function, Nelder & Wedderburn (1972), p. 372). der Poisson-Verteilung etwa wurde $\mu = \lambda = e^y$ gewählt; Dann ist $H(\mu) = \log \lambda = y$ die Link-Funktion. Bei der Gauß-Verteilung ergibt sich die Probit-Beziehung $H(\mu) = \Phi^{-1}(p_g) = y$, $p_g = p_g(Y)$ als Link-Funktion, und bei der logistischen Verteilung hat man

$$H(\mu) = \log\left(\frac{p_g}{1-p_g}\right) = y, \quad p_g = p_g(y).$$

Diese Definition entspricht (4.1.11) oder (4.1.12), aber man kann y über die Reparametrisierung der logistischen Funktion mit dem Erwartungswert μ in der Verteilungsfunktion

$$P(\eta \leq \kappa) = \frac{1}{1 + \exp\left(\frac{\kappa-\mu}{\sigma}\right)}$$

in Beziehung setzen, also $y = \mu$; der Faktor $\sqrt{3}/\pi$ ist bereits in σ absorbiert worden (vergl. (2.6.85), Seite 77).

Die logistische bzw. kategoriale Regression ist eine der bekanntesten Anwendungen des GLM; für die Wahrscheinlichkeit eines Ereignisses, das über eine Indikatorfunktion $X = \{0, 1\}$ signalisiert wird und die von unabhängigen Variablen $X_1, \ldots, X_r$ abhängt, hat man bekanntlich

$$P(X = 1|X_1, \ldots, X_r) = \frac{1}{1 + \exp(b_1 X_1 + \cdots b_r X_r)}, \tag{4.1.14}$$

wobei $P(X = 1|X_1, \ldots, X_r)$ durch entsprechende relative Häufigkeiten geschätzt wird. (4.1.14) ist ein Beispiel dafür, wie unabhängige Variablen $X_1, \ldots, X_r$ auf Häufigkeiten oder relative Häufigkeiten bezogen werden können: auf jeden Fall nicht durch unbeschränkte Anwendung des ALM

$$P(X = 1|X_1, \ldots, X_r) = b_1 X_1 + \cdots b_r X_r.$$

Auch diese Beziehung kann gelten, – aber nur als Approximation in der Nachbarschaft eines Punktes P_0, in dem P in guter Näherung eine lineare Funktion der $X_1, \ldots, X_r$ ist, meistens um $P_0 = .5$ herum. Für die Analyse des allgemeinen Zusammenhanges zwischen den unabhängigen Variablen und P ist die lineare Approximation aber nicht hinreichend und führt zu verzerrten Schätzungen für die b_j.

Im Folgenden wird der Spezialfall, dass $X_1 = \theta$ ein Personenparameter und $X_2 = \kappa$ ein Itemparameter betrachtet. Will man die Auswirkungen irgendwelcher Maßnahmen auf θ untersuchen, ist man also automatisch bei der logistischen Regression oder der Logit-Analyse. Auch eine Probit-Anlyse könnte gewählt werden, es zeigt sich aber, dass die Logit-Analyse, d.h. das logistische Modell, einige grundsätzliche Vorzüge hat, die sie für die Anwendung in der Theorie psychometrischer Tests besonders geeignet macht.

4.2 Modelle auf der Basis der logistischen Funktion

Es werden zunächst dichotome Items betrachtet. Wie in Abschnitt 3.8.2 ausgeführt wurde, impliziert der Fall dichotomer Items Inkonsistenzen für die KTT. Die Diskussion dichotomer Items führt vielmehr direkt auf das Generalisierte Lineare Modell (GLM), wie es im vorangegangenen Abschnitt vorgestellt wurde. Statt also das Modell $X = \tau + \varepsilon$ zu betrachten, wird man auf die Notwendigkeit geführt, eine geeignete Itemfunktion für $P(X_g = 1|\theta, \kappa_g)$ zu bestimmen. Naheliegend sind das Ogive-Modell, oder das logistische Modell, die Weibull-Funktion. Aber auch andere Funktionen, die sich aus anderen Verteilungsfunktionen ergeben, sind denkbare Kandidaten für die Definition eines probabilistischen Modells im Rahmen des GLM, man denke an Extremwerteverteilungen oder die Gompertz-Verteilung. Es zeigt sich aber, dass das logistische Modell eine Reihe von für die Testtheorie interessanten Eigenschaften hat, so dass die Betrachtungen auf dieses Modell fokussiert werden.

Das Birnbaum-Modell: In Abschnitt 2.6, Gleichung (2.6.6), Seite 36, wurde das logistische

$$P(X_g = 1|\theta, \kappa_g, \gamma_g) = \gamma_g + (1 - \gamma_g)\frac{\exp(\alpha_g(\theta - \kappa_g))}{1 + \exp(\alpha_g(\theta - \kappa_g))} \tag{4.2.1}$$

Modell eingeführt. Wie im Kommentar zu (2.6.85) ausgeführt wurde, kann man die Interpretation $\alpha_g = 1/\sigma_g$ und $b_g = \kappa_g$ wählen, wobei σ_g die Streuung einer unterliegenden Variable ist, die das fluktuierende Merkmal repräsentiert. Das Modell ist auch als Birnbaum-Modell bekannt; Birnbaum (1968) hat es in einem messtheoretischen Rahmen eingeführt. Bei ihm wird β_g statt α_g geschrieben und kein Bezug zu einer unterliegenden Variablen hergestellt. Die Einführung des Faktors β_g erlaubt itemspezifische Trennschärfen, die bei dem im folgenden Abschnitt besprochenen Modell nicht möglich sind. In (4.2.1) wird noch die Möglichkeit, dass eine Testperson rät, in Rechnung gestellt. Dies ist das *dreiparametrische logistische Modell* (3PL-Modell); die drei Parameter sind γ_g, α_g und κ_g, – θ hat hier die Bedeutung einer unabhängigen Variable. Für $\gamma_g = 0$ erhält man das *zweiparametrige Modell* (2PL-MOdell), und für $\alpha_g = 1$ ergibt sich das *einparametrige Modell* (1PL-Modell). Das 1PL-Modell ist ein wichtiger Spezialfall, das unter dem Namen *Rasch-Modell* bekannt ist; es wird im folgenden Abschnitt behandelt.

4.3 Das Rasch-Modell

Das Rasch-Modell wurde bereits auf Seite 36 erwähnt und wird hier noch einmal vorgestellt.

Definition 4.3.1 *Es sei $X_g = \{0, 1\}$ eine Indikatorvariable definiert: $X_g = 1$, wenn die Aufgabe gelöst oder der Meinungsaussage zugestimmt wurde, und $X_g = 0$ sonst. Die Itemfunktion sei durch*

$$F_g(\theta) = P(X_g = 1|\theta) = \frac{e^{\theta-\kappa_g}}{1 + e^{\theta-\kappa_g}}, \tag{4.3.1}$$

gegeben und es gelte lokale Unabhängigkeit. Das Testmodell heißt dann Rasch-Modell*.*

Ein wichtiger Begriff, der in Bezug auf (4.3.1) eingeführt werden muß, ist der der *Rasch-Homogenität*: intuitiv ist damit gemeint, dass die Items, die (4.3.1) genügen, alle nur ein und dasselbe Merkmal messen, dessen Ausprägung bei einer Person eben durch θ bzw. θ_a repräsentiert wird. Die Rasch-Homogenität läßt sich auch formal definieren, und aus dieser Definition folgt dann (4.3.1):

Definition 4.3.2 *Es seien X_g, X_h Indikatorvariablen, $g, h = 1, \ldots, n$, für dichotome Items. Irgendzwei Items I_g und I_h sind* Rasch-homogen*, wenn*

$$\theta - \kappa_g = \log\left[\frac{P(X_g = 1|\theta)}{P(X_g = 0|\theta)}\right] = \log\left[\frac{P(X_h = 1|\theta)}{P(X_h = 0|\theta)}\right] + \delta_{gh}. \tag{4.3.2}$$

Anmerkung: Formal bedeutet (4.3.2), dass sich die Logits für irgendzwei Items nur um eine additive Konstante – δ_{gh} – unterscheiden sollen; für fixen θ-Wert bedeutet dies, dass sich die additive Konstante nur auf eine Differenz der Schwierigkeiten beziehen kann, was wiederum impliziert, dass θ und die Schwierigkeiten sich auf das gleiche Merkmal beziehen müssen.

> **Nachweis der Äquivalenz von (4.3.2) und (4.3.1):** Gilt nun (4.3.1), so folgt leicht, dass auch (4.3.2) gilt, mit $\delta_{gh} = \kappa_h - \kappa_g$.
>
> Um zu sehen, dass die Beziehung (4.3.2) auch (4.3.1)impliziert, betrachte man zunächst
>
> $$\theta - \kappa_g = \log\left[\frac{P(X_g = 1|\theta)}{P(X_g = 0|\theta)}\right].$$
>
> Dies ist gleichbedeutend mit
>
> $$e^{\theta-\kappa_g} = \frac{P(X_g = 1|\theta)}{P(X_g = 0|\theta)} = \frac{P(X_g = 1|\theta)}{1 - P(X_g = 1|\theta)},$$
>
> woraus sofort
>
> $$P(X_g = 1|\theta) = \frac{e^{\theta-\kappa_g}}{1 + e^{\theta-\kappa_g}}$$

folgt. Weiter folgt in Bezug auf die rechte Seite von (4.3.2)

$$e^{\theta-\kappa_g} = \frac{P(X_h=1|\theta)e^{\delta_{gh}}}{P(X_h=0|\theta)} = \frac{P(X_h=1|\theta)e^{\delta_{gh}}}{1-P(X_h=1|\theta)},$$

woraus sich

$$P(X_h=1|\theta) = \frac{e^{\theta-\kappa_g-\delta_{gh}}}{1+e^{\theta-\kappa_g-\delta_{gh}}} = \frac{e^{\theta-\kappa_h}}{1+e^{\theta-\kappa_h}}$$

ergibt. □

Schreibweisen: Die Ausdrücke für $P(X_g=1|\theta,\kappa_g)$ und $P(X_g=0|\theta,\kappa_g)$ sind verschieden. Die Schreibweise

$$P(X_g) = \frac{e^{x_g(\theta-\kappa_g)}}{1+e^{\theta-\kappa_g}} \tag{4.3.3}$$

fasst beide Ausrücke zusammen: für $x_g=1$ ergibt sich der Ausdruck für $P(X_g=1|\theta,\kappa_g)$, und für $x_g=0$ erhält man $P(X_g=0|\theta,\kappa_g)$.

Parametrisierungen: Die Form (4.3.1) des Rasch-Modells heißt *subtraktive Parametrisierung*, da die Wahrscheinlichkeit einer Lösung oder Beantwortung ($X_g=1$) von der Differenz $\theta-\kappa_g$ abhängt. Es sei noch einmal darauf hingewiesen, dass sowohl die Person (mit dem Parameter θ) wie auch das Item I_g (mit dem Parameter κ_g) auf der gleichen Skala positioniert werden. Nun gilt aber auch

$$\frac{e^{\theta-\kappa_g}}{1+e^{\theta-\kappa_g}} = \frac{e^{\theta}e^{-\kappa_g}}{1+e^{\theta}e^{-\kappa_g}},$$

d.h. das Modell kann in der Form

$$F_g(\theta) \to G_g(\vartheta) = \frac{\vartheta\sigma_g}{1+\vartheta\sigma_g}, \quad \vartheta=e^{\theta}, \quad \sigma_g=e^{-\kappa_g} \tag{4.3.4}$$

geschrieben werden; diese Form des Modells heißt *multiplikative Parametrisierung*. Für manche Betrachtungen ist diese Art der Parametrisierung vorteilhaft.

Ein Vorteil dieser Parametrisierung ist, dass sie auf Beziehungen des Rasch-Modells mit anderen Modellen verweist. Dividiert man auf der rechten Seite von (4.3.4) Zähler und Nenner durch σ_g, so erhält man

$$G_g(\vartheta) = \frac{\vartheta}{1/\sigma_g+\vartheta} = \frac{\vartheta}{\zeta_g+\vartheta}, \quad \zeta_g = 1/\sigma_g. \tag{4.3.5}$$

Diese Form des Rasch-Modells korrespondiert zum *Bradley-Terry-Modell* bzw. zum *Bradley-Terry-Luce-Modell* für den Paarvergleich, womit eine Beziehung des Rasch-Modells zur Messtheorie hergestellt wird (vergl. Suppes & Zinnes (1963), insbesondere Abschnitt 4, s.a. Beispiel B.2.3, Seite 295 im Anhang).

Parallele Itemfunktionen: Gegeben seien zwei Items I_g und $I_{g'}$, mit $\kappa_{g'} = \kappa_g + \Delta\kappa$. Für die Itemfunktionen F_g und $F_{g'}$ gilt nun

$$F_g(\theta) = \frac{e^{\theta-\kappa_g}}{1+e^{\theta-\kappa_g}}, \quad F_{g'}(\theta) = \frac{e^{\theta-\kappa_g-\Delta\kappa}}{1+e^{\theta-\kappa_g-\Delta\kappa}} = \frac{e^{(\theta-\Delta\kappa)-\kappa_g}}{1+e^{(\theta-\Delta\kappa)-\kappa_g}},$$

d.h. aber

$$F_g(\theta) = F_{g'}(\theta'), \quad \theta' = \theta - \Delta\kappa \tag{4.3.6}$$

Dies bedeutet, dass sich die Itemfunktionen $F_{g'}$ und F_g für die Items I_g und $I_{g'}$ nur durch eine Verschiebung um $\Delta\kappa$ auf der θ-Skala unterscheidet. Die beiden Itemfunktionen sind *parallel*. Dies folgt natürlich schon aus der Tatsache, dass die Itemfunktionen beim Rasch-Modell gleiche Trennschärfe anzeigen.

Zulässige Transformationen: Die Gleichung (4.3.6) verweist darauf, dass die θ- und der κ-Skalen offenbar nur eindeutig auf Translationen sind. Dies folgt eben aus

$$\theta' = \theta + \alpha,\ \kappa'_g = \kappa_g + \alpha \Rightarrow \theta - \kappa_g = \theta' - \kappa'_g, \tag{4.3.7}$$

d.h. die Skala, auf der θ und κ_g Werte annehmen, ist eindeutig bis auf Addition einer beliebigen, endlichen Konstanten α (die positiv oder negativ sein kann). Hat man also Skalenwerte θ und κ_g gefunden, die dem Modell genügen, so kann man

$$\alpha = -\sum_{g=1}^{n} \kappa_g \tag{4.3.8}$$

setzen und zu Werten $\theta' = \theta + \alpha$, $\kappa'_g = \kappa_g + \alpha$ übergehen. Anschließend kann man θ' und κ'_g wieder in θ und κ_g umbenennen, und für die κ_g-Werte gilt dann

$$\sum_{g=1}^{n} \kappa_g = 0. \tag{4.3.9}$$

Wettquotienten und Logits: Aus (4.3.1) folgt sofort

$$P(X_g = 0|\theta) = 1 - P(X_g = 1|\theta) = \frac{1}{1+e^{\theta-\kappa_g}}.$$

Daraus ergibt sich die als *Wettquotient* bekannte Größe

$$\frac{P(X_g = 1)}{P(X_g = 0)} = e^{\theta-\kappa_g}, \tag{4.3.10}$$

und durch Logarithmieren[54] erhält man das *Logit*

$$\text{Logit}P(X_g = 1|\theta) = \log\left(\frac{P(X_g = 1)}{P(X_g = 0)}\right) = \theta - \kappa_g. \tag{4.3.11}$$

[54]Es ist hier stets der *natürliche Logarithmus*, also der Logarithmus zur Basis e, gemeint.

Der Wettquotient gibt an, wie die Chancen stehen, dass z.B. die Person a das Item I_g löst. Es sei etwa $P(X_{ag} = 1) = .75$. Dann ist $P(X_{ag} = 0) = .25$ der Wettquotient beträgt $.75/.25 = 3/1$, d.h. die Chancen stehen 3 zu 1, dass a die Aufgabe löst. Beträgt die Wahrscheinlichkeit, das Item zu beantworten oder zu lösen, nur .65, so ist $.65/.35 \approx 1.8 \approx 9/5$; - Wettquotienten werden üblicherweise als Quotienten ganzer Zahlen ausgedrückt und sind deswegen im Allgemeinen Näherungswerte.

Spezifische Objektivität: Gegeben seien zwei Probanden $a, a' \in \mathcal{P}$ mit den Parameterwerten θ_a und $\theta_{a'}$. Für ein gegebenes Item I_g werde nun die Differenz der Logits für die beiden Probanden betrachtet:

$$\log\left(\frac{P(X_{ag}=1)}{P(X_{ag}=0)}\right) - \log\left(\frac{P(X_{ga'}=1)}{P(X_{ga'}=0)}\right) = \theta_a - \theta_{a'} - \kappa_g + \kappa_g = \theta_a - \theta_{a'}. \quad (4.3.12)$$

Die Differenz der Logits ist also nur durch die Differenz der Probandenparameter definiert und unabhängig vom Itemparameter κ_g. Dieses Ergebnis gilt natürlich für ein beliebiges Item I_g, also für alle I_g. Die Gleichung (4.3.12) erlaubt den Vergleich zweier Personen unabhängig von den Items, mit denen sie getestet werden. Diese Eigenschaft eines Testmodells, also der Vergleich von Personen unabhängig von den im Test verwendeten Items, wird die *spezifische Objektivität* eines Tests genannt.

Auf analoge Weise sieht man, dass die Differenz der Logits für zwei Items unabhängig von den Personen ist:

$$\log\left(\frac{P(X_{ag}=1)}{P(X_{ag}=0)}\right) - \log\left(\frac{P(X_{g'a}=1)}{P(X_{g'a}=0)}\right) = \theta_a - \theta_a - \kappa_g + \kappa_{g'} = \kappa_{g'} - \kappa_g. \quad (4.3.13)$$

Diese Gleichung bedeutet, dass man die Differenz der Schwierigkeit zweier Items unabhängig von irgendwelchen Personenparametern bestimmen kann, d.h. die Items können unabhängig von der Stichprobe oder gar Population von Probanden hinsichtlich ihrer Schwierigkeit miteinander verglichen werden. Die spezifische Objektivität und die Populationsunabhängigkeit der Itemparameter sind charakteristisch für das Rasch-Modell und machen einen großen Teil seiner Attraktivität aus.

4.4 Die Schätzung der Modellparameter

Werden Personen befragt oder getestet, so möchte man i. A. etwas über die Ausprägung der untersuchten Merkmals bei den Personen erfahren. Darüber hinaus kann es von Interesse sein, zu erfahren, ob die Items das Merkmal in unterschiedlichem Maße erfassen oder nicht, ob sie also gleich schwierig sind oder nicht. Dazu müssen im Falle des Rasch-Modells die Parameter θ_a und κ_g geschätzt werden.

Das Problem bei Modellen wie denen von Birnbaum und Rasch ist, dass die Parameter nicht in direkter Beziehung zu beobachtbaren Größen stehen. Denn es

werden nur Wahrscheinlichkeiten "vorausgesagt", um im Jargon der Regressionsanalyse zu bleiben: Die Gleichung

$$P(X_{ag} = 1|\theta_a, \kappa_g) = \frac{e^{\theta_a - \kappa_g}}{1 + e^{\theta_a - \kappa_g}}$$

besagt ja, dass die rechte Seite eben nur die Wahrscheinlichkeit $P(X_{ag} = 1|\theta_a, \kappa_g)$ bestimmt. Die wird aber nicht beobachtet, sondern nur die manifeste Variable X_{ag}, die bei der Person a und dem Item I_g bei einer Testvorgabe entweder den Wert 1 oder 0 annimmt. Man könnte die Wahrscheinlichkeit durch wiederholte Vorlage abschätzen, nur wird das nicht funktionieren: wenn die Person etwa die Lösung erinnert, wird die manifeste Variable stets den Wert 1 annehmen. Abgesehen davon wird die Anzahl der Vorlagen des Tests kaum jemals hinreichend groß sein können, um eine vernünftige Schätzung von $P(X_{ag} = 1)$ anhand relativer Häufigkeiten zu erlauben.

Die Methode der Wahl ist demnach die *Maximum-Likelihood-Methode. Likelihood* heißt zunächst einfach nur – wie 'probability' – 'Wahrscheinlichkeit', meint hier aber speziell die Wahrscheinlichkeit der beobachteten Daten oder Messungen, unter der Bedingung, dass die Parameter der entsprechenden Wahrscheinlichkeitsverteilung bestimmte Werte haben. Es seien die Daten einer Stichprobe von m Personen gegeben, die auf n dichotome Items antworten. Die Daten können in einer Matrix zusammengefasst werden; man spricht auch von einer *Datenstruktur*:

$D =$

Person	Items					Σ_a
1	0	0	0	0	0	0
2	0	0	0	0	1	1
3	0	1	0	0	0	1
4	1	1	0	1	0	3
5	0	0	1	1	1	3
6	1	1	0	0	0	2
7	1	0	1	1	1	4
8	0	0	0	0	1	1
9	1	1	0	1	1	4
10	0	0	1	1	0	2
Σ_g	4	4	3	5	5	

(4.4.1)

Hier gibt es $m = 10$ befragte oder getestete Personen $a = 1, \ldots, m$, und fünf dichotome Items, $g = 1, \ldots, n$. Eine 1 oder eine 0 repräsentiert die Realisierung der zufälligen Veränderlichen $X_{ag} = \{0, 1\}$. $n_g = \sum_a X_{ag}$ ist die Häufigkeit, mit der das g-te Item gelöst bzw. positiv beantwortet wurde, $m_a = \sum_g X_{ag}$ ist die Anzahl der Items, die von der Person a gelöst wurden.

Man kann nun die Likelihood dieser Daten definieren, vorausgesetzt, die X_{ag} sind alle stochastisch unabhängig. Die Likelihood der Daten ist durch

$$L(X_{11}, \ldots, X_{mn}) = \prod_{a=1}^{m} \prod_{g=1}^{n} \frac{e^{X_{ag}(\theta_a - \kappa_g)}}{1 + e^{\theta_a - \kappa_g}}, \quad X_{ag} = \{0, 1\} \tag{4.4.2}$$

gegeben.

> **Anmerkung:** Man spricht hier bewußt von Likelihood und nicht von der Wahrscheinlichkeit, weil einerseits der Begriff der Likelihood sich auf die Wahrscheinlichkeit von Daten, gegeben bestimmte Bedingungen (hier: die Geltung des Rasch-Modells), bezieht, zum anderen, weil L nicht notwendig eine Wahrscheinlichkeit im strengen Sinn des Wortes ist: ist etwa f die Dichtefunktion einer stetigen zufälligen Veränderlichen, so ist bekanntlich $f(x)$ nicht die Wahrscheinlichkeit, dass X den Wert x annimmt. Diese Wahrscheinlichkeit ist gleich Null. Nicht notwendig gleich Null ist die Wahrscheinlichkeit, dass X einen Wert in einem Intervall $x, x + \Delta x$ annimmt. Diese Wahrscheinlichkeit ist durch $F(x + \Delta x) - F(x)$ gegeben, und f ist der Grenzwert eines Quotienten
>
> $$\lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = f(x).$$
>
> Dementsprechend ist
>
> $$F(x + \Delta x) - F(x) = f(x)\Delta x$$
>
> eine Wahrscheinlichkeit. Für beliebig kleines $\Delta x = dx$ heißt $f(x)dx$ *Wahrscheinlichkeitselement.*
>
> Andererseits kann L auch eine Wahrscheinlichkeit repräsentierten, wenn nämlich X eine diskrete zufällige Veränderliche ist.

Für die Zwecke der Parameterschätzung kann aber, trotz der vorangegangenen Anmerkung, L wie eine Wahrscheinlichkeit behandelt werden. Denn das Wahrscheinlichkeitselement $f(x)dx$ wird maximal, wenn $f(x)$ maximal wird, und es wird gleich Null, wenn $f(x) = 0$. In diesem Sinne maximiert man die Wahrscheinlichkeit der Daten in Abhängigkeit von den Werten der freien Parameter in einem Ausdruck für L, wenn man L maximiert.

Bei der Likelihood-Funktion (4.4.2) sind die Parameter θ_a und κ_g die freien Parameter, deren Werte bestimmt werden sollen. Der Bestimmung dieser Werte liegt die folgende 'Philosophie' zugrunde: Es folgt ja aus dem Begriff der Wahrscheinlichkeit, dass ein zufälliges Ereignis E, dessen Wahrscheinlichkeit höher höher ist als die des Ereignisses $E' \neq E$, eben eher und damit auch häufiger eintritt als E'. Hat man umgekehrt ein zufälliges Ereignis E beobachtet, so kann man davon ausgehen, dass E eine höhere Wahrscheinlichkeit hat als die Ereignisse E', deren Wahrscheinlichkeit geringer ist. Das kann im Einzelfall durchaus falsch sein, aber *im Allgemeinen* liegt man mit dieser Vermutung richtig. Da nach (4.4.2) die Wahrscheinlichkeit L der Daten von den Werten der Parameter θ_a und κ_g abhängt, ist es dementsprechend eine plausible Annahme, dass sie L maximieren. Man kann L als Funktion von θ_a und κ_g auffassen. Die Aufgabe ist nun, diejenigen Werte $\hat{\theta}_a$ und $\hat{\kappa}_g$ zu bestimmen, für die L maximal wird. Diese Werte bestimmt man, indem man die *Likelihood-Funktion* einmal nach θ_a differenziert, und einmal nach κ_g; man spricht von *partiellen Ableitungen.* Für diejenigen Werte

$\hat{\theta}_a$ und $\hat{\kappa}_g$, für die die beiden partiellen Ableitungen gleich Null werden, nimmt L einen maximalen Wert an. Man fasst dementsprechend die beiden partiellen Ableitungen als Gleichungen in den Unbekannten $\hat{\theta}_a$ und κ_g auf, mit $a = 1, \ldots, m$ und $g = 1, \ldots, n$, d.h. es müssen $n + m$ Gleichungen gelöst werden.

Die Herleitung dieser Gleichungen soll hier nicht durchgeführt werden; das partielle Differenzieren von L nach den θ_a und κ_g etc. ist ein wenig länglich und trägt zum Verständnis des Folgenden nichts weiter bei. Es genügt, das Resultat dieser Rechnungen anzugeben:

$$m_a = \sum_{g=1}^{n} \frac{\exp(\hat{\theta}_a - \hat{\kappa}_g)}{1 + \exp(\hat{\theta}_a - \hat{\kappa}_g)}, \quad a = 1, \ldots, m \tag{4.4.3}$$

$$n_g = \sum_{a=1}^{m} \frac{\exp(\hat{\theta}_a - \hat{\kappa}_g)}{1 + \exp(\hat{\theta}_a - \hat{\kappa}_g)}, \quad g = 1, \ldots, n. \tag{4.4.4}$$

Es ist nicht möglich, diese Gleichungen explizit nach den θ_a und κ_g aufzulösen, so dass eine numerische Lösung gefunden werden muß. Programme für die Lösung der Gleichung sind verfügbar.

Anmerkungen zu den Schätzungen: Schätzmethoden und einige wesentliche Eigenschaften von Schätzern sind kurz in Abschnitt 2.6.10 behandelt worden.

1. **Suffizienz:** Zunächst fällt auf, dass die Daten nur über die Summen m_a und n_g in die Gleichungen eingehen. m_a ist die Summe – hier die Anzahl – der Punktwerte für die gelösten Items bei der Person a. Der Punktwert der Person a wird also ohne jede Gewichtung der einzelnen Antworten berechnet. Jede "korrekte" Antwort liefert einen Punkt, unabhängig von der Schwierigkeit des beantworteten Items. Offenbar liefern die m_a – neben den n_g – also bereits hinreichende Information über die zu schätzenden Parameter. Die m_a sind dementsprechend auch *suffiziente Statistiken* (vergl. Gleichung (2.6.100) in Definition 2.6.3, 82).

2. **Verteilung der Schätzungen:** Die Schätzungen $\hat{\theta}_a$ und $\hat{\kappa}_g$ sind Schätzungen auf der Basis einer Stichprobe und damit fehlerbehafted. Möchte man einer Person einen Parameterwert θ zuordnen, so ist die Schätzung $\hat{\theta}_a$ nicht notwendig gleich dem "wahren" Wert θ_a. Deswegen ist man daran interessiert, z.B. ein Konfidenzintervall angeben zu können, dessen Bestimmung aber Kenntnis der Verteilung der Schätzungen $\hat{\theta}_a$ voraussetzt. Maximum-Likelihood-Schätzungen haben nun die Eigenschaft, *asymptotisch normal* zu sein. Damit hat man eine Basis für die Konstruktion von Konfidenzintervallen.

3. **Erwartungstreue und Konsistenz:** Die Schätzung $\hat{\theta}$ ist erwartungstreu und konsistent;das gleiche gilt für die Schätzungen von κ. Im Falle eines Testmodells bedeutet dies, dass nicht nur die Anzahl der Personen, sondern

auch die Anzahl der Items hinreichend groß sein muß, damit die Schätzungen $\hat{\theta}$ und $\hat{\kappa}$ hinreichend nahe an den wahren Werten θ und κ sind.

4. **Extreme Reihen- bzw. Spaltensummen:** Es kann geschehen, dass eine Person a keines der Items beantwortet, so dass $m_a = 0$ ist. Ebenso kann es geschehen, dass ein Item I_g von keiner Person beantwortet wird, so dass $n_g = 0$. Betrachtet man nun die Gleichungen (4.4.3) und (4.4.4), so sieht man, dass für diese Fälle Schwierigkeiten entstehen. Denn $m_a = 0$ impliziert $\exp(\hat{\theta}_a - \hat{\kappa}_g) = 0$, und dies ist nur möglich, wenn $\theta_a = -\infty$. Analog dazu impliziert $n_g = 0$, dass $\kappa_g = \infty$. Ein solches Item liefert aber keinerlei Information mehr über das zu untersuchenden Merkmal.

   Eine einfache Lösung besteht nun darin, das Item aus dem Test zu entfernen. Ebenso ist die Person a nicht weiter interessant, – abgesehen von der Frage, was es bedeuten soll, eine Merkmalsausprägung von $-\infty$ zu haben.

   Alternative Möglichkeiten der Schätzung, bei denen die hier genannte Problematik umgangen werden kann, beruhen auf dem Bayes-Theorem, vergl. Rost (2004), Abschnitt 4.2.

5. **Eindeutigkeit der Schätzungen:** Die Schätzung der Parameter θ_a und κ_g beruht auf der Lösung nichtlinearer Gleichungen. Im Prinzip ist es möglich, dass solche Gleichungen mehr als nur einen Satz von Lösungen zulassen (dies kann schon bei linearen Gleichungen der Fall sein). Wenn mehrere Lösungen möglich sind, ergibt sich sofort die Frage, welche Lösung die richtige ist: so könnte es ja sein, dass für eine Person a eine Schätzung $\hat{\theta}_a^{(1)}$ und eine weitere, davon verschiedene Schätzung $\hat{\theta}_a^{(2)}$ existieren, was bedeutet, dass die Person nicht beurteilt werden kann. Analog dazu kann es mehr als eine Schätzung $\hat{\kappa}_g$ für ein Item I_g geben.

   Es läßt sich zeigen, dass unter einer bestimmten Randbedingung die Parameterschätzungen für das Rasch-Modell eindeutig sind. Ob diese Bedingung erfüllt ist, läßt sich anhand der Datenmatrix feststellen. Man muß prüfen, ob sich die Items in zwei Klassen $\mathcal{I}_1$ und $\mathcal{I}_2$ aufteilen lassen und, korrespondierend dazu, die Probanden ebenfalls in zwei Klassen $\mathcal{K}_1$ und $\mathcal{K}_2$ derart, dass alle Probanden der Klasse $\mathcal{K}_1$ alle Items der Klasse $\mathcal{I}_1$ lösen, und alle Probanden der Klasse $\mathcal{K}_2$ *kein* Item der Itemklasse $\mathcal{I}_2$ lösen. Sind Aufteilungen dieser Art *nicht* möglich, sind die Schätzungen der Parameter eindeutig. (Rost (2004), p. 317)

4.5 Der Test des Modells

4.5.1 Allgemeines Vorgehen

Über die Gleichungen (4.4.3) und (4.4.4) lassen sich die Personen- sowie die Itemparameter schätzen. Die Interpretation der geschätzten Werte ist natürlich nur

sinnvoll, *wenn* das Rasch-Modell auch tatsächlich angemessen ist. Ob es überhaupt angemessen ist, muß allerdings entschieden werden. Denn es liegen dem Modell ja starke Annahmen zugrunde: (i) die Items müssen homogen sein, d.h. sie sollen nur ein Merkmal erfassen, (ii) die Itemfunktion muß durch die logistische Funktion gegeben sein, und (iii) muß alle Items dieselbe Trennschärfe haben. Keine dieser Annahmen muß den tatsächlichen Gegebenheiten entsprechen. Die Frage ist also, wie geprüft werden kann, ob das Rasch-Modell überhaupt adäquat ist.

Die Homogenität läßt sich sicherlich faktorenanalytisch untersuchen. Es sollte eine Rotation der Achsen derart geben, dass die Items auf einer Achse liegen und nur vernachlässigbare Ladungen auf zweiten oder gar dritten Achsen haben. In der Tat ist der Weg über eine Faktorenanalyse ein Weg, Items zu erhalten, von denen angenommen werden kann, dass sie nur ein Merkmal messen. Die Frage ist natürlich, ob das Merkmal, dass durch eine solche Achse repräsentiert wird, auch dasjenige Merkmal ist, das gemessen werden soll.

Darüber hinaus muß geprüft werden, ob die Itemfunktion durch eine logistische Funktion repräsentiert werden kann. Es gibt viele Funktionen, deren Verlauf dem der logistischen Funktion ähnlich ist, so dass dann, wenn eine logistische Funktion an eine empirische Itemfunktion angepasst werden kann, noch kein Beweis vorliegt, ob die "wahre" Funktion tatsächlich die logistische Funktion ist. Auf jeden Fall müssen die Itemfunktionen für verschiedene Items durch Parallelverschiebung ineinander überführbar sein, denn alle Itemfunktionen müssen ja parallel sein.

Gleichheit der κ_g in verschiedenen Subpopulationen: Gemäß der Gleichung (4.3.13) ist die Differenz der Schwierigkeits- oder Itemparameter unabhängig von den Personenparametern θ. Bestimmt man also die κ_g in verschiedenen Subpopulationen (unter der Nebenbedingung $\sum_g \kappa_g = 0$), so sollten sie gleich groß sein. Trägt man also die κ_g-Werte aus Subpopulation I auf der x-Achse und die aus der Subpopulation II auf der y-Achse eines Koordinatensystems ab, so sollten die Punkte $(\kappa_g(I), \kappa_g(II))$ auf oder in unmittelbarer Nachbarschaft einer Regressionsgerade mit der Steigung 1 und einer additiven Konstanten gleich Null liegen.

Gleichheit von Wahrscheinlichkeitsquotienten: Es läßt sich zeigen, dass die Differenzen $\kappa_g - \kappa_{g'}$ durch einen Quotienten von Wahrscheinlichkeiten definiert sind, also

$$\kappa_g - \kappa_{g'} = \log \left[\frac{P(X_g = 1 \cap X_{g'} = 0)}{P(X_g = 0 \cap X_{g'} = 1)} \right]. \tag{4.5.1}$$

Da die Differenzen $\kappa_g - \kappa_{g'}$ populationsunabhängig sind, müssen die Quotienten auf der rechten Seite, wie sie in verschiedenen (Sub-)Populationen bestimmt

werden können, gleich sein, d.h. es muß bei Geltung des Rasch-Modells

$$\frac{P^{(I)}(X_g = 1 \cap X_{g'} = 0)}{P^{(I)}(X_g = 0 \cap X_{g'} = 1)} = \frac{P^{(II)}(X_g = 1 \cap X_{g'} = 0)}{P^{(II)}(X_g = 0 \cap X_{g'} = 1)}, \tag{4.5.2}$$

gelten, wobei (I) und (II) die beiden Subpopulationen anzeigen.

Die Gleichung (4.5.2) ergibt sich direkt aus der Gleichung (4.3.13), Seite 216, also

$$\log\left(\frac{P(X_{ag} = 1)}{P(X_{ag} = 0)}\right) - \log\left(\frac{P(X_{g'a} = 1)}{P(X_{g'a} = 0)}\right) = \theta_a - \theta_a - \kappa_g + \kappa_{g'} = \kappa_{g'} - \kappa_g.$$

Denn

$$\log\left[\frac{P(X_g = 1)}{P(X_g = 0)}\right] - \log\left[\frac{P(X_{g'} = 1)}{P(X_{g'} = 0)}\right] = \log\left[\frac{P(X_g = 1)}{P(X_g = 0)}\frac{P(X_{g'} = 0)}{P(X_{g'} = 1)}\right] = \kappa_{g'} - \kappa_g,$$

und wegen der lokalen Unabhängigkeit gilt

$$\begin{aligned} P(X_g = 1)P(X_{g'} = 0) &= P(X_g = 1 \cap X_{g'} = 0) \\ P(X_g = 0)P(X_{g'} = 1) &= P(X_g = 0 \cap X_{g'} = 1). \end{aligned}$$

□

4.5.2 Genauigkeit der Personenparameter

Hat man den Parameter θ_a einer Person $a \in \mathcal{P}$ geschätzt, so hat man eben eine Schätzung $\hat{\theta}_a$, die aber im Allgemeinen nicht identisch it θ_a ist; als Folge von "Messfehlern" wird $\hat{\theta}_a$ mehr oder weniger von θ_a abweichen. Die Frage ist, ob sich wenigstens ein Intervall angeben läßt, in dem der wahre Wert θ_a mit einer hinreichend großen Wahrscheinlichkeit, etwa $p = .95$, liegt.

Die Frage geht dann über in die, wie die Schätzungen $\hat{\theta}_a$ denn verteilt sind. Werden die θ_a mit der Maximum-Likelihood-Methode geschätzt, so ist bekannt, dass die Schätzungen *asymptotisch normalverteilt* sind, wobei unter bestimmten Randbedingungen $\mathbb{E}(\hat{\theta}_a) = \theta_a$ ist, d.h. unter bestimmten Bedingungen ist (wenigstens) der Erwartungswert der Schätzungen gleich dem wahren Wert. Kennt man nun noch die Varianz dieser Verteilungen von Schätzungen, kann man ein Konfidenzintervall für den wahren Wert angeben. Das Konfidenzintervall ist dann durch

$$\text{Konfidenzintervall } = \hat{\theta}_a \pm z_\alpha \sqrt{\mathbb{V}(\hat{\theta}_a)} \tag{4.5.3}$$

gegeben. Die Frage ist, wie $\mathbb{V}(\theta_a)$ zu berechen ist. In (4.4.2), Seite 217, wurde der allgemeine Ausdruck für die Likelihood-Funktion gegeben. Da $\mathcal{L}$ von den Daten abhängt, ist $\mathcal{L}$ eine zufällige Veränderliche. Dann sind auch die Ableitungen (Differentialquotienten) von $\mathcal{L}$ zufällige Veränderliche. Es läßt sich nun zeigen (Kendall & Stuart, 1973, p. 44 ff), dass die Information über den zuschätzenden

Parameter – hier also θ_a, im Erwartungswert der zweiten partiellen Ableitung von $\mathcal{L}$ enthalten ist, d.h. es gilt

$$I(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right). \tag{4.5.4}$$

$I(\theta)$ ist die Information über θ in den Daten. Weiter läßt sich zeigen, dass die Varianz $\mathbb{V}(\hat{\theta}_a)$ gerade durch den Reziprokwert der Wurzel aus $I(\theta)$ gegeben:

$$\mathbb{V}(\hat{\theta}_a) = \frac{1}{\sqrt{I(\theta_a)}}. \tag{4.5.5}$$

Die Größe $I(\theta_a)$ und damit $\mathbb{V}(\hat{\theta}_a)$ wird mit den Schätzungen berechnet, so dass explizite Formeln für $I(\theta_a)$ hier nicht angegeben werden müssen.

4.6 Mehrkategoriale und mehrdimensionale Verallgemeinerungen

4.6.1 Raschs multikategoriales Modell

Rasch (1961) betrachtete den Fall, dass Items in Bezug auf eine Anzahl vorgegebener Kategorien beantwortet werden. Rost (2004) illustriert das Modell anhand eines Fragebogens zur Einstellung zur Umwelt, bei dem vier Antwortkategorien vorgegeben werden: (i) 0: Habe ich schon getan bzw. tue ich bereits, (ii) 1: Kann ich mir gut vorstellen, (iii) 2: Würde ich tun, wenn geeignete Bedingungen geschaffen würden, (iv) 4: Halte ich für ungeeignet, um die Umwelt zu schützen. Ein mögliches Item wäre: 'Ich fahre mit dem Fahrrad zur Arbeit'. Die Probanden kreuzen dann diejenige Kategorie an, die ihrer Einstellung am ehesten entspricht. Wesentlich bei diesem Modell ist, dass die Kategorien für alle Items gelten. Eine befragte Person muß mindestens eine der Kategorien angeben. Für jede der befragten Personen können dann die Häufigkeiten bestimmt werden, mit denen die einzelnen Kategorien angekreuzt wurden, d.h. es können die Summensscores für jede Person a bestimmt werden, die dann als *Vektor*

$$\vec{r}_a = (r_{a0}, r_{a1}, r_{a2}, r_{a3})' \tag{4.6.1}$$

angeschrieben werden können: r_{ak}, $k = 0, 1, 2, 3$ ist die Häufigkeit, mit der die k-te Kategorie angegeben wurde. Man kann fragen, wie viele mögliche Vektoren es überhaupt geben kann. Dies ist die Frage nach der Anzahl der möglichen Verteilungen der n Items auf K Kategorien. So kann es sein, dass eine Person alle Items in genau eine Kategorie sortiert, etwa die erste, so dass der Antwortvektor

$$\vec{r} = \vec{r}_a = (n, 0, 0, 0)'$$

resultiert. Sortiert sie alle Items in die zweite Kategorie, ergibt sich der Antwortvektor

$$\vec{r} = \vec{r}_a = (0, n, 0, 0)';$$

es gibt gerade K Vektoren, die eine Verteilung angeben, bei der alle Items in genau eine Kategorie sortiert wurden. Alternativ können $n-1$ Items in eine Kategorie und das restliche Item in irgendeine der übrigen Kategorien sortiert werden. Werden die $n-1$ Items in die erste Kategorie eingeordnet, so gibt es noch $K-1$ Möglichkeiten für das restliche Item. Natürlich kann es sein, dass $n-1$ Items in die zweite Kategorie sortiert werden und das restliche wieder einer der übrigen Kategorien zugeordnet wird. Insgesamt gibt es also $K(K-1)$ Vektoren, bei denen $n-1$ Items in eine K Möglichkeiten) und ein Item in eine der übrigen $(K-1)$ Kategorien sortiert werden. Formal gleicht die Frage, wie man n Items auf K Kategorien aufteilen kann, der Frage, auf wie viele Möglichkeiten man n Bälle auf K Zellen aufteilen kann. Man kann zeigen, dass die Anzahl M dieser Möglichkeiten durch

$$M = \binom{n+K-1}{K} \tag{4.6.2}$$

gegeben[55]. Für den Fall $n = 5$ und $K = 4$ Kategorien erhält man[56]

$$M = \binom{5+4-1}{4} = \frac{8 \cdot 7 \cdot 6 \cdot 5}{1 \cdot 2 \cdot 3 \cdot 4} = 14 \cdot 5 = 70.$$

mögliche Vektoren $\vec{r}$. Die Anwortvektoren, die man für eine gegebene Stichprobe von Personen erhält, ist eine Teilmenge der Menge der möglichen Vektoren $\vec{r}$. Man kann dann für die Stichprobe die Häufigkeit, mit der bestimmte Vektoren auftreten, bestimmen (vergl. Rost (2004), p. 191):

$\vec{r}$				$n(\vec{r})$
4	1	0	0	53
4	0	1	0	35
4	0	0	1	6
1	4	0	0	15
1	0	4	0	2
1	0	0	4	2
0	4	1	0	19
etc				

(4.6.3)

Die Annahme ist, dass die Summe, die eine Person für eine bestimmte Kategorie liefert, die Ausprägung des Merkmals, dass diese Kategorie repräsentiert, abbildet. Man muß dabei aber berücksichtigen, dass für jede Person a die Beziehung

$$\sum_{k=1}^{K} r_{ak} = n \ = \text{Anzahl der Items} \tag{4.6.4}$$

[55]Eine Herleitung dieser Formel findet man in Mortensen, U. *Wahrscheinlichkeitstheorie*, p. 15, Uwe-Mortensen.de, Skripten. Eine alternative und elegante Herleitung findet man auch in Feller (1968), p. 38

[56]Rost (2004), p. 191 gibt hier ohne weitere Begründung die Anzahl 56 an; vermutlich hat er versehentlich $4 \cdot 14 = 56$ gerechnet.

gilt. Je häufiger eine Person eine bestimmte Kategorie angibt, desto weniger häufig kann sie andere Kategorien ankreuzen. Damit sind die Summenscores einer Person nicht unabhängig voneinander: sind bei einer Person $K-1$ Summenscores bereits gegeben, so liegt der K-te Score damit fest. Scores mit dieser Eigenschaft haben einen besonderen Namen:sie heißen *ipsative Scores* oder *ipsative Messwerte*. Ipsative Messwerte reflektieren die Ausprägung eines Merkmals einer Person nur relativ zu den Ausprägungen anderer Merkmale *bei dieser Person*. Im Unterschied zu den ipsativen Messwerten reflektieren die *normativen Messwerte* die Ausprägung eines Merkmals bei einer Person relativ zu den Ausprägungen dieses Merkmals *bei anderen Personen*. Für ipsative Messungen gilt, dass sie negativ miteinander korrelieren; für K Kategorien erhält man die Korrelation

$$\rho_{\text{ipsativ}} = \frac{1}{K-1}. \tag{4.6.5}$$

Will man die Korrelationen zwischen den Kategorien bestimmen, so geht die Korrelation (4.6.5) verzerrend ein, erschwert also die inhaltliche Diskussion der Korrelationen (über die Personen) zwischen den Kategorien. Diese Verzerrung vererbt sich auch auf die Korrelationen mit externen Variablen, wie sie bei der Bestimmung der Validität berechnet werden.

Rost (204) betrachtet nun die die Wahrscheinlichkeit, mit der eine Person a bei einem ItemI_g gerade die k-te Kategorie wählt. Er nimmt an, dass sie durch die Differenz $\theta_{ak} - \kappa_{gk}$ bestimmt wird, wobei θ_{ak} die Neigung der a-ten Person, die k-te Kategorie zu wählen und κ_{gk} die k-te Kategorie beim g-ten Item charakterisiert. Man kann κ_{gk} als *kategorienspezifische Schwierigkeit* bezeichnen. Es sei $X_{ag} = k$, $k = 1, \ldots, K$ eine Indikatorvariable ist, die angibt, welche Kategorie die Person a beim g-ten Item angegeben hat. Rost macht nun die Annahme, dass die Wahrscheinlichkeit $P(X_{ag} = k)$ durch den Ausdruck

$$P(X_{ag} = k|\theta_{ak}, \kappa_{gk}) = A\, e^{\theta_{ak} - \kappa_{gk}} \tag{4.6.6}$$

gegeben ist, wobei A eine Konstante ist derart, dass

$$\sum_{k=1}^{K} P(X_{ag} = k|\theta_{ak}, \kappa_{gk}) = 1, \tag{4.6.7}$$

ist, woraus folgt, dass

$$A = \frac{1}{\sum_{k=1}^{K} e^{\theta_{ak} - \kappa_{gk}}} \tag{4.6.8}$$

gelten muß. Man erhält dann

$$P(X_{ag} = k|\theta_{ak}, \kappa_{gk}) = \frac{e^{\theta_{ak} - \kappa_{gk}}}{\sum_{k=1}^{K} e^{\theta_{ak} - \kappa_{gk}}}. \tag{4.6.9}$$

Diese Gleichung definiert das mehrdimensionale Raschmodell.

Kommentare:

1. Zunächst muß festgestellt werden, dass sich das Modell offenbar nicht auf die Messung eines mehrdimensionalen Merkmals bezieht, sondern nur auf ein nicht-dichotom, also multikategorial erfasstes Merkmal. Als solches stellt es allerdings eine wichtige Verallgemeinerung des Rasch-Modells für dichotome Items dar, da Items mit mehr als nur einer Antwortkategorie z.B. in Fragebögen häufig vorkommen und das Modell eine Rasch-Analyse entsprechender Daten erlaubt.

2. Die Definition des Parameters θ_{ak} ist ein wenig unklar. Erklärt werden soll ja die Wahrscheinlichkeit, dass beim g-ten Item die k-te Kategorie gewählt wird. Diese Wahrscheinlichkeit soll eine Funktion der Differenz von Person- und Itemparameter sein. Dazu müsste aber θ_a nicht auch noch als für die k-te Kategorie spezifisch angenommen werden: $\theta_a - \kappa_{gk}$ drückt ja bereits die Relation zwischen der Position der a-ten Person auf das g-te Item aus. Die Situation ist vergleichbar mit der beim Thurstoneschen Modell des Paarvergleichs, wo in Gleichung (2.6.39) (Seite 59) der Vergleich zweier Items von der Position θ_a der a-ten Person auf der Skala abhängig gemacht wird.

   Wenn man aber gleichwohl einen für die Person a und die Kategorie k spezifischen Parameter θ_{ak} einführt, so postuliert man damit die Existenz einer Tendenz, eine bestimmte Kategorie anzukreuzen, unabhängig vom Item (denn θ_{ak} ist ja nicht auch noch für ein Item charakteristisch). Denn angenommen, die k-te Kategorie repräsentiert eine neutrale Position, wie etwa "Ich weiß es nicht": bekommt man ein Item wie "Sollen die Steuern verringert werden?" vorgelegt und repräsentieren die Kategorien ein Spektrum zwischen scharfer Ablehnung und enthusiastischer Zustimmung, so kann die neutrale Kategorie die von den meisten favorisierte Kategorie sein, denn kein Mensch außer einigen Ideologen mit ausgeprägter Fähigkeit zur Autosuggestion kennt die wirklich "wahre" Antwort. □

Die Annahme (4.6.6) ist natürlich motiviert durch den Wunsch, das Rasch-Modell für den 1-dimensionalen Fall zu verallgemeinern, eine *inhaltliche* Überlegung liegt dieser Annahme nicht zugrunde, wenn man von messtheoretischen Betrachtungen absieht. (4.6.9) ist ein mögliches Modell für das Antwortverhalten, – ob es im konkreten Fall zutrifft oder nicht ist eine empirische Frage.

Die Parameter im 1-dimensionalen Rasch-Modell sind nur eindeutig bis auf Translationen, d.h. der Nullpunkt etwa der θ-Skala ist willkürlich. Deswegen kann insbesondere

$$\sum_{g=1}^{n} \theta_{ak} = 0 \text{ für alle } a \in \mathcal{P} \tag{4.6.10}$$

gesetzt werden, so dass nur $K - 1$ Parameter für jede Person geschätzt werden müssen. Die Itemparameter κ_{gk} können wegen der genannten Eindeutigkeit nur

bis auf Translation ebenfalls normiert werden, indem man

$$\sum_{g=1}^{n} \kappa_{gk} = 0 \tag{4.6.11}$$

setzt. Hat man etwa n Items mit je K Kategorien, so sind, da ja noch (4.6.7) gelten soll, $(n-1)(K-1)$ freie Parameter κ_{gk} zu schätzen.

Spezifische Objektivität: Die wesentliche Eigenschaft des 1-dimensionalen Rasch-Modells ist die spezifische Objektivität: dieses Merkmal erlaubt es, die Personen- und die Itemparameter unabhängig voneinander zu schätzen, was wiederum bedeutet, dass man die Schwierigkeit unabhängig von der Stichprobe von Personen und die Fähigkeit der Personen unabhängig von den Items beurteilen kann. Die Frage ist nun, ob sich diese Eigenschaft auf das mehrdimensionale Rasch-Modell überträgt.

Dazu betrachtet man zunächst die Definition von $P(X_{gk} = k)$ in (4.6.9). $P(X_{gk} = k)$ hängt nicht nur von θ_{ak} und κ_{gk} ab, sondern wegen des Nenners auch von allen anderen Parametern. Dies impliziert, dass der Vergleich irgend zweier Itemparameter nicht mehr spezifisch objektiv ist. Es läßt sich aber zeigen, dass die *Vergleiche von Differenzen von Itemparametern* spezifisch objektiv sind.

Es bleibt zu fragen, wie bei diesem Modell denn Itemfunktionen definiert sind und wie die unbekannten Parameter geschätzt werden können. Man kann sagen, dass jeder Parameter θ_{ak} für ein Merkmal steht, das eben durch die k-te Kategorie repräsentiert wird. Dementsprechend sind K Itemfunktionen zu betrachten. Der Parameter für eine Itemfunktion ist dann κ_{gk}, die Schwierigkeit der k-ten Kategorie beim g-ten Item. Je größer der Wert von κ_{gk}, desto weiter rechts wird die Itemfunktion positioniert.

Die Parameter werden über die Maximum-Likelihood-Methode geschätzt. Die Likelihood-Funktion ist

$$\begin{aligned} L &= \prod_{a=1}^{m} \prod_{g=1}^{m} \frac{\exp(\theta_{ak} - \kappa_{gk})}{\sum_{j=1}^{K} \exp(\theta_{gj} - \kappa_{jk})} \\ &= \frac{\exp\left[\sum_a \sum_g r_{ak}\theta_{ak} - \sum_i n_{gk}\kappa_{gk}\right]}{\prod_a \prod_g \sum_k \exp(\theta_{ak} - \kappa_{gk})} \end{aligned} \tag{4.6.12}$$

Man sieht, dass die Likelihood L – abgesehen von den zu schätzenden Parametern – nur von den Häufigkeiten r_{ak} und n_{gk} abhängt. Es ist also nicht wichtig, zu wissen, bei welchem Item die k-te Kategorie gewählt wurde, *falls das Modell gilt*. Die unterschiedlichen Antwortmuster sind also für die Interpretation der Ergebnisse nicht von Bedeutung.

Diese Parameterschätzung setzt voraus, dass jede Person für jedes Item auch mindestens eine Kategorie angekreuzt hat. Kommt es vor, dass Personen bei bestimmten Items keine Kategorie angekreuzt haben, so kann man, wenn es einem

nur auf die Itemparameter ankommt, diese Person einfach weglassen. Nur ist dies keine gute Strategie, wenn die Anzahl der Personen, die bei einem Item keine Kategorie gewählt haben, zu groß wird. Ein praktikabler Ausweg besteht darin, in solchen Fällen einfach einen kleinen Wert für r_{gk} einzusetzen, etwa $r_{gk} = .01$.

4.6.2 Faktorenanalytische Ansätze

In Abschnitt 2.6.9 wurden Elemente der Theorie der Faktorenanalyse vorgestellt. Die Leistung in einem Test wird dabei als eine lineare Kombination von verschiedenen Komponenten, die bei der Beantwortung von Fragen bzw. beim Lösen von Aufgaben benötigt werden, dargestellt. Angenommen, es werden r solche latenten Komponenten benötigt. Eine Person a ist dann in Bezug auf diese Komponenten durch Faktorwerte $q_{a1}, \ldots, q_{ar}$ charakterisiert, wobei eine Komponente q_{ak} das Ausmaß der k-ten Komponente repräsentiert, über die die Person a verfügt. Das Item I_g wird durch die "Ladungen" $\alpha_{g1}, \ldots, \alpha_{gr}$ repräsentiert, wobei α_{gk} das Ausmaß angibt, mit dem das Item I_g die k-te Dimension – also nicht die k-te Kategorie, wie im vorangegangenen Abschnitt – erfaßt. Wie in Gleichung (4.6.13); Seite 228, angeschrieben, kann man θ_a mit den Faktorwerten in Beziehung setzen.

$$\theta_{ak} = q_{ak}, \quad k = 1, \ldots, r \tag{4.6.13}$$

Der standardisierte Score z_{ag} der a-ten Person bei der g-ten Aufgabe ist dann, entsprechend (**??**) auf Seite **??**

$$z_{ag} = \alpha_{1g}\theta_{a1} + \cdots + \alpha_{rg}\theta_{ar} = \sum_{k=1}^{r} \alpha_{kg}\theta_{ak}, \tag{4.6.14}$$

wobei ein Fehler vernachlässigt wurde.

Hat man nun dichotome Items, so kann man vom Ansatz der logistischen Regression Gebrauch machen; dort werden ja die Logits zu den unabhängigen Variablen in Beziehung gesetzt: Ist p die Wahrscheinlichkeit einer Antwort in Abhängigkeit von den Prädiktoren $X_1, \ldots, X_n$, so ist

$$\frac{p}{1-p} = b_0 + b_1 X_1 + \cdots + b_n X_n$$

setzen, was natürlich

$$p = \frac{\exp(b_0 + b_1 X_1 + \cdots + b_n X_n)}{1 + \exp(b_0 + b_1 X_1 + \cdots + b_n X_n)}$$

bedeutet. Die Regressionsparameter können als ML-Schätzungen bestimmt werden. Setzt man nun

$$p = P(X_{ag} = 1 | \theta_{a1}, \ldots, \theta_{ar}, \kappa_g)$$

und ersetzt $b_0 + b_1X_1 + \cdots + b_nX_n$ durch $z_{ag} - \kappa_g$, so erhält man wegen (4.6.14)

$$P(X_{ag} = 1|\theta_{a1}, \ldots, \theta_{ar}, \kappa_g) = \frac{\exp\left(\sum_{k=1}^{r} \alpha_{kg}\theta_{ak} - \kappa_g\right)}{1 + \exp\left(\sum_{k=1}^{r} \alpha_{kg}\theta_{ak} - \kappa_g\right)}. \tag{4.6.15}$$

Damit hat man eine mögliche Verallgemeinerung des Rasch-Modells auf r Dimensionen. Ein ähnlicher Ansatz wird auch in McDonald (1999), p. 310, diskutiert.

Interpretation: Es wird die Kombination verschiedener, von einander unabhängiger Fähigkeiten angenommen, die kompensatorisch eingesetzt werden können, um eine Aufgabe zu lösen. Nimmt man der Einfachheit halber an, dass einerseits eine numerische Fähigkeit (nF) – repräsentiert durch θ_1 – und andererseits räumliches Vorstellungsvermögen (rV) – repräsentiert durch θ_2 – benötigt werden, soll also $\theta = \alpha_1\theta_1 + \alpha_2\theta_2$ gelten. Die Wahrscheinlichkeit, eine Aufgabe zu lösen, hängt nur ab von der Differenz $\theta - \kappa_g$. Für fixen Wert θ ergibt sich ein "trade-off"

$$\theta_2 = \frac{1}{\alpha_2}(\theta - \alpha_1\theta_1); \tag{4.6.16}$$

Für alle Personen mit gleichem θ-Wert ist die Wahrscheinlichkeit einer Aufgabenlösung gleich groß, auch wenn sie sich hinsichtlich ihrer θ_1- und θ_2-Werte unterscheiden, so lange diese Werte nur der Bedingung (4.6.16) genügen.

Ein solches Modell kann sicherlich nur für bestimmte Aufgaben gelten. Es gibt Aufgaben, die sich nicht nur mit geometrischen Methoden lösen lassen, und andererseits kann es hilfreich für das Lösen algebraischer Aufgaben sein, geometrische Betrachtungen anzustellen. Aufgaben, für die ein Trade-off wie in (4.6.16) gilt, dürften eine sehr kleine Klasse bilden. Plausibler ist vermutlich ein Ansatz, bei dem gewisse Minimalfähigkeiten zur Lösung einer Aufgabe gefordert werden, ein Trade-off wie in (4.6.16) also nur sehr partiell gilt. Man mag einwenden, dass (4.6.16) zwar eine formale Implikation des faktorenanalytischen Ansatzes sei, die aber psychologisch gar nicht gemeint sei; (4.6.16) sei nur eine irrelevante Kollateralimplikation. Mit dem faktorenanalytischen Ansatz sei nur gemeint, dass für die *gegebenen* Aufgaben das Modell

$$z_{ag} = \sum_{k=1}^{r} \alpha_{kg}\theta_{ak}$$

gelten soll. Für eine gegebene Aufgabe betrüge der Anteil der notwendigen numerischen Fähigkeit gerade α_1 und der für die räumliche Vorstellung α_2, und für eine gegebene Person seien die Fähigkeiten gerade θ_1 und θ_2. Eine solche Darstellung der Scores z_{ag} lässt sich ja schon wegen der allgemeinen Anwendbarkeit der Singularwertzerlegung stets finden (vergl. die Fußnote zu Gleichung (2.6.63), Seite 65). Aber für sich genommen ist die Repräsentation dann trivial, zumal man nicht weiß, wie sie zu deuten ist. Interessant wird sie erst, wenn man sie als Modell ernst nimmt und etwa fordert – wie es bei faktorenanalytischen Modellen ja geschieht –, dass die α_j bzw. α_{gj} zu verschiedenen Fähigkeitsanteilen korrespondieren, die von jeder Person verlangt werden, wenn sie die Aufgabe lösen will.

Will man dann das Lösungsverhalten verschiedener Personen diskutieren, deren Fähigkeitsprofile verschieden sind, aber die Aufgaben mit gleicher Wahrscheinlichkeit lösen, so landet man notgedrungen doch wieder beim Trade-off-Modell (4.6.16).

Takane & de Leeuw (1987) weisen eine Äquivalenzbeziehung zwischen den *marginal likelihoods* für das 2-Parameter Ogiven-Modell einer IRT und der Faktorenanalyse für dichotome Items nach und verallgemeinern ihr Resultat für den Fall mehrkategorialer Items.

Das Modell von Muraki & Carlson: Einen weiteren Ansatz findet man bei Muraki & Carlson (1995). Der Personenparameter wird zu einem Vektor $\vec{\theta_a} = (\theta_{a1}, \theta_{a2}, \ldots, \theta_{ar})'$. Für das g-te Item I_g existieren Faktorladungen $\alpha = (\alpha_{g1}, \alpha_{g2}, \ldots, \alpha_{gr})$. Dann wird eine Antwortvariable $y_{ga} = \alpha_{g1}\theta_{a1} + \cdots + \alpha_{gr}\theta_{gr} + \varepsilon_{ga}$ definiert, wobei ε_{ga} eine Fehlervariable ist, von der angenommen wird, dass sie standardnormalverteilt ist. Die übliche Faktorenanalyse geht von Messungen y_{ga} aus, die direkt den Antwortprozess repräsentieren. Bei kategorialen Variablen ist aber die Antwortvariable y_{ga} latent, von ihrem Wert hängt die Wahl der Kategorie ab. Es sei w_g eine Indikatorvariable: $w_g = j$ genau dann, wenn die Antwort auf das Item I_g in die j-te Kategorie fällt. Eine Antwort fällt in die j-te Kategorie, wenn

$$\gamma_{g,j-1} \leq y_{ga} < \gamma_{g,j}, \tag{4.6.17}$$

mit $\gamma_{g,0} = -\infty$, $\gamma_{gk} = \infty$. Die γ_{gj} sind Schwellenparameter. Die Wahrscheinlichkeit, dass eine Antwort in die j-te Kategorie fällt, ist

$$P(w_g = j|\theta_a) = \int_{\gamma_{g,j-1}}^{\gamma_{g,j}} f_{ag}(y)dy, \quad f_{ag}(y) = \frac{1}{\sigma_g\sqrt{2\pi}} \exp\left[-\frac{(y - y_{ag})^2}{2\sigma_g^2}\right]. \tag{4.6.18}$$

Die Antwort auf die Frage nach der Parameterschätzung ist länglich und wird hier übergangen (s. Muraki & Carlson 1995). Für die Interpretation gelten die gleichen Anmerkungen wie oben. Der Unterschied zu dem vorangegangenen Modell besteht eigentlich nur in der Wahl der Wahrscheinlichkeitsverteilung.

Whitleys Modell: Whitley (1980) diskutiert drei Modelle: (i) unabhängige Komponenten, (ii) sequentiell abhängige Komponenten, (iii) wiederholbar abhängige Komponenten. Es genügt hier, das erste Modell vorzustellen. Es repräsentiere X_{agT} die Antwort der a-ten Person auf das g-te Item. $X_T = \{0, 1\}$ signalisiert, ob ein Item korrekt beantwortet wurde oder nicht. Die Antwort hänge von K "unterliegenden" Komponenten ab, von denen jede durch eine entsprechende, nur diese Komponente erfassende Aufgabe erfasst werden könne. X_{agk} repräsentiere die Antwort auf ein Item für die k-te Komponente, und $X_k = \{0, 1\}$indiziere, ob die Antwort auf dieses Item korrekt war oder nicht. Es gelte

$$P(X_k = 1|\theta_k, b_k) = \psi(\theta_k - b_k). \tag{4.6.19}$$

Dabei ist θ_k ein Personenparameter für das k-te Merkmal, und b_k sei die Schwierigkeit des Items. Es wird lokale Unabhängigkeit angenommen, und für ψ kann insbesondere die logistische Funktion angenommen werden. Man erhält dann für die k-te Komponente wieder ein Rasch-Modell. Es wird weiter angenommen, dass die K Komponenten stochastisch unabhängig sind, gegeben die Parameter θ_k und b_k. Die Wahrscheinlichkeit, dass die Aufgabe korrekt beantwortet wird, ist dann

$$P(X_T = 1|\theta, b) = \prod_k \frac{\exp(\theta_k - b_k)}{1 + \exp(\theta_k - b_k)}, \tag{4.6.20}$$

wo jetzt explizit die logistische Funktion postuliert wurde. Die Wahrscheinlichkeit einer korrekten Antwort ist die Wahrscheinlichkeit, dass alle Komponenten in korrekter Weise zur Beantwortung der Aufgabe herangezogen wurden.

4.6.3 Das linear-logistische Modell

Suppes, Jerman & Brian (1968) versuchten, die Abhängigkeit der Wahrscheinlichkeit der Lösung arithmetischer Aufgaben von der Anzahl der Aufgabenkomponenten, die zur Lösung der Aufgabe zu bewältigen waren, zu modellieren. Ist p_g die Wahrscheinlichkeit, die g-te Aufgabe zu lösen, so soll

$$\text{Logit}(p_g) = \log \frac{p_g}{1 - p_g} = \sum_j c_{gj}\kappa_{gj} + c \tag{4.6.21}$$

gelten, wobei η_j ein "Leichtigkeitsparameter" für die j-te Komponente der g-ten Aufgabe ist; dieser Parameter ist unabhängig von der speziellen Aufgabe I_g, die entsprechende Komponente hat also die gleiche Leichtigkeit in allen Aufgaben. c_{gj} ist die Anzahl der Komponenten in I_g mit der Leichtigkeit κ_{gj}, und c ist ein freier Parameter. Scheiblechner (1972) hat das Modell auf die Durchführung einfacher logischer Operationen angewendet und dabei einen zusätzlichen Parameter ξ_a, der die Fähigkeit der a-ten Person repräsentiert, eingeführt, also $\xi_a = \theta_a$ in der in diesem Skript verwendeten Terminologie. Damit konstruierte er einen Übergang zum Rasch-Modell, wenn man

$$\kappa_g = \sum_j c_{gj}\kappa_{gj} + c \tag{4.6.22}$$

setzt. Das Modell für das Lösen einer Aufgabe wird dann

$$P(X_{ag} = 1|\theta_a, \kappa_g) = \frac{\exp[\theta_a - (\sum_j c_{gj}\kappa_{gj} + c)]}{1 + \exp[\theta_a - (\sum_j c_{gj}\kappa_{gj} + c)]}. \tag{4.6.23}$$

Fischer (1972) diskutiert dieses Modell unter dem Namen *linear-logistisches Modell*.

Interpretation: Zunächst sei angemerkt, dass eine notwendige Voraussetzung des linear-logistischen Modells die Gültigkeit des Rasch-Modells ist. Das linear-logistische Modell kann als eine weitere Spezifikation des Rasch-Modells angesehen werden, da angenommen wird, dass der Schwierigkeitsparameter κ_g als Summe von Schwierigkeiten, die sich auf Komponenten der Aufgabe beziehen, dargestellt werden kann.

Eine naheliegende Frage ist, aus welchen Annahmen über das Lösungsverhalten die Darstellung (4.6.22) von κ_g gefolgert werden kann. Wie es aussieht, sind solche Annahmen nicht formuliert worden. Der Ansatz (4.6.21) von Suppes et al. scheint ein ad-hoc-Ansatz zu sein: die Wahrscheinlichkeit der Lösung einer Aufgabe wird als logistische Regression auf die Schwierigkeiten der Komponenten dargestellt, die hier die Rolle von Prädiktoren einnehmen. Scheiblechner (1972) hat nur die Fähigkeit θ einer Testperson als weiteren Prädiktor hinzugefügt. Der Ansatz folgt also nicht aus psychologischen Betrachtungen über das Lösungsverhalten, sondern ist einfach ein Versuch, die Abhängigkeit der Lösungswahrscheinlichkeit von den Teilschwierigkeiten und der Fähigkeit darzustellen, ganz analog zu der Art und Weise, wie man versucht, den Effekt verschiedener Risikofaktoren auf die Wahrscheinlichkeit, in einen Autounfall verwickelt zu werden, oder sich bei einem Krankenhausbesuch zu infizieren. Korrespondierend zu den Betrachtungen in Abschnitt 2.6.3 dieses Skriptums kann man zeigen, dass der Ansatz der logistischen Regression auf die Annahme einer logistisch verteilten "unterliegenden Variablen" zurückgeführt werden kann[57]. Im hier gegebenen Zusammenhang würde man

$$P(X_g = 1|\theta, \kappa_g) = P(\eta > \kappa|\theta, \kappa_g) = \frac{1}{1 + \exp\left(\frac{\kappa_g - \theta}{\sigma}\frac{\pi}{\sqrt{3}}\right)}, \tag{4.6.24}$$

wobei η wieder die Fähigkeit einer Person zum Zeitpunkt des Testens ist. Den Faktor $\pi/\sqrt{3}$ kann man etwa in den Parameter σ absorbieren, indem man $\sigma^* = \sigma\sqrt{3}/\pi$ setzt, und weiter

$$\frac{\kappa_g - \theta}{\sigma^*} = -(\theta^* - \kappa_g^*) \quad \theta^* = \theta/\sigma^*,\ \kappa_g^* = \kappa_g/\sigma^*$$

setzt. Mit dieser Reparametrisierung erhält man, wenn man zur Abkürzung $p_g = P(X_g = 1|\theta, \kappa_g)$ schreibt,

$$\frac{p_g}{1 - p_g} = \theta - \kappa_g,$$

wobei θ^* und κ_g^* einfach wieder in θ und κ_g umbenannt wurden. In (4.6.24) spielt θ die Rolle eines Erwartungswertes (vergl. (2.6.6), Seite 36), und es wird angenommen, dass Größen, die die Lösungswahrscheinlichkeiten beeinflußen, auf θ einwirken. Diese Annahme entspricht dem allgemeinen Ansatz, der vielen statistischen Verfahren zugrunde liegt, man denke etwa an die Varianzanalyse, wo ja

[57] Dieser Ansatz wird explizit vorgeführt im Skript *Kategoriale Regression und loglineare Modelle*, Abschnitt 2.2.2. WWW.Uwe-Mortensen.de, Skripten.

auch postuliert wird, dass die unabhängigen Variablen auf den Erwartungs- bzw. Mittelwert der abhängigen Variablen einwirken. In Suppes et al.'s Ansatz (4.6.22) dagegen wird – nicht unplausibel – davon ausgegangen, dass die Darbietung einer Aufgabe ja nicht den Erwartungswert θ, also hier der Personenparameter, verändert, sondern dass *für gegebenen θ-Wert* der Schwierigkeitsparameter κ_g bzw. die Schwierigkeitsparameter der Komponenten der Aufgaben die Antwortwahrscheinlichkeit bestimmt. Dabei wird kein explizites Modell des Lösungsverhaltens aufgestellt. In einem solchen Modell würde sich κ_g als irgendeine, von den speziellen Modellannahmen abhängende, vermutlich komplizierte Funktion der Teilschwierigkeiten ergeben. Solche komplizierten Funktionen lassen sich aber, einem Theorem der Mathematik zufolge[58] durch geeignet gewählte Polynome beliebig genau *approximieren*. Für kleine Bereiche der unabhängigen Variablen genügt oft schon ein Polynom ersten Grades, also eine lineare Funktion, und das ist der Ansatz von Suppes et al. Potenzen höheren Grades und Produktterme (Polynome können auch in mehreren Variablen definiert werden) führen dann zu immer genaueren Approximationen. Im Rahmen von Regressionsanalysen verschwinden die nichlinearen Teile aber oft im "Rauschen", also in den Messfehlern, wie auch immer diese zustande kommen, so dass man mit einem linearen Ansatz im Allgemeinen sehr weit kommt.

In den üblichen Anwendungen der logistischen Regression werden nun die unabhängigen Variablen, von denen man vermutet, dass sie einen Einfluß auf die Wahrscheinlichkeit des Eintretens des jeweils interessierenden Ereignisses haben, explizit erhoben. So kann man die Wahrscheinlichkeit, dass eine Frau, die in einer Klinik ein Kind entbindet, sich eine Infektion zuzieht, als von den Faktoren "Kaiserschnitt" (ja oder nein), "Diabetikerin" (ja oer nein), "Übergewicht" (ja oder nein) abhängig vermuten und die Hypothese eines Einflusses dieser Faktoren über die logistische Regression testen[59]. Für jede gebärende Frau kann objektiv festgestellt werden, ob sie sich infiziert hat oder nicht und welche der genannten Faktoren auf sie zutreffen und welche nicht. Regressionsgewichte in der Nachbarschaft von 0 legen nahe, dass der korrespondierende Faktor keinen Einfluß hat.

Im linear-logistischen Modell wird die objektive Bestimmung der Werte der unabhängigen Variablen ("Prädiktoren") ersetzt durch *Annahmen* über Teilkomponenten. Es sind Annahmen über die Gewichte c_{gj}. Muß in einer Aufgabe eine bestimmte Differentiationsregel angewendet werden, wird der entsprechende c_{gj}-Wert gleich 1, sonst gleich Null gesetzt. Auf diese Weise entsteht eine $n \times r$-Matrix C von Koeffizienten $c_{gj} = \{0, 1\}$, n die Anzahl der Aufgaben, r die Maximalzahl

[58] Gemeint ist der Weierstraßsche Approximationssatz, wie man ihn in Lehrbüchern zur Differential- und Integralrechnung findet, etwa in Courant (1955), p. 289, oder Heuser (2002), p. 63. Ein Polynom (n-ten Grades) ist ein Ausdruck der Form $f(x) = a_0 + a_1x^1 + a_2x^2 + \cdots + a_nx^n$, ein Polynom ersten Grades hat dann die Form $f(x) = a_0 + a_1x$. Die Definition läßt sich auf Funktionen von mehr als einer Variablen verallgemeinern.

[59] vergl. Skript *Kategoriale Regression und loglineare Modelle*, Beispiel 17, p. 48.

von Komponenten in einer Aufgabe. Mögliche Wechselwirkungen zwischen den Komponenten werden nicht berücksichtigt. Geschätzt werden dann die κ_{gj}-Werte.

Die Interpretation läuft dann wie bei der üblichen logistischen Regression: Aus (4.6.21) folgt

$$\frac{p_g}{1-p_g} = e^{c_{g1}\kappa_{g1}+\cdots+c_{gr}\kappa_{gr}} = e^{c_{g1}\kappa_{g1}} e^{c_{g2}\kappa_{g2}} \cdots e^{c_{gr}\kappa_{gr}}.$$

Ist insbesondere $c_{hj} = 0$, so ist $\exp(c_{gj}\kappa_{gj} = 1$, d.h. die j-te Komponente hat keinen Effekt auf die Wettchance ("odds"), dass das Item gelöst wird. Für $c_{gj} \neq 0$ gibt der Faktor $\exp(c_{gj}\kappa_{gj}$ den Einfluß an, den j-te Komponente auf die Wettchance hat. Das Ausmaß des Einflusses einer Teilkomponente einer Aufgabe hängt auch von dem in (4.6.21) angenommenen Ansatz ab; nimmt man nichtlineare Komponenten und Wechselwirkungen ebenfalls an, kann man u. U. zu anderen Abschätzungen kommen.

Um eine bessere Intuition über das linear-logistische Modell zu bekommen, soll kurz auf die Arbeit von Fischer (1973) eingegangen werden. Fischer (1973) diskutierte die Lösungen von Aufgaben aus der elementaren Differentialrechnung im Rahmen des linear-logistischen Modells. Es wurden Aufgaben verschiedenen Schwierigkeitsgrades präsentiert und sowohl die θ_a- wie auch die κ_{gj} für die verschiedenen Komponenten, die bei der Lösung zu bewältigen sind, geschätzt. Eine Testaufgabe ist eine Funktionen vom Typ

$$x^3(x^2+1)^2,\ \exp[3(x^2-x-2)/2],\ \sin\left[\sqrt{\frac{1-x^3}{1+x^3}}\right],\ \text{etc.},$$

für die die erste Ableitung nach x zu finden ist. Offenbar sind die Aufgaben verschieden schwierig. Es müssen die verschiedenen Differentiationsregeln in geeigneter Verkettung durchgeführt werden. Man betrachte dazu die Differentiation der Funktion $f(x) = \sqrt{x}$. Da $\sqrt{x} = x^{1/2} = x^p$ mit $p = 1/2$ läßt sich die Differentiation auf die Differentiation der Funktion $g(x) = x^p$, $p > 0$ zurückführen: bekanntlich gilt $dg(x)/dx = g'(x) = px^{p-1}$. Um f zu differenzieren, schreibt man sie in die Form g mit $p = 1/2$ um. Dann ist

$$g'(x) = \frac{1}{2}x^{1/2-1} = \frac{1}{2}x^{-1/2} = \frac{1}{2\sqrt{x}}.$$

Man muß also (i) den Übergang $f \to g$ bewältigen, dann (ii) die Regel für die Differentiation von x^p anwenden, dann (iii) berücksichtigen, dass $x^{p-1} = 1/x^{1-p}$ ist, etc. Dabei können Vorzeichenfehler gemacht werden, der Faktor $1/2$ kann am Ende vergessen werden, etc. Ist die Funktion $f(x) = \sqrt{ax^3}$ zu differenzieren, muß zusätzlich die Kettenregel angewendet werden, wobei weitere Fehlermöglichkeiten entstehen, und die Vermeidung der Fehler kann, abgesehen von der Notwendigkeit, die Regeln zu erinnern, die Schwierigkeiten ausmachen.

Die beschriebene Zerlegung der Aufgabe in Teilaufgaben ist eine denkbare Zerlegung, – ob sie auch zwingend in dieser Form vorgenommen wird, ist eine andere Frage. Wer mehr Übung hat, sieht die Struktur gewissermaßen auf einen Blick und die einzelnen Differentiationsregeln haben keine unterschiedliche Schwierigkeit, Fehler ergeben sich allenfalls als Flüchtigkeits- oder Schreibfehler. Ob die vom Autor vorgegebenen c_{gj}-Werte also für alle Probanden gelten, ist keineswegs gesichert. Die Schätzungen der κ_{gj}-Werte müssen nicht notwendig das repräsentieren, was von ihnen angenommen wird. Als Test hat Fischer einen Plot der ebenfalls geschätzten Gesamtschwierigkeiten κ_g versus die auf der Basis der Einzelschätzungen κ_{gj} über die Summe $\hat{\kappa}_g = \sum_j c_{gj}\kappa_{gj}$ vorhergesagten κ_g-Werte durchgeführt. Zwar kann man die Hypothese einer linearen Regression von $\hat{\kappa}_g$ auf κ_g beibehalten, aber die Daten zeigen doch eine erhebliche Varianz.

Hornke & Habon (1986) betrachteten die *Konstruktion* von Testaufgaben vom Raven-Matrizen-Typ: ein Feld wird in drei Zeilen und drei Spalten aufgeteilt. Jedes Feld enthält eine Figur: einen Kreis, oder ein Dreieck, oder ein Quadrat. In der dritten Zeile fehlt eine Figur. Sie muß nach Maßgabe der Regeln, nach der die übrigen Reihen besetzt wurden, aufgefüllt werden. Das Problem besteht darin, die Regel bzw. die Regeln zu erkennen. Denn die Figuren können mit bestimmten Mustern gefüllt werden, so dass es eine Regel bezüglich der Muster und eine bezüglich der Figuren zu finden gibt. Das Problem, solche Aufgaben zu konstruieren, entspricht dem Problem, solche Aufgaben zu lösen, – aber das ist nur ein Nebenaspekt, der hier nicht weiter von Interesse ist. Die Autoren schlagen zuer Analyse der Daten aus solchen Tests das linear-logistische Modell vor. Die Begründung ist interessant: da verschiedene mentale Operationen vorgenommen müssen, um eine solche Aufgabe zu lösen, sei die Gesamtschwierigkeit einer Aufgabe gleich der Summe der Teilschwierigkeiten für die einzelnen Operationen. Es wird also nicht der oben vorgeführte Weg über die logistische Regression, wie er durch Suppes et al.s Ansatz nahegelegt wird, gewählt, sondern einfach nur eine additive Komposition der Teilschwierigkeiten angenommen. Eine Begründung für diese Annahme findet man nicht.

Betrachtet man die Fischerschen und die Hornke et al.schen Untersuchungen, so erscheint fraglich, ob der unterstellte Anspruch, über die Parameterschätzungen im Rahmen des linear-logistischen Modells tatsächlich Aufschluß über die Komponenten des jeweiligen Problemlöseprozesses bekommen zu haben. Es werden einerseits keinerlei Annahmen über die Art und Weise, in der die einzelnen Operationen vorgenommen werden, gemacht, andererseits werden ihnen individuelle Schwierigkeiten zugeordnet. Man kann sich fragen, unter welchen Bedingungen sich eine solche Additivität ergibt. Man könnte annehmen, dass die Regelanwendungen unabhängig voneinander korrekt oder inkorrekt erfolgen. Nimmt man der Einfachheit halber einmal an, dass nur zwei Regeln – und jede nur einmal bei einer Aufgabe – angewendet werden müssen, so sind zwei latente Variable η_1 und η_2 zu betrachten. Die Aufgabe wird korrekt gelöst, wenn $\eta_1 > \kappa_1$ und $\eta_2 > \kappa_2$

ist. Die Wahrscheinlichkeit dieses Ereignisses ist durch das Integral

$$P(X_g = 1|\theta, \kappa_1, \kappa_2) = \int_{\kappa_1}^{\infty} \int_{\kappa_2}^{\infty} f(\eta_1, \eta_2) d\eta_1 d\eta_2 \tag{4.6.25}$$

gegeben, wobei $f(\eta_1, \eta_2)$ die gemeinsame Dichte von η_1 und η_2 ist. Es ist eine interessante Aufgabe, diejenige Dichte f zu finden, für die dieses Doppelintegral gleich der logistischen Funktion ist, durch die das linear-logistische Modell definiert wird. Ein Versuch, sie zu lösen, soll hier nicht gemacht werden. Es ist aber interessant, den Spezialfall dreier Aufgaben zu betrachten: In zweien von ihnen sei nur jeweils eine Regel anzuwenden, die die Schwierigkeiten κ_1 bzw. κ_2 haben mögen. Die dritte Aufgabe verlange die Anwendung beider Regeln. Es gelte jeweils das linear-logistische Modell. Die Antwortwahrscheinlichkeiten sind dann

$$P(X_j = 1|\theta, \kappa_j) = \frac{\exp(\theta - \kappa_j)}{1 + \exp(\theta - \kappa_j)}, \quad j = 1, 2 \tag{4.6.26}$$

$$P(X_3 = 1|\theta, \kappa_1, \kappa_2) = \frac{\exp(\theta - (\kappa_1 + \kappa_2))}{1 + \exp(\theta - (\kappa_1 + \kappa_2))}, \tag{4.6.27}$$

denn nach Voraussetzung gilt $\kappa_3 = c_1\kappa_1 + c_2\kappa_2$, und da jede Regel in der dritten Aufgabe genau einmal vorkommt, gilt $c_1 = c_2 = 1$. Es werde nun die Annahme gemacht, dass die dritte Aufgabe gelöst werde, indem die beiden Regeln unabhängig voneinander korrekt angwendet werden. Die Wahrscheinlichkeit einer korrekten Antwort ist dann gleich der Wahrscheinlichkeit, dass beide Regeln korrekt angwendet werden, so dass

$$P(X_3 = 1|\theta, \kappa_1, \kappa_2) = P(X_1 = 1|\theta, \kappa_1) P(X_2 = 1|\theta, \kappa_2)$$

gelten müßte. Nun ist aber

$$\begin{aligned} P(X_1 = 1|\theta, \kappa_1) P(X_2 = 1|\theta, \kappa_2) &= \left[\frac{\exp(\theta - \kappa_1)}{1 + \exp(\theta - \kappa_1)}\right] \left[\frac{\exp(\theta - \kappa_2)}{1 + \exp(\theta - \kappa_2)}\right] \\ &= \frac{e^{2\theta - (\kappa_1 + \kappa_2)}}{1 + e^{\theta - \kappa_1} + e^{\theta - \kappa_2} + e^{2\theta - (\kappa_1 + \kappa_2)}} \end{aligned} \tag{4.6.28}$$

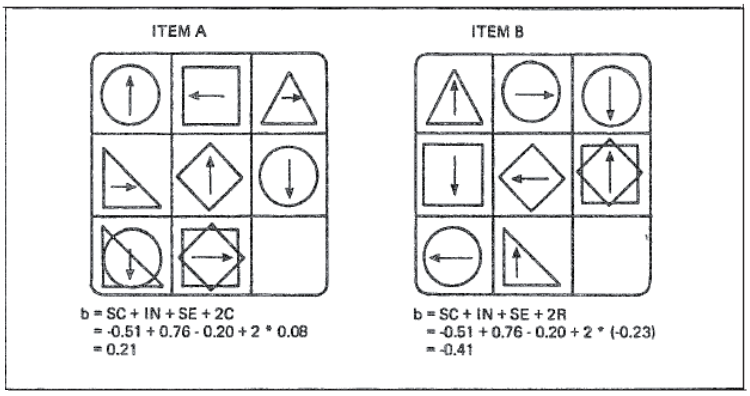
Offenbar ist die rechte Seite ungleich der Wahrscheinlichkeit (4.6.27), schon weil im Nenner von (4.6.28) die Summe $e^{\theta - \kappa_1} + e^{\theta - \kappa_2}$ auftritt, die in (4.6.27) nicht existiert; hinzu kommt, dass in (4.6.28) der Term 2θ auftritt. Der zugehörige Logit-Ausdruck ist nicht linear in den κ_j-Werten, denn für den Wettquotienten findet man

$$\frac{P(X_3 = 1|\theta, \kappa_1, \kappa_2)}{1 - P(X_3 = 1|\theta, \kappa_1, \kappa_2)} = \frac{e^{2\theta - (\kappa_1 + \kappa_2)}}{1 + e^{\theta - \kappa_1} + e^{\theta - \kappa_2}},$$

so dass

$$\text{Logit}(P(X_3 = 1|\theta, \kappa_1, \kappa_2)) = 2\theta - (\kappa_1 + \kappa_2) - \log(1 + e^{\theta - \kappa_1} + e^{\theta - \kappa_2}).$$

Abbildung 22: Aufgaben mit verschiedenen Komponenten, aus Hornke und Habon (1986). Mit Schwierigkeitsberechnungen.



Eine lineare Approximation ergibt sich nur, wenn

$$\log(1 + e^{\theta-\kappa_1} + e^{\theta-\kappa_2}) \approx 0$$

angenommen werden kann, was wiederum

$$e^{\theta-\kappa_1} + e^{\theta+\kappa_2} \approx 0$$

bedeutet (denn $\log 1 = 0$), und dazu müssen die Werte von κ_1 und κ_2 *groß* im Vergleich zu θ sein, – was bedeutet, dass die Wahrscheinlichkeiten für die korrekte Lösung einer Komponente beide nahezu Null sein müssen. Unter der Bedingung, dass beide Komponenten nach dem logistischen Modell und unabhängig voneinander gelöst werden ist die Annahme einer linearen Logit-Funktion also nicht sehr vernünftig. Soll eine lineare Logit-Funktion für beliebige Wahrscheinlichkeiten gelten, schließt man unabhängige Rasch-Modelle für die Einzelkomponenten aus.

Das linear-logistische Modell impliziert also Abhängigkeiten, die bei der ursprünglichen Spezifikation des Modells gar nicht explizit gemacht worden sind. Welcher Art die Abhängigkeiten sein müssen, damit sowohl für Aufgaben mit nur einer Komponente als auch für Aufgaben mit mehr als einer Komponente das Rasch-Modell gilt, ist eine offene Frage. Damit ist aber auch die Frage, was überhaupt über den Lösungsprozess angenommen wird, offen. Verwunderlich ist diese Unklarheit nicht, wenn man an die Herkunft des linear-logistischen Modells aus den Annahmen über die logistische Regression denkt, wo ja nur postuliert wird, dass die Wettchance $p_g/(1-p_g)$ durch ein Polynom ersten Grades approximierbar sein soll, – welche Funktion auch immer dabei approximiert wird. Abbildung 22 zeigt zwei Aufgaben aus Hornke und Habons Sammlung; man kann sich anhand

dieser Aufgaben Gedanken über die Anwendung des linear-logistischen Modells machen und zu eigenen Schlußfolgerungen gelangen.

Zum Schluß sei noch auf das faktorenanalytische Modell (4.6.15), Seite 229, zurückgekommen. Natürlich ist es prinzipiell möglich, dieses Modell mit dem linear-logistischen Modell zu kombinieren. Eventuell würde die Anwendung des fakorenanalytischen Modells Hinweise auf voneinander unabhängige Komponenten des Lösungsprozesses liefern, die dann bei der Definition der Teilschwierigkeiten hilfreich sein können.

4.6.4 Tests der Eindimensionalität und Item Cluster

Auch im mehrdimensionalen Fall lassen sich die Parameter über die Maximum-Likelihood-Methode schätzen. Man hat aber damit noch keinen Test, ob Ein- oder Mehrdimensionalität gilt. Eine wichtige Arbeit zu dem Problem, die Hypothese der Eindimensionaltiät zu testen, wurde von Martin-Löf (1973) vorgelegt. Er testet die Hypothese der Eindimensionalität gegen die Alternativhypothese, dass die Menge der Items in zwei Teilmengen $\mathcal{I}_1$ und $\mathcal{I}_2$ mit je n_1 und n_2 Items zerlegt werden kann. Die Items in diesen beiden Klassen messen verschiedene Merkmale. Für jede Klasse gelte das Rasch-Modell.

In Gleichung (2.6.59), Seite 64, wurde der Begriff des Logits eingeführt. Definiert man

$$P_g(\theta) = P(X_g = 1|\theta),$$

so kann man das Rasch-Modell in der knappen Form

$$\text{Logit}(P_g(\theta)) = \theta - \kappa_g \tag{4.6.29}$$

schreiben. Diese Form entspricht der Nullhypothese, die Martin-Löf betrachtet. Die Alternativhypothese läßt sich dann in der Form

$$\text{Logit}(P_g(\theta)) = \sum_d \delta_{gd}\theta_d - \kappa_g, \quad d = 1, 2;\ g = 1, \ldots, m \tag{4.6.30}$$

anschreiben. $\delta_{gd} = 1$, wenn $I_g \in \mathcal{I}_1$, und $\delta_{gd} = 0$, wenn $I_g \in \mathcal{I}_2$. κ_g ist, wie üblich, die Schwiergkeit des g-ten Items. Es sei nun T der Gesamtscore, und T_d, $d = 1, 2$, seien die Teilscores für die Items aus $\mathcal{I}_1$ bzw. $\mathcal{I}_2$, d.h.

$$T = \sum_g X_g, \quad T_d = \sum_{g \in \mathcal{I}_d} X_g, \quad d = 1, 2 \tag{4.6.31}$$

Weiter sei n_t die Anzahl von Probanden mit dem Gesamtscore t, und $n_{t_1 t_2}$ sei die Gesamtzahl von Probanden mit den Teilscores t_1 und t_2. Die folgende Größe ist eine Teststatistik:

$$U_0(\Theta) = -2 \log \left\{ \frac{\hat{L}_c \prod_{t=0}^{m} (n_t/n)^{n_t}}{\hat{L}_{c_1} \hat{L}_{c_2} \prod_{t_1=0}^{n_1} \prod_{t_2=0}^{n_2} (n_{t_1 t_2}/n)^{n_{t_1 t_2}}} \right\}. \tag{4.6.32}$$

Dabei ist $\hat{L}_c$ die bedingte Likelihood unter der Nullhypothese, $\hat{L}_{c_1}$ und $\hat{L}_{c_2}$ sind die entsprechenden Likelihoods für die Teilklassen $\mathcal{I}_1$ und $\mathcal{I}_2$, und Δ ist eine $(m \times 2)$-Matrix mit den Elementen δ_{gd}. Δ definiert das Modell, nach dem die Items in zwei Klassen aufgeteilt werden. Es ist also nicht so, dass die zwei Klassen explorativ aus den Daten geschätzt werden, sondern es wird eine vom Untersucher postulierte Aufteilung der Items getestet. $U_0(\theta)$ ist asymptotisch χ^2-verteilt mit $n_1 n_2 - 1$ Freiheitsgraden.

Das Problem mit Martin-Löfs Ansatz ist das Rasch-Postulat: beim Rasch-Modell wird ja angenommen, dass alle Items die gleiche Trennschärfe haben. Wenn aber die verwendeten Items dieser Bedingung *nicht* genügen, kann der Wert von $U_0(\Theta)$ dazu führen, dass die Nullhypothese verworfen wird, obwohl sie korrekt ist, d.h. obwohl alle Items der gleichen Klasse angehören. Der Ansatz Martin-Löfs läßt sich nun verallgemeinern, indem man ungleiche Trennschärfen zuläßt. In Abschnitt (4.2), Seite 212, ist das verallgemeinerte (2PL) Modell

$$F_g(\theta) = \frac{\exp((\theta - \kappa_g)/\sigma_g^*)}{1 + \exp((\theta - \kappa_g)/\sigma_g^*)} = \frac{\exp(\alpha_g(\theta - b_g))}{1 + \exp(\alpha_g(\theta - b_g))},$$

d.h.

$$\text{Logit} P_g(\theta) = \alpha_g(\theta - b_g) \tag{4.6.33}$$

eingeführt worden; sind die α_g verschieden für verschiedene Items, so hat man ein Modell, das ungleiche Trennschärfen zuläßt. Bartolucci (2007) hat dieses Modell für verschiedene Annahmen über die α_g numerisch getestet und zeigt, dass im Falle ungleicher α_g-Werte die Nullhypothese eines einzigen Itempools häufiger fälschlich zurückgewiesen wird als im Fall gleicher α_g-Werte.

Bartolucci (2007) diskutiert eine Reihe weiterer Modelle der Form

$$\text{Logit}(P_g(\theta)) = \alpha_g \sum_d \delta_{gd}\theta_d - \kappa_g, \quad d = 1, 2;\ g = 1, \ldots, m \tag{4.6.34}$$

Auf die Details kann hier nicht eingegangen werden. Es sei aber darauf hingewiesen, dass Bartolucci zeigt, dass die Analyse entsprechend formulierter Modelle ein *hierarchisches Clustering* erlaubt, d.h. man kann zu einem Dendogramm der Items wie bei einer Cluster-Analyse kommen.

4.7 Das ordinale Rasch-Modell

Das Merkmal sei kontinuierlich, aber in Kategorien aufgeteilt. Die Probanden liefern Einschätzungen (Ratings) von Zustimmung oder Ablehnung, von sich oder anderen Personen, etc. Wie bei der Thurstone-Skala kann man davon ausgehen, dass die Wahrnehmung des Merkmals nicht konstant im strengen Sinne ist, sondern in einem bestimmten Bereich zufällig fluktuiert. In anderen Worten, man hat es mit den *discriminal dispersions* zu tun, die von Thurstone eingeführt wurden, um aus Präferenz- oder Dominanzurteilen eine Skala herzuleiten, die das Merkmal

abbildet. Die Frage ist, ob man das Rasch-Modell auf mehr als zwei, möglicherweise ordinale Kategorien verallgemeinern kann. Eine solche Verallgemeinerung würde es u. a. zulassen, Teillösungen von Aufgaben zu bewerten. Denn dichotome Items, für die also nur "gelöst" versus "nicht gelöst" notiert werden kann, liefern nur ein vergröbertes Bild, wenn es um komplexere Aufgaben geht.

Eine erste Verallgemeinerung des dichtomen Rasch-Modells geht auf Rasch (1961) zurück. Betrachtet werden Items, für die es r Antwortkategorien gibt. Für das g-te Item soll gelten

$$P(X_g = k|\theta) = \frac{\exp[\delta_k + \phi_k(\theta - \kappa_g)]}{1 + \exp[\delta_k + \phi_k(\theta - \kappa_g)]}, \quad k = 0, \ldots, r \tag{4.7.1}$$

δ_k ist ein Parameter, der für die k-te Kategorie des Items I_g charakteristisch ist, ϕ_k ist ein "scoring coefficient". Rasch (1961) hat bewiesen, dass (4.7.1) das einzige Modell für mehrere Kategorien ist, das eine spezifisch objektive Schätzung von Item- und Personenparametern zuläßt. Anderson (1977) bewies, dass unter der Bedingung, dass die Koeffizienten $\phi_0, \phi_1, \ldots, \phi_r$ gleichabständig sind, der einfache Summenscore

$$s_a = \sum_{g=1}^{n} x_{ag} \tag{4.7.2}$$

eine suffiziente Statistik für θ_a, also für den Personenparameter der Person a, ist. Eine Interpretation der δ_k geht auf Andrich (1978) zurück.

Samejima (1969) hat bereits ein Modell der gestuften Antwort (*graded response model*) vorgeschlagen. Demnach soll gelten

$$P(X_{gak} = 1) = \frac{\exp(\alpha_g(\theta_a - \lambda_{gk}))}{1 + \exp(\alpha_g(\theta_a - \lambda_{gk}))}. \tag{4.7.3}$$

Hierin ist λ_{gk} die Grenze zwischen der $(k-1)$-ten und der k-ten Kategorie beim g-ten Item I_g. α_g definiert die Trennschärfe des g-ten Items. Samejima nahm weiter an, dass die Person a, wenn sie mit dem Item I_g konfrontiert wird, nach Maßgabe einer zufälligen Veränderlichen ε_{ag} reagiert; die Wahrscheinlichkeit, mit der ε_{ag} einen Wert größer als die k-te Kategoriengrenze animmt, hängt vom Wert der Parameter λ_{gk}, θ_a und α_g ab. Ist ε_{ag} logistisch verteilt, so ist die Wahrscheinlichkeit, dass die Person eine Antgwort liefert, die entweder in die k-te oder eine größere Kategorien liefert, ist dann durch (4.7.3) gegeben. Daraus läßt sich dann die Wahrscheinlichkeit, dass die Antwort genau in die k-te Kategorie fällt, ableiten.

Masters (1982) hat ein Modell für partielle Leistungen vorgeschlagen, dass unter dem Namen *partial credit scoring model* bekannt ist. Die r Kategorien im Samejima-Modell entsprechen in diesem Modell den verschiedenen Schritten, die zur Lösung des Items notwendig sind. Masters liefert als Beispiel die Aufgabe

$$\sqrt{\frac{7.5}{.25} - 16} = ?$$

Man sieht sofort, dass offenbar drei Schritte notwendig sind: man muß zunächst den Quotienten 7.5/.25 bewältigen, ohne einen Kommafehler zu begehen. Die Teillösung für diesen Schritt ist 7.5/.25 = 25. Dann muß man die Differenz 25 - 16 bestimmen, also 25 - 16 = 9. Als letzten Schritt muß man die Wurzel aus 9 bestimmen, und damit hat man die Lösung – 3 – gefunden. Offenbar kommt es auf die Reihenfolge dieser Schritte an, – man kann keine Wurzel ziehen, bevor man nicht die Differenz bestimmt hat, und die läßt sich nur bestimmen, wenn man den Quotienten bestimmt hat. Generell können Items mit r Kategorien betrachtet werden, etwa Items, die Meiungen abbilden, etwa von "ausgesprochen nicht einverstanden" ($X_g = 0$), "nicht einverstanden" ($X_g = 1$), "einverstanden" ($X_g = 2$) bis zu "sehr einverstanden" ($X_g = 3$).

Ist die Anzahl der notwendigen Schritte gleich 1, so kann man das Rasch-Modell annehmen. Um die Betrachtungen einfach notieren zu können, führt Masters die folgende Schreibweise ein. Mit κ_{g1} wird die Schwierigkeit des ersten Schritts bezeichnet. Die Wahrscheinlichzkeit, dass die Person a den ersten Schritt bewältigt, werde mit π_{1ag} bezeichnet. $\pi_{0ga} = 1 - \pi_{1ga}$ ist dann die Wahrscheinlichkeit, dass sie ihn nicht bewältigt; natürlich ist $\pi_{1ga} + \pi_{0ga} = 1$, so dass man

$$\pi_{1ga} = \frac{\pi_{1ga}}{\pi_{0ga} + \pi_{1ga}} = \frac{\exp(\theta_a - \kappa_{g1})}{1 + \exp(\theta_a - \kappa_{g1})} \tag{4.7.4}$$

schreiben kann. Dieses Modell gilt dann auch für den Übergang von $X_g = 0$ zu $X_g = 1$. Im dichotomen Rasch-Modell ist die Anzahl der Personen, die ein Item "lösen", eine suffiziente Statistik für den Itemparameter dieses Items. Diese Eigenschaft überträgt sich auf das hier behandelte Modell. Die Anzahl S_k der Personen, die der k-ten Kategorie zustimmen, ist höchstens so groß wie die Anzahl S_{k-1} der Personen, die der $(k-1)$-ten Kategorie zustimmen. Mithin gilt generell

$$S_{g1} \leq S_{g2} \leq \cdots \leq S_{gr}$$

und S_{gk} ist eine suffiziente Statistik für λ_{gk}.

Damit die Person die zweite Kategorie wählt, muß sie mindestens der Ausprägung der ersten Kategorie zustimmen bzw. die erste Stufe lösen. Die Wahrscheinlichkeit für das Lösen der zweiten Stufe oder der Zustimmung zur zweiten Stufe ist dann

$$\frac{\pi_{2ag}}{\pi_{1ag} + \pi_{2ag}} = \frac{\exp[\theta_a - \kappa_{g2}]}{1 + \exp[\theta_a - \kappa_{g2}]}. \tag{4.7.5}$$

Für die dritte Stufe erhält man analog

$$\frac{\pi_{3ag}}{\pi_{2ag} + \pi_{3ag}} = \frac{\exp[\theta_a - \kappa_{g3}]}{1 + \exp[\theta_a - \kappa_{g3}]}, \tag{4.7.6}$$

etc. Der allgemeine Ausdruck ist

$$\pi_{kag} = \frac{\exp\left[\sum_{j=0}^{k}(\theta_a - \kappa_{gj})\right]}{\sum_{k=0}^{r_g}\exp\left[\sum_{j=0}^{k}(\theta_a - \kappa_{gj})\right]} \tag{4.7.7}$$

Masters (1982), p. 159, zeigt, dass die Parameter θ_a und κ_{gk} unabhängig voneinander geschätzt werden können.

Insgesamt muß gelten

$$\pi_{0ag} + \pi_{1ag} + \cdots + \pi_{rag} = 1. \tag{4.7.8}$$

4.8 Die Messung von Veränderungen

Allgemeine Veränderungen: Die Messung von Veränderungen nach einer Therapie, einem Training etc ist stets von besonderem Interesse. Es werde angenommen, dass die Messungen mit den Items $I_1, \ldots, I_n$ dem Rasch-Modell genügen. Die Messungen zum Zeitpunkt t_1 mögen die Parameter $\theta_1, \ldots, \theta_m$ für die Personen und $\kappa_1, \ldots, \kappa_n$ für die Items ergeben. Zum Zeitpunkt t_2 wird der Test noch einmal vorgelegt. Sollten sich überhaupt Veränderungen ergeben, so sollten sie nur bei dem Personenparametern nachweisbar sein, denn die κ_g, $g = 1, \ldots, n$ sollen ja nur die Items charakterisieren.

Im Rahmen der Testtheorie sind Datenstrukturen im Allgemeinen zweidimensional, dh sie können in einer Personen $\times$ Itemmatrix angeschrieben werden. Nimmt man nun noch die Zeit als weitere Dimension hinzu, entstehen dreidimensionale Datenstrukturen. Es gibt einen weiteren Parameter δ_t, der den Effekt des Zeitpunkts der Messung abbildet. Das Modell nimmt demnach die Form

$$P(X_{gat} = 1) = \frac{\exp(\theta_a + \kappa_g + \delta_t)}{1 + \exp(\theta_a + \kappa_g + \delta_t)} \tag{4.8.1}$$

an. Man erhält sofort die Logit-Form

$$\log\left[\frac{P(X_{gat} = 1)}{P(X_{gat} = 0)}\right] = \theta_a + \kappa_g + \delta_t. \tag{4.8.2}$$

Rost (2004) (p. 282) weist explizit auf den offenkundigen Sachverhalt hin, dass es sich – analog zur Varianzanalyse – um ein Logit-Modell ohne Wechselwirkungen handelt. Der Wert von δ_t, der den Effekt auf die latente Variable zum Zeitpunkt t repräsentiert, ist für alle Personen und alle Items gleich. Das Modell (4.8.2) ist mathematisch einfach, psychologisch deswegen aber keineswegs trivial: es ist ja nicht unplausibel, anzunehmen, dass die zeitlichen Effekte für verschiedene Personen verschieden sind. Das Modell ist äußerst restriktiv. Die naheliegende Verallgemeinerung ist demnach, statt δ_t Parameter δ_{at} zu postulieren und damit zuzulassen, dass die zeitlichen Effekte personenspezifisch sind. Dann kann man natürlich die Parameter θ_a und δ_{at} zu einem Parameter $\alpha_{at} = \theta_a + \delta_{at}$ zusammenfassen und erhält

$$P(X_{gat} = 1) = \frac{\exp(\alpha_{at} + \kappa_g)}{1 + \exp(\alpha_{at} + \kappa_g)}. \tag{4.8.3}$$

Die Frage nach der Schätzung der Parameter wird gelöst, indem man das Modell (4.8.3) auf den bekannten Fall eines Zweiparametermodells zurückführt. Dies geschieht, indem man eine Datenstruktur erzeugt, bei der die Datenstrukturen für die einzelnen Zeitpunkte untereinander angeschrieben werden:

$$X = \left[\begin{array}{c|c|ccc} \text{Zeitpunkt} & \text{Person} & \multicolumn{3}{c}{\text{Items}} \\ & & 1 & \cdots & n \\ \hline t=1 & \begin{array}{c}1\\ \vdots \\ m\end{array} & & & \\ \hline t=2 & \begin{array}{c}1\\ \vdots \\ m\end{array} & & & \\ \hline \vdots & & & & \\ \hline t=k & \begin{array}{c}1\\ \vdots \\ m\end{array} & & & \end{array}\right] \tag{4.8.4}$$

Die Personen sind zwar stets dieselben, werden aber bei der Schätzung der Parameter wie verschiedene Personen "behandelt". Es zeigt sich, dass die Schätzungen für die α_{at} i. A. nicht sehr reliabel und mit dem Parameter für $t = 1$ negativ korreliert sind. Von zentraler Bedeutung für die Interpretierbarkeit der Parameter ist die Konstanz der κ_g-Schätzungen. Kann diese nicht angenommen werden, können die Daten nicht im Rahmen des Rasch-Modells gedeutet werden.

Formal ist es möglich, die Personenparameter θ_a als konstant anzunehmen und dafür die κ_g-Werte als zeitabhängig aufzufassen. Rost (2004) betrachtet dazu als Beispiel den Symptomverlauf während einer Therapie, also den Fall, dass sich die Symptome im Verlauf der Therapie verändern. Man kann darüber hinaus noch Modelle betrachten, bei denen Wechselwirkungen zwischen Personen und Items, Personen und Zeitpunkten etc angenommen werden, vergl.Rost (2004), p. 284.

Lernprozesse: Eine spezielle Art der Veränderung sind Lernprozesse während der Testbearbeitung. Sie können sich personenspezifisch, itemspezifisch oder in Form von Wechselwirkungen äußern. Ein Modell, bei dem die Veränderung der Personenparameter angenommen wird, ist

$$P(X_{ai} = 1) = \frac{\exp(\theta_a - \kappa_g + (i-1)\delta_a)}{1 + \exp(\theta_a - \kappa_g + (i-1)\delta_a)}, \quad i = 1, \ldots, n \tag{4.8.5}$$

wobei δ_a jetzt ein Parameter ist, der die Lernfähigkeit repräsentiert. Das Modell stellt aus lerntheoretischer Sicht eine starke Vereinfachung dar, weil der "Lernparameter" für alle Items gleich groß und insofern nicht itemspezifisch und unabhängig von der Zeit ist. Darüber hinaus wird angenommen, dass die Art der Antwort ("Reaktion") keinen Einfluß auf den Lernvorgang hat.

Die Schätzungen der Parameter verweisen auf Interpretationsprobleme, die die Genauigkeit der Schätzungen für die Parameter θ_a und δ_a sich als gering erweist; die Schätzfehler erweisen sich als korrreliert. Klauer und Sydow (1992) haben versucht, die Probleme zu umgehen, indem sie Parameter nicht einzeln schätzten, sondern annahmen, dass die Werte in der Population 2-dimensional normalverteilt sind. Dann müssen nur die Parameter dieser Verteilung geschätzt werden.

Auch itemspezifisches Lernen läßt sich modellieren. Dazu kann man das linear-logistische Testmodell (4.6.23), Seite 231, heranziehen. Man erhält das Modell

$$P(X_{ag} = 1) = \frac{\exp\left(\theta_a - \sum_{j=1}^{h} q_{gj}\eta_j - c\right)}{1 + \exp\left(\theta_a - \sum_{j=1}^{h} q_{gj}\eta_j - c\right)}. \tag{4.8.6}$$

Zur Erinnerung: Nach (4.6.22) entspricht der Summe

$$\sum_{j=1}^{h} q_{gj}\eta_j - c$$

die Schwierigkeit κ_g eines Items; diese Charakteristik des Items wird in Komponenten aufgespalten und der Lernvorgang bezieht sich auf die Komponenten. Die Komponenten q_{gj} werden in einer Matrix Q zusammengefasst:

	Komponenten j							q-Gewichte	
	1	2	3	4	5	6	7	8	9
1	1							0	0
2		1						−1	0
3			1					−1	−1
4				1				−1	0
5					1			−2	−2
6						1		−2	0
7							1	−3	−3

Items (4.8.7)

(nach Rost (2004), p. 289). Hier enthält der Test sieben Items, die durch die Zeilen repräsentiert werden. Das erste Item hat einen Lerneffekt, der sich auf das zweite Item auswirkt; das Gewicht ist -1, weil die Itemschwierigkeit *verringert* und die Lösungswahrscheinlichkeit damit erhöht wird. Der Effekt wirkt sich auf die folgenden Items aus, und beim vierten Item tritt erneut ein Lerneffekt ein, so dass das Gewicht für das fünfte Item nun -2 ist. Ein weiterer Lerneffekt tritt bei Item 6 auf, so dass beim siebten Item ein Gewicht von -3 entsteht.

Die neunte Spalte reflektiert Lerneffekte bei den ungeraden Items I_3, I_5 und I_7.

4.9 Faktorenanalytische Modelle in der Testtheorie

4.9.1 Überblick

Wie am Ende des Abschnitts 3.4.3 angemerkt wurde, sind die Beziehungen zwischen der Reliabilität und der Validität eines Tests komplex: größere Validität läßt sich mit größerer Heterogenität erreichen, während eien größere Reliabilität eher mit einer größeren Homogenität zu erzielen ist. Von Bedeutung sind beide Parameter. Eine Möglichkeit, einen Einblick in das Ausmaß der Heterogenität zu bekommen, ist die Faktorenanalyse der Items, – im Prinzip erlaubt diese Methode auch, homogene, d.h. 1-dimensionale (Sub-)Skalen zu gewinnen.

Die Faktorenanalyse geht üblicherweise von einem linearen Modell aus: Gemessene Variablen $X_1, \ldots, X_p$ sollen durch ein Modell der Art

$$X_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jr}F_r + e_j, \quad r \leq p \tag{4.9.1}$$

erklärt werden, wobei $F_1, \ldots, F_r$ die "Faktoren", d.h. bestimmte latente Variable sind. Die Anzahl r der Faktoren ist, wie diese selbst, unbekannt. Man hofft, dass r möglichst klein ist im Vergleich zu dem von p. Die Koeffizienten ("Ladungen") a_{jk} sind ebenfalls unbekannt. Zumindest bei einem explorativen Ansatz können die a_{jk} sowie die F_k allerdings stets aus den Daten geschätzt werden. Dies folgt aus bestimmten Sätzen der Linearen Algebra, – die unabhängig von der Anwendung des Modells (4.9.1) und damit von psychologischen, medizinischen etc Sachverhalten gelten. In diesem Sinne ist das Modell (4.9.1) auch nicht direkt testbar. Das Modell ist allenfalls indirekt testbar: es erscheint allenfalls dann als unvernünftig, wenn sich keine sinnvolle Deutung der F_k finden läßt. Eine wesentliche Annahme in (4.9.1) ist, dass es keine Interaktionen zwischen den latenten Variablen gibt. Die Betrachtung psychologischer, biologischer und anderer Prozesse zeigt aber, dass es keineswegs zwingend ist, sich auf rein additiv wirkende latente Variablen zu beschränken, weshalb im Folgenden auch einige nichtlineare Ansätze vorgestellt werden sollen.

Der gängige faktorenanalytische Ansatz geht davon aus, dass die $X_1, \ldots, X_p$ Messungen von kontinuierlichen Variablen sind. Viele psychometrische Tests und Fragebögen enthalten aber dichotome Items, d.h. die X_j können etwa nur die Werte 0 oder 1 annehmen. Die direkte Anwendung der Faktorenanalyse, die von Korrelationen zwischen den gemessenen Variablen ausgeht, führt aber zu Schwierigkeiten. Seit Christofferson (1975) existieren allerdings Ansätze, die hier entstehenden Probleme zu umgehen. Diese Ansätze werden ebenfalls vorgestellt. Ein weiterer Ansatz besteht darin, LC-Modelle (LC – Latent Classes) anzuwenden. Sie werden im letzten Abschnitt kurz vorgestellt.

4.9.2 Das Lineare Faktorenmodell

Für jede der Variablen $X_1, \ldots, X_p$ mögen m Messungen vorliegen; sie können jeweils zu einem m-dimensionalen Vektor

$$\vec{X}_j = \begin{pmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{mj} \end{pmatrix} \tag{4.9.2}$$

zusammengefasst werden. Es sei $\bar{x}_j = \sum_i X_{ij}/m$ der Mittelwert der Messungen der j-ten Variablen. Üblicherweise betrachtet man die Abweichungen $x_{ij} = X_{ij} - \bar{x}_j$, also die Vektoren $\mathbf{x}_j$ mit den Komponenten x_{ij}. Schreibt man die $\mathbf{x}_j$ nebeneinander an, entsteht eine Matrix X. Geht man noch zu standardisierten Werten $z_{ij} = x_{ij}/s_j$ über, wobei s_j die Standarabweichung der der Messwerte der j-ten Variablen ist, so erhält man die Matrix Z der standardisierten Messwerte. Offenbar ist $\Sigma = X'X/m$ die Matrix der Varianzen und Kovarianzen der Variablen, und $R = Z'Z/m$ ist die Matrix der Korrelationen zwischen ihnen.

Das Modell der Faktorenanalyse wird häufig in Bezug auf die Matrix X dargestellt. Haben die X_j allerdings verschiedene Maßeinheiten, ist es sinnvoll, zu standardisierten Werten überzugehen, da die z_{ij}-Werte frei von Maßeinheiten sind.

Zur Bestimmung der Faktoren: im Allgemeinen wird man von Null verschiedene Kovarianzen (und damit Korrelationen) zwischen den X_j finden. Dies bedeutet, dass es zwischen irgendzwei Variablen X_j und $X_{j'}$ eine Regressionsbeziehung der Art

$$X_{j'} = \alpha_{jj'} X_{j'} + \beta_{jj'} + e_{jj'} \tag{4.9.3}$$

gibt. Noch allgemeiner kann man etwa für X_1 die Gleichung

$$\vec{X}_1 = \alpha_2 \vec{X}_2 + \cdots + \alpha_p \vec{X}_p + \vec{\varepsilon}_1 \tag{4.9.4}$$

schreiben, d.h. $\vec{X}_j = \vec{X}_1$ läßt sich als Linearkombination der restlichen Vektoren der Messwerte plus einem Fehlervektor anschreiben, wobei die $\alpha_2, \ldots, \alpha_p$ geeignet zu wählende Koeffizienten sind. Eine entsprechende Gleichung läßt sich für jeden anderen Datenvektor anschreiben. Um die Faktoren F_k zu bestimmen, geht man von einem Resultat der Linearen Algebra aus. Hier wird zwischen *linear abhängigen* (l.a.) und *linear unabhängigen* (l.u.) Vektoren unterschieden. Sind $\mathbf{x}_1, \ldots, \mathbf{x}_n$ m-dimensionale Vektoren, so heißen sie linear unabhängig, wenn keiner von ihnen als Linearkombination der übrigen darstellbar ist. Formal bedeutet dies, dass

$$0 = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n \tag{4.9.5}$$

nur dann gilt, wenn $a_1 = a_2 = \cdots = a_n = 0$ gilt. Denn wären etwa zwei der Koeffizienten ungleich Null, etwa a_1 und a_2, so hätte man $0 = a_1\mathbf{x}_1 + a_2\mathbf{x}_2$ und

damit etwa

$$\mathbf{x}_2 = -\frac{a_1}{a_2}\mathbf{x}_1,$$

d.h. $\mathbf{x}_2$ ist eine Linearkombination von $\mathbf{x}_1$ (und umgekehrt, $\mathbf{x}_1$ ist eine von $\mathbf{x}_2$), und damit wären die $\mathbf{x}_j$ nicht mehr linear unabhängig. Gleichungen der Form (4.9.4) mit von Null verschiedenen α_j-Werten bedeuten also Lineare Abhängigkeit der $\vec{X}_j$. Faktoren sollen voneinander unabhängige Dimensionen repräsentieren. Eine Dimension repräsentiert ein Merkmal (das eine Kombination von Merkmalen sein kann), dass nicht durch andere Dimensionen erklärt werden kann. Gemeint ist damit genau das, was durch den Begriff der linearen Unabhängigkeit ausgedrückt wird. Repräsentiert man Faktoren durch Vektoren $\vec{F}_k$ (in einem noch genauer zu spezifierenden Sinn), so sollen diese Vektoren l. u. sein, d.h. keiner der $\vec{F}_k$ als Linearkombination der übrigen $\vec{F}_{k'}$ darstellbar sein. Inhaltlich bedeutet dies, dass sich die Merkmale, die durch die $\vec{F}_k$ repräsentiert werden, nicht durch Kombination der übrgien Merkmale darstellen lassen. Das Ziel ist demnach, l.u. Vektoren $\vec{F}_1, \ldots, \vec{F}_r$ zu finden, d.h. die gemessenen Vektoren als Linearkombinationen einer bestimmten Menge von l.u. Vektoren darzustellen.

Die $\vec{X}_j$ und damit die $\mathbf{x}_j$ und $\mathbf{z}_j$ sind m-dimensionale Vektoren, – einfach, weil sie m Komponenten haben. Es läßt sich zeigen, dass sich stets *alle* m-dimensionalen Vektoren als Linearkombinationen von m l.u. Vektoren $\mathbf{u}_1, \ldots, bu_m$ darstellen lassen. Man nennt $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ dann eine *Basis*. Formal gesehen besteht die Aufgabe der Faktorenanalyse also darin, eine solche Basis zu finden. Die Basis ist nicht eindeutig bestimmt, es gibt beliebig viele Basen, so dass zusätzliche Kriterien formuliert werden müssen, die auf die Wahl einer bestimmten Basis führen.

Zunächst ist festzuhalten, dass im Allgemeinen die Anzahl p der gemessenen Variablen kleiner als die Anzahl m der Messungen pro Variable ist (die Anzahl m der Probanden ist größer als die Anzahl p der Items). Dies impliziert algebraisch, dass alle p Vektoren $\mathbf{x}_j$ in einem p-dimensionalen Teilraum des m-dimensionalen Raumes liegen müssen, – und man hofft, dass sie in einem r-dimensionalen Teilraum ($r < p$) liegen. Demnach müssen r l.u. Vektoren $\mathbf{u}_1, \ldots, \mathbf{u}_r$ ($\vec{F}_j = \mathbf{u}_j$) gefunden werden. Sie bilden eine Teilbasis des m-dimensionalen Raums.

Einen wichtigen Spezialfall erhält man, wenn man nicht nur die lineare Unabhängigkeit, sondern die Orthogonalität der Basisvektoren fordert. Das Skalarprodukt zweier orthogonaler Vektoren $\mathbf{u}$ und $\mathbf{v}$ ist gleich Null, $\mathbf{u}'\mathbf{v} = 0$[60]. Orthogonale Vektoren sind stets l.u., aber l.u. Vektoren sind nicht notwendig auch orthogonal. Fordert man, dass die $\vec{F}_k$ paarweise orthogonal sind, fasst man sie in einer Matrix F zusammen und die Koeffizienten in einer Matrix A, so kann man insbesondere

$$X = FA' \tag{4.9.6}$$

[60] Allgemein gilt für das Skalarprodukt $\mathbf{u}'\mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta$, wobei $\|\mathbf{u}\|$ und $\|\mathbf{v}\|$ die Längen der Vektoren $\mathbf{u}$ bzw. $\mathbf{v}$ sind und θ der Winkel zwischen ihnen ist. $\mathbf{u}'\mathbf{v} = 0$ bedeutet dann $\cos\theta = 0$, und dies ist der Fall, wenn $\theta = \pi/s$ oder 90^o ist, wenn die Vektoren also senkrecht aufeinander stehen.

schreiben, d.h. alle Spaltenvektoren der Matrix X sind Linearkombinationen der orthogonalen Spaltenvektoren in F, wobei F eine (m, r)-Matrix und A eine (p, r)-Matrix ist. Man hat dann $X'X = AF'FA' = A\Lambda A'$, wobei Λ wegen der postulierten Orthogonalität der Spalten von F eine Diagonalmatrix sein muß. Damit hat man aber schon eine Lösung für die unbekannten Koeffizienten (Ladungen) A: A enthält die (orthogonalen) Eigenvektoren von $X'X$, und Λ die zugehörigen Eigenwerte. Normiert man die Spaltenvektoren von A auf die Länge 1, erhält man sofort $XA = F$ und damit eine Lösung für F. Diese Lösungen für A und F sind eindeutig bis auf eine Rotation, die nach Belieben gewählt werden kann. Rechnerisch werden stets p Eigenvektoren und Eigenwerte bestimmt, der Wert von r wird u.a. durch Inspektion bestimmt (wenn man von Kriterien wie dem Scree-Test oder dem Kaiser-Kriterium einmal absieht). Die verbleibenden $p - r$ Eigenvektoren repräsentieren dann den "Fehler".

Will man keine orthognalen Faktoren, so enthält $F'F$ die "Korrelationen" zwischen den "obliquen" Faktoren. Auf die Art der Schätzung solcher Faktoren kann hier nicht eingegangen werden. Da die Anzahl r der als "bedeutsam" eingestuften Faktoren unabhängig von der Wahl orthogonaler oder obliquer Faktoren ist und man stets von einer Basis zu einer anderen übergehen kann, kann man etwa mit einer orthogonalen Basis (den Eigenvektoren von $X'X$ oder $Z'Z$) ausgehen und dann die Eigenvektoren nach irgendwelchen Kriterien rotieren.

Eine Verteilungsannahme mußte bisher nicht gemacht werden. Die Berechnung von Eigenvektoren setzt auch kein spezielles Schätzverfahren oder eine Annahme über die Verteilung der $\mathbf{x}_j$ voraus. Andererseits kann man die Annahme, dass die Daten multivariat normalverteilt sind machen. Dies ermöglicht es, die Faktorenladungen und Faktorenscores als Maximum-Likelihood-Schätzungen zu gewinnen. Im Falle dichotomer Variablen führt der Ansatz, die Faktoren über die Bestimmung der Eigenvektoren der Matrix der Korrelationen zu bestimmen, i. A. zu Verschätzungen einmal des Wertes von r, und zum anderen der Ladungen. Einen Ausweg erhält man, wenn man die Annahme der multivariaten Normalvereteilung macht, wie weiter unten ausgeführt wird.

4.9.3 Dichotome Variable I

Eine unmittelbare Anwendung der Faktorenanalyse auf dichtotome Daten ergibt sich, wenn man von den Korrelationen zwischen den Items ausgeht. Es gibt zwei Möglichkeiten: man geht vom ϕ-Koeffizienten oder vom tetrachorischen Koeffizienten aus. Die damit verbundenen Probleme werden im Folgenden vorgestellt.

Der ϕ-Koeffizient: Es seien X und Y dichotome Variable, d.h. es gelte $X = \{0, 1\}$ und $Y = \{0, 1\}$. Es werden N Messungen gemacht, d.h es wird bei N Probanden der Wert sowohl von X als auch von Y bestimmt. Gesucht ist die Korrelation zwischen X und Y. Das Ergebnis der Messungen kann in einer Vier-

feldertafel zusammengefasst werden:

	Y		
X	1	0	Σ
1	a	b	a + b
0	c	d	c + d
Σ	a + c	b + d	N

(4.9.7)

Für die folgenden Betrachtungen ist es nützlich, relative Häufigkeiten zu betrachten. Man hat dann

	Y		
X	1	0	Σ
1	a/N	b/N	$(a+b)/N = p_x$
0	c/N	d/N	$(c+d)/N = q_x$
Σ	$(a+c)/N = p_y$	$(b+d)/N = q_y$	1

(4.9.8)

wobei natürlich $q_x = 1 - p_x$, $q_y = 1 - p_y$. Wendet man die gewöhnliche Formel für den Produkt-Moment-Korrelationskoeffizienten auf die Daten in (4.9.7) an, so ergibt sich nach einigen Vereinfachungen der Ausdruck

$$r_{xy} = \phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{4.9.9}$$

Für die relativen Häufigkeiten findet man

$$\phi = \frac{(ad - bc)/N^2}{\sqrt{p_x p_y q_x q_y}} \tag{4.9.10}$$

Natürlich sind die Ausdrücke (4.9.9) und (4.9.10) für ϕ numerisch identisch.

Der Produkt-Moment-Korrelationskoeffizient r_{xy} für Messwerte liegt bekanntlich zwischen den Grenzen -1 und +1, d.h. $-1 \leq r_{xy} \leq 1$. Für ϕ ist dies allerdings nur unter bestimmten Randbedingungen der Fall: allgemein gilt $-1 \leq \phi_{\min} \leq \phi \leq \phi_{\max} \leq 1$:

Satz 4.9.1 *Für gegebene Randverteilungen p_x und p_y gilt*

$$\phi_{\min} = \max\left(-\sqrt{\frac{p_x p_y}{q_x q_y}}, -\sqrt{\frac{q_x q_y}{p_x p_y}}\right) \leq \phi \leq \min\left(\sqrt{\frac{p_x q_y}{q_x p_y}}, \sqrt{\frac{p_y q_x}{p_x q_y}}\right) = \phi_{\max} \tag{4.9.11}$$

und

$$\phi_{\max} = 1 \iff p_x = p_y \tag{4.9.12}$$

$$\phi_{\min} = -1 \iff p_x = 1 - p_y \tag{4.9.13}$$

$$-1 \leq \phi \leq 1 \iff p_x = p_y = .5. \tag{4.9.14}$$

Beweis: Die Fälle $b = c = 0$ ($\phi_{\max} = 1$) und $a = d = 0$ ($\phi_{\min} = -1$) werden am Ende des Beweises betrachtet.

Es werde zuerst $\phi_{\max}$ bestimmt. Aus (4.9.10) folgt, dass *für gegebene Randverteilungen* p_x und p_y ϕ dann maximal wird, wenn $bc = 0$ ist. Es werden zunächst die Fälle $b = 0$, $c > 0$ und $b > 0$, $c = 0$ betrachtet. Es wird zuerst $\phi_{\max}$ bestimmt. **Der Fall** $b = 0$, $c > 0$**:** Aus der Tabelle 4.9.8 liest man ab, dass nun $a = p_x$ und $d = q_y$ gilt, so dass

$$\phi_{b=0,c>0} = \frac{p_x q_y}{\sqrt{p_x p_y (1-p_x)(1-p_y)}} = \sqrt{\frac{p_x q_y}{q_x p_y}}. \tag{4.9.15}$$

Der Fall $b > 0$, $c = 0$**:** aus der Tabelle 4.9.8) ergeben sich $a = p_y$ und $d = q_y$, so dass

$$\phi_{b>0,c=0} = \frac{p_y q_x}{\sqrt{p_x p_y (1-p_x)(1-p_y)}} = \sqrt{\frac{p_y q_x}{p_x q_y}}. \tag{4.9.16}$$

$\phi_{b=0,c>0}$ und $\phi_{b>0,c=0}$ sind mögliche Ausdrücke für ϕ. Inspektion dieser Ausdrücke zeigt, dass $\phi_{b=0,c>0} = 1/\phi_{b>0,c=0}$. ϕ kann keinen Wert größer als 1 annehmen. Gilt also $\phi_{b>0,c=0} \leq 1$, so folgt $\phi_{b=0,c>0} \geq 1$, und umgekehrt. Also folgt

$$\phi_{\max} = \min(\phi_{b=0,c>0}, \phi_{b>0,c=0}). \tag{4.9.17}$$

Nun werde $\phi_{\min}$ bestimmt. ϕ wird minimal, wenn $ad = 0$ gilt. Für $a = d = 0$ folgt sofort $\phi = -1$, wie man sich wieder anhnd von (4.9.9) klarmacht.

Der Fall $a = 0$**,** $d > 0$**:** Aus der Tabelle 4.9.8 ergibt sich $b = p_x$ und $c = p_y$ und

$$\phi_{a=0,\,d>0} = -\frac{p_x p_y}{\sqrt{p_x p_y (1-p_x)(1-p_y)}} = -\sqrt{\frac{p_x p_y}{q_x q_y}}. \tag{4.9.18}$$

Der Fall $a > 0$**,** $d = 0$**:**, Tabelle 4.9.8 liefert $b = q_y$, $c = q_x$, so dass

$$\phi_{a>0,\,d=0} = -\frac{q_x q_y}{\sqrt{p_x p_y (1-p_x)(1-p_y)}} = -\sqrt{\frac{q_x q_y}{p_x p_y}}. \tag{4.9.19}$$

Wieder gilt $\phi_{a=0,\,d>0} = 1/\phi_{a>0,\,d=0}$, und $\phi_{a>0,\,d=0} \leq 1$ impliziert $\phi_{a=0,\,d>0} \geq 1$, und umgekehrt. Daraus folgt

$$\phi_{\min} = \max(\phi_{a=0,\,d>0},\ \phi_{a>0,\,d=0}). \tag{4.9.20}$$

Damit sind die untere und die obere Grenze für ϕ bestimmt.

Nun sei $\phi_{\max} = 1$. Es gilt

$$\phi_{\max} = \min\left(\sqrt{\frac{p_x q_y}{q_x p_y}}, \sqrt{\frac{p_y q_x}{p_x q_y}}\right),$$

Dann gilt entweder

$$\phi_{\max} = \sqrt{\frac{p_x q_y}{q_x p_y}} = 1,$$

woraus $p_x q_y = q_x p_y$, d.h. $p_x(1-p_y) = (1-p_x)p_y$ folgt, d.h. $p_x - p_x p_y = p_y - p_x p_y$, also $p_x = p_y$. Sei umgekehrt $p_x = p_y$. Dann folgt aus $p_x q_y = q_x p_y$

$$p_x(1-p_x) = (1-p_x)p_x \Rightarrow \phi_{\max} = 1.$$

Der Fall $\phi_{\max} = \sqrt{p_y q_x / p_x q_y}$ impliziert die gleichen Folgerungen.

Für $\phi_{\min} = -\sqrt{p_x p_y / q_x q_y}$ (s. (4.9.18)) hat man $\phi_{\min} = -1$ wenn $p_x p_y = q_x q_y$, d.h.

$$p_x p_y = (1-p_x)(1-p_y) = 1 = p_x - p_y + p_x p_y,$$

woraus $1 - p_x - p_y = 0$, d.h. $1 = p_x + p_y$ folgt (s. (4.9.13)). Der Fall $q_x q_y / p_x p_y = 1$ führt auf $(1-p_x)(1-p_y) = 1 - p_x - p_y + p_x p_y = p_x p_y$, also wieder auf $1 = p_x + p_y$.

Nun gelte $1 = p_x + p_y$; dann gilt

$$\frac{p_x p_y}{q_x q_y} = \frac{p_x(1-p_x)}{(1-p_x)(1-p_y)} = \frac{p_x(1-p_x)}{(1-p_x)p_x} = 1,$$

so dass $\phi_{min} = -1$. Die Schlußfolgerung für $\phi_{min} = -\sqrt{q_x q_y / p_x p_y}$ ist analog.

$\phi_{\max} = 1$ und $\phi_{\min} = -1$ führen also zusammen auf $p_x = p_y$ und $p_x + p_y = 1$, woraus $p_x = p_y = .5$, also (4.9.14) folgt. □

Sind also die Bedingungen (4.9.14), also $p_x = p_y = .5$, nicht erfüllt, ist der Bereich möglicher ϕ-Werte auf ein Teilintervall von $[-1, 1]$ eingeschränkt, d.h. es ist unabhängig vom tatsächlichen Zusammenhang zwischen den Variablen $|\phi\| < 1$. Die Faktorenanalyse wird dann im Allgemeinen mehr laténte Dimensionen anzeigen, als es tatsächlich gibt. Da p_x und p_y die Schwierigkeiten des X- bzw-Y-Items sind, reflektieren die überzähligen Faktoren nur die unterschiedlichen Schwierigkeiten der Items; man spricht von *Schwierigkeitsfaktoren*, die eben nicht qualitativ verschiedene latente Dimensionen darstellen. Die Beziehung (4.9.14) besagt, dass alle dichotomen Items eines Tests die Schwierigkeit .5 haben sollten, um Schgwierigkeitsfaktoren zu vermeiden. Diese Forderung ist in aller Strenge schwer zu erfüllen. McDonald & Ahlawat (1974) haben allerdings nachgewiesen, dass das Problem der Schwierigkeitsfaktoren nicht überbewertet werden sollte: sind die Unterschiede zwischen den Schwierigkeiten (und damit den Randverteilungen) nicht allzu drastisch und sind die Regressionen zwischen den Items linear, so sind keine Schwierigkeitsfaktoren zu erwarten!

Der tetrachorische Koeffizient: Gegeben seien die zufälligen Veränderlichen X_1 und X_2, die bivariat normalverteilt seien, d.h. die gemeinsame Dichtefunktion sei durch

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \right.\right. \\ &\quad \left.\left. + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right]. \end{aligned} \tag{4.9.21}$$

Dabei sind μ_1 und μ_2 die Erwartungswerte von X_1 bzw. X_2, und σ_1^2 und σ_2^2 sind die Varianzen von X_1 und X_2. ρ ist die Korrelation zwischen den beiden Variablen.

Nun werde angenommen, dass X_1 und X_2 nicht direkt beobachtet werden können, sondern nur Dichotomisierungen: es seien

$$Y_1 = \begin{cases} 0, & X_1 \leq \gamma_1 \\ 1, & X_1 > \gamma_1 \end{cases}, \quad Y_2 = \begin{cases} 0, & X_2 \leq \gamma_2 \\ 1, & X_2 > \gamma_2 \end{cases} \tag{4.9.22}$$

Weiter sei

$$\pi_1 = P(Y_1 = 1) = P(X_1 > \gamma_1) = \int_{\gamma_1}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_d x_2 \tag{4.9.23}$$

$$\pi_2 = P(Y_2 = 1) = P(X_2 > \gamma_2) = \int_{-\infty}^{\infty} \int_{\gamma_2}^{\infty} f(x_1, x_2) dx_1 dx_2. \tag{4.9.24}$$

Weiter kann man die Wahrscheinlichkeit bestimmen, mit der sowohl $X_1 > \gamma_1$ als auch $X_2 > \gamma_s$ gilt; diese ist durch

$$\pi_{12} = P(X_1 > \gamma_1 \cap X_2 > \gamma_2) = \int_{\gamma_1}^{\infty} \int_{\gamma_2}^{\infty} f(x_1, x_2) dx_1 dx_2 \tag{4.9.25}$$

gegeben.

$$\bar{\pi}_{12} = P(X_1 \leq \gamma_1 \cap X_2 \leq \gamma_2) = \int_{\infty}^{\gamma_1} \int_{\infty}^{\gamma_2} f(x_1, x_2) dx_1 dx_2 \tag{4.9.26}$$

Die Vierfeldertafel illustriert diese Wahrscheinlichkeiten:

	$X_2 > \gamma_2$	$X_2 \leq \gamma_2$	Σ
$X_1 > \gamma_1$	π_{12}	$\pi_1 - \pi_{12}$	π_1
$X_1 \leq \gamma_1$	$1 - \pi_1 - \bar{\pi}_{12}$	$\bar{\pi}_{12}$	$1 - \pi_1$
Σ	π_2	$1 - \pi_2$	1

(4.9.27)

Die Wahrscheinlichkeiten π_1, π_2, π_{12} und $\bar{\pi}_{12}$ können als Anteile aus den Daten geschätzt werden. Daraus ergibt sich die Schätzung r_{12} für ρ aus den obigen Gleichungen. Natürlich können die Gleichungen für die π-Werte nicht nach ρ und den anderen Parametern aufgelöst werden; die Schätzungen müssen numerisch bestimmt werden. Die Schätzung r_{12} ist der *tetrachorische Korrelationskoeffizient*.

Die Problematik des tetrachorischen Korrelationskoeffizienten besteht in der Annahme der bivariaten Normalverteilung. Gilt diese Annahme nicht, erhält man verzerrte Korrelationskoeffizienten. Gilt diese Annahme, so kann es natürlich sein, dass verschiedene Gruppen von Personen sich durch die Erwartungswerte μ_1 und μ_2 unterscheiden. Der Korrelationskoeffizient ρ ist aber nicht von diesen beiden Parametern abhängig, so dass die Schätzungen r_{xy} für ρ *invariant* gegenüber Veränderungen der Erwartungswerte sein sollten. Carroll (1961) hat gezeigt, dass die Normalverteilung die einzige Verteilung ist, die diese Invarianz zuläßt.

4.9.4 Dichotome Variable II

Im Falle dichotomer Items nehmen die Komponenten von $\mathbf{x}_j$, $j = 1, \ldots, p$ nur die Werte 1 oder 0 an; beantwortet eine Person das j-te Item "positiv" (löst sie die Aufgabe oder gibt sie das Vorhandensein eines Merkmals an, oder findet man das Merkmal bei der Person), so wird $x_{ij} = 1$ gesetzt, sonst $x_{ij} = 0$. Der folgende Ansatz wurde zuerst von Christofferson (1975) vorgeschlagen.

Man kann die Messungen x_{ij} mit den Werten einer entsprechenden latenten Variablen ξ in Verbindung bringen: Ist, für die i-te Person, $\xi_i \geq \tau_j$, so löst die Person die Aufgabe bzw. beantwortet sie positiv, und für $\xi_i < \tau_j$ löst sie die j-te Aufgabe nicht bzw. beantwortet sie negativ. τ_j ist ein Schwellenparameter, der für das j-te Item charakteristisch ist.

In Gleichung (4.9.6) wurden die (Mess-)Werte in X über die Beziehung $X = FA'$ zu latenten Variablen in F mit den Ladungen A in Beziehung gesetzt. Insbesondere hat man für die j-te Variable X_j die Gleichung

$$\mathbf{x}_j = F\mathbf{a}_j,$$

wobei $\mathbf{a}_j$ die j-te Spalte von A', d.h. die j-te Zeile von A ist. Ist

$$\vec{\xi_j} = \begin{pmatrix} \xi_{1j} \\ \xi_{2j} \\ \vdots \\ \xi_{mj} \end{pmatrix}$$

der j-te Vektor für die latenten Variablen mit ξ_{ij} der Wert der i-ten Person für diese Variable, so kann man analog zu (4.9.6)

$$\vec{\xi_j} = F\mathbf{a}_j + \varepsilon_j, \quad \xi = FA' + \varepsilon \tag{4.9.28}$$

schreiben, wobei jetzt ξ die aus den Spaltenvektoren $\vec{\xi_j}$ zusammengesetzte Matrix ist, und die Zeilen von A durch die $\mathbf{a}_j$ gegeben sind. Fasst man $(\xi_1, \ldots, \xi_p)$ als zufälligen Vektor auf (die ξ_{ij} sind dann Realisationen der Komponente ξ_j) und nimmt man an, dass er multivariat normalverteilt sei, so dass

$$f(\xi) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{p/2}} \exp(-\frac{1}{2}\varepsilon'\Sigma^{-1}\varepsilon) \tag{4.9.29}$$

so ist Σ durch

$$\Sigma = \xi'\xi = A\Lambda A' + \Psi \tag{4.9.30}$$

gegeben. $F'F = F'F$ ist die Matrix der Korrelationen zwischen den Faktoren, falls der oblique Fall zugelassen wird. Im Falle orthogonaler Faktoren ist Λ eine Diagonalmatrix. Da die beobachteten Messungen nur die Werte 0 oder 1 annehmen kann, folgt, dass die Diagonalwerte von Σ nicht identifizierbar sind. Man setzt dann die Diagonalelemente gleich 1 und erhält dann

$$\Psi = I - \mathrm{diag}(A\Lambda A'). \tag{4.9.31}$$

Um das Modell an die Daten anzupassen, müssen (i), der Vektor $\tau = (\tau_1, \ldots, \tau_p)'$ der Schwellen sowie (ii) die Matrizen A und F aus den Daten geschätzt werden. Um die Maximum-Likelihood-Methode anwenden zu können, muß zunächst ein Ausdruck für die Wahrscheinlichkeit der Daten gefunden werden; für den Zufallsvektor $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)'$ findet man

$$g(\mathbf{x}_1, \ldots, \mathbf{x}_p) = \int_{\tau_1}^{\infty} \int_{-\infty}^{\tau_2} \cdots \int_{\tau_p}^{-\infty} f(\mathbf{x}) d\mathbf{x} \tag{4.9.32}$$

Ist hierin τ_j die untere Grenze eines Integrals, so bedeutet dies, dass die latente Variable einen Wert größer als τ_j hat, demenstprechend $x_{ij} = 1$ ist. Ist τ_j die obere Grenze des Integrals. Dies ist der Ausdruck für eine Person; die entsprechenden Ausdrücke für alle Personen müssen dann ebenfalls miteinander multipliziert werden. Es ist klar, dass die Minimalisierung der Gesamt-Likelihood eine formidable Aufgabe ist, zumal die Mehrfachintegrale in (4.9.32) implizieren einen hohen Rechenaufwand. Christoffersons Ansatz wurde von Muthén (1978) aufgenommen, der eine verbesserte Schätzung vorschlug; auf die Details kann hier nicht eingegangen werden.

4.9.5 Faktorenanalyse und Item-Response-Theorie

In diesem Abschnitt wird ein von Takane und De Leeuw (1987) gelieferter Beweis der Äquivalenz des faktorenanalytischen Ansatzes von Christofferson (1975) und dem IRT-Modell präsentiert.

Es sei $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_p)'$ ein Zufallsvektor für Antwortmuster für p dichotome Items; für eine gegebene Person ist $\tilde{x}$ eine Folge von Nullen und Einsen. $\tilde{x}_j = 1$, wenn das j-te Item gelöst oder "positiv" beantwortet wurde, und $\tilde{x}_j = 0$ sonst. Weiter sei $\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_m)'$ ein m-dimensionaler Vektor ($m < p$ die Anzahl der PRobanden in einer Stichprobe), dessen Komponenten die Fähigkeiten der Probanden abbilden. $\tilde{u}$ ist nicht direkt beobachtbar und insofern latent und sei in der Population multivariat normalverteilt ($\tilde{u} \sim N(0, I)$). Es sei $P(\tilde{x} = \mathbf{x}|\mathbf{u})$ die bedingte Wahrscheinlichkeit, das spezielle Antwortmuster $\mathbf{x}$ uz beobachten, wenn der Vektor der Fähigkeiten $\tilde{u}$ ist. Die unbedingte Wahrscheinlichkeit für $\mathbf{x}$ ist dann

$$P(\tilde{x} = \mathbf{x}) = \int_U P(\tilde{x} = \mathbf{x}|\tilde{u}) g(\mathbf{u}) d\mathbf{u}, \tag{4.9.33}$$

$g(\mathbf{u})$ die Dichte für $\mathbf{u}$. Akzeptiert man die Annahme der lokalen Unabhängigkeit, so hat man

$$P(\tilde{x} = \mathbf{x}|\mathbf{u}) = \prod_{i=1}^{p} (p_i(\mathbf{u}))^{x_i} (1 - p_i(\nu))^{1-x_i}, \quad x_i \in (0, 1) \tag{4.9.34}$$

mit

$$p_i(\mathbf{u}) = \int_{-\infty}^{\mathbf{a}'\mathbf{u}+b} \phi(z) dz = \Phi(\mathbf{a}'\mathbf{u} + b). \tag{4.9.35}$$

ϕ ist die Dichte der Standardnormalverteilung, Φ ist die dazu korrespondierende Verteilungsfunktion.

In Christoffersons (1975) Modell ist

$$P(\tilde{x} = \mathbf{x}) = \int_R h(\mathbf{y}) d\mathbf{y} \tag{4.9.36}$$

mit

$$\tilde{y} = C\tilde{u} + \tilde{e}. \tag{4.9.37}$$

C ist die Matrix der Faktorladungen und $\tilde{u}$ ist der Vektor der Faktorscores (= der Vektor der Fähigkeiten). Es wird angenommen, dass $\tilde{u} \sim N(0, Q^2)$, wobei Q^2 eine Diagonalmatrix ist (was der Annahme der lokalen Unabhängigkeit entspricht). $\tilde{u}$ und $\tilde{e}$ sollen unabhängig voneinander sein. Dann hat man

$$\tilde{y} \sim N(0, CC' + Q^2), \quad \tilde{y}|\mathbf{u} \sim N(C\mathbf{u}, Q^2), \tag{4.9.38}$$

($\tilde{y}|\mathbf{u}$ ist die bedingte Verteilung von $\tilde{y}$, gegeben $\mathbf{u}$). Die $\tilde{y}$ werden dichotomisiert:

$$\tilde{x}_i = \begin{cases} 1, & \tilde{y}_i \geq \tau_i \\ 0, & \tilde{y}_i < \tau_i \end{cases} \tag{4.9.39}$$

wobei die τ_i wieder Schwellenparameter sind.

Die Ausdrücke (4.9.33) und (4.9.36) sind äquivalent. Dazu zeigt man zuerst, dass (4.9.36) den Ausdruck (4.9.33) impliziert:

$$\begin{aligned} P(\tilde{x} = \mathbf{x}) &= \int_R h(\mathbf{y}) d\mathbf{y} \\ &= \int_R \left(\int_U f(\mathbf{y}|\mathbf{u}) g(\mathbf{u}) \right) d\mathbf{y} \\ &= \int_R g(\mathbf{u}) \left(\int_R f(\tilde{y}|\mathbf{u}) d\mathbf{y} \right) d\mathbf{u}. \end{aligned} \tag{4.9.40}$$

Die Beziehung (4.9.34) impliziert dann

$$\begin{aligned} \int_R f(\mathbf{y}|\mathbf{u}) d\mathbf{y} &= \prod_{i=1}^{p} \int_{R_i} f_i(y_i|\mathbf{u}) dy_i \\ &= \prod_i \left(\int_{\tau_i}^{\infty} f_i(y_i|\mathbf{u}) dy_i \right)^{x_i} \left(1 - \int_{\tau_i}^{\infty} f_i(y_i|\mathbf{u}) dy_i \right)^{1-x_i} \end{aligned} \tag{4.9.41}$$

mit

$$\int_{\tau_i}^{\infty} f_i(y_i|\mathbf{u}) dy_i = \Phi\left(\frac{\mathbf{c}_i'\mathbf{u} - \tau_i}{q_i} \right), \quad i = 1, \ldots, p \tag{4.9.42}$$

Dieser Ausdruck entspricht (4.9.35) mit

$$\mathbf{a}_i = \mathbf{c}_i / q_i, \quad b_i = -\tau_i / q_i. \tag{4.9.43}$$

Die Implikation (4.9.36) $\Rightarrow$ (4.9.33) zeigt man, indem man die eben gegebene Herleitung umgekehrt laufen läßt.

Die Varianz der $\tilde{y}_i$ kann wegen der dichotomen Daten nicht geschätzt werden, so dass den q_i ein beliebiger Wert zugeordnet werden kann. Die Modelle (4.9.36) und (4.9.33) haben dann die gleiche Anzahl freier und damit zu schätzender Parameter. Damit ist die Äquivalenz nachgewiesen. Das faktorenanalytische Modell und das IRT-Modell sind demnach nur zwei verschiedene FOrmulierungen ein und desselben Modells. Ein Unterschied in den Schätzungen entsteht allenfalls dadurch, dass beim IRT-Modell die Maximum-Likelihood-Methode angewendet wird, während beim faktorenanalytischen Modell die Verallgemeinerte KQ-Methode benützt wird.

4.9.6 Nichtlineare Modelle

Zunächst sei daran erinnert, dass der Ansatz (4.9.1) im Kern ein multipler Regressionsansatz ist, allerdings ein spezieller. Nimmt man an, dass x_j durch $r \leq p$ latente Variablen bestimmt wird, so kann allgemein $X_j = f_j(F_1, \ldots, F_r)$ gelten, wobei f_j irgendeine Funktion ist. f_j ist nicht bekannt. Aber unter sehr allgemeinen Bedingungen kann eine unbekannte Funktion durch ein geeignet gewähltes Polynom approximiert werden[61]. Die einfachste Approximation ist dann (4.9.1). Darüber hinaus kann man Terme hinzufügen, die durch Potenzen der F_k sowie durch Produkte $F_k F_{k'}$ etc definiert sind. Die Kunst ist dann, die zugehörigen Koeffizienten zu schätzen.

Wie Busemeyer und Jones (1983) ausführen, sind viele psychologische Gesetzmäßigkeiten durch Produkte von Variablen definiert. So ist etwa die Arbeitsmotivation dem Modell von Vroom (1964) durch das Produkt von Erwartung und Valenz (expectency $\times$ valence) definiert, und die Performanz einer Person durch das Produkt von Fähigkeit und Motivation (ability $\times$ motivation). Sofern das Modell korrekt ist, sollten Tests etwa der Motivation oder der Performanz die korrespondierenden Produkte von entsprechenden latenten Variablen erfassen. Die Autoren zeigen dann allerdings, dass es außerordentlich schwierig ist, Modelle dieser Art über hierarchische Regressionsmodelle zu testen: sind die Prädiktorvariablen nicht Messfehlerfrei, so werden interaktive Trends unterschätzt, und die Reliabilität von Produkttermen ist eine Funktion des Produktes der Einzelreliabilitäten, – da diese Zahlen kleiner als 1 sind, ist die Gesmatrealiabilität kleiner als die Einzelreliabilitäten. Hinzu kommen Probleme, die sich aus den Skalenniveaus ergeben: Produktterme können u. U. durch geeignete Transformationen in additive Terme transformiert werden.

Kenny und Judd (1984) haben aber gezeigt, dass es u. U. möglich ist, tatsächlich ist, Modelle mit nichtlinear wirkenden latenten Variablen zu schätzen.

[61] Weierstraßscher Approximationssatz

Sie illustrieren ihren Ansatz zweier Gleichungen:

$$y = a_1 x + a_2 x^2 + u \tag{4.9.44}$$
$$y = b_1 x + b_2 z + b_3 xz + v \tag{4.9.45}$$

Hierin x und z Variable, die auf y wirken, und a_1, a_2 sowie b_1, b_2 und b_3 sind Regressionskoeffizienten, u und v sind Residuen ("Fehler"). xz repräsentiert eine interaktive Wirkung von x und z auf y. Die Frage ist nun, wie diese Koeffizienten geschätzt werden können, wenn x und z nicht direkt gemessen werden können und als latente Variable in die Größe y eingehen. Es wird angenommen, dass alle Variablen Abweichungen vom jeweiligen Mittelwert repräsentieren.

Zuerst wird die Schätzung der Parameter für den Fall der Gleichung (**??**) illustriert. Dazu definieren Kenny & Judd zwei "Indikatoren" x_1 und x_2 für die latente Variable x:

$$x_1 = x + u_1 \tag{4.9.46}$$
$$x_2 = cx + u_2 \tag{4.9.47}$$

wobei c wieder ein Parameter ist. Dann wird angenommen, dass die Residuen u_1, u_2 und u und die latente Variable x alle unabhängig voneinander sind. Um den Effekt von x^2 schätzen zu können, müssen entsprechende Indikatoren bestimmt werden. Dazu bieten sich die Größen x_1^2, x_2^2 und $x_1 x_2$ an. Aus (4.9.46) und (4.9.47) erhält man

$$x_1^2 = x^2 + u_1^2 + 2xu_1 \tag{4.9.48}$$
$$x_2^2 = c^2 x^2 + u_2^2 + 2cxu_2 \tag{4.9.49}$$
$$x_1 x_2 = cx^2 + cxu_1 + xu_2 + u_1 u_2 \tag{4.9.50}$$

Man kann nun die zu schätzenden Parameter (= Ladungen) in einer Tabelle zusammenfassen Offenbar muß nur ein einzelner Paramter, c, geschätzt werden.

Tabelle 14: Ladungen für das nichtlineare Faktorenmodell

Variable	x	x^2	u_1	u_2	u_1^2	u_2^2	xu_1	xu_2	u_1u_2
x_1	1	0	1	0	0	0	0	0	0
x_2	c	0	0	1	0	0	0	0	0
x_1^2	0	1	0	0	1	0	2	0	0
x_2^2	0	c^2	0	0	0	1	0	$2c$	0
x_1x_2	0	c	0	0	0	0	c	1	1

Die Gesamtheit der latenten Variablen ist $x, x^2, u_1, u_2, u_1^2, u_2^2, xu_1, xu_2, u_1u_2$. Die Struktur der Kovarianzmatrix für diese Variablen hängt allerdings von Annahmen

über deren Verteilung ab. Unter der Annahme, dass x, u_1 und u_2 normalverteilt sind, ergeben sich die folgenden Beziehungen

$$\begin{array}{cc} \sigma^2_{x^2} = 2\sigma^4_x, & \sigma^2_{u_1^2} = 2\sigma^4_{u_1} \\ \sigma^2_{u_2^2} = 2\sigma^4_{u_2}, & \sigma^2_{xu_1} = \sigma^2_x\sigma^2_{u_1} \\ \sigma^2_{xu_2} = \sigma^2_x\sigma_{u_2}, & \sigma^2_{u_1u_2} = \sigma_{u_1}\sigma_{u_2} \end{array}$$

Die Kovarianzen erweisen alle als gleich Null, und die Varianzen sind Funktioen von σ^2_x, $\sigma^2_{u_1}$ und $\sigma^2_{u_2}$. Aus der Annahme der Normalverteilung für x, u_1 und u_2 folgt allerdings *nicht*, dass auch x^2 normalverteilt ist, so dass ein Maximum-Likelihood-Ansatz auf der Basis einer multivariaten Normalverteilung nicht zur Schätzung der Parameter herangezogen werden kann. Einem Vorschlag von McDonald (1978) folgend verwenden die Autoren eine Verallgemeinerte Kleinste-Quadrate-Schätzung mit der Inversen der Stichprobenkovarianzmatrix als Gewichtsmatrix (die Verallgemeinerte KQ-Schätzung wird im Anhang vorgestellt).

In analoger Weise können Interaktionen zwischen latenten Variablen diskutiert werden. Dazu werden wieder Indikatoren für die Interaktionen definiert und dann wird die dazu korrespondierende Ladungsmatrix hergeleitet. Die Matrix für die Kovarianzen zwischen den latenten Variablen wird unter der Annahme der multivariaten Normalverteilung hergeleitet. Für den Ansatz (4.9.45) ergeben sich

$$\begin{aligned} x_1 &= x + u_1 && (4.9.51)\\ x_2 &= d_1x + u_2 && (4.9.52)\\ x_3 &= z + u_3 && (4.9.53)\\ x_4 &= d_2z + u_4 && (4.9.54) \end{aligned}$$

Daraus ergeben sich die Indikatoren für das Produkt xz:

$$\begin{aligned} x_1z_1 &= xz + xu_3 + zu_1 + u_1u_3 && (4.9.55)\\ x_1z_2 &= d_2xz + xu_4 + d_2zu_1 + u_1u_4 && (4.9.56)\\ x_2z_1 &= d_1xz + d_1xu_3 + zu_2 + u_2u_3 && (4.9.57)\\ x_2z_2 &= d_1d_2xz + d_1xu_4 + d_2zu_2 + u_2u_4 && (4.9.58) \end{aligned}$$

Damit hat man insgesamt 15 latente Variable: x, z, xz, xu_3, xu_4, zu_1, zu_2, u_1, u_2, u_3, u_4, u_1u_3, u_1u_4, u_2u_3, u_2u_4. Wird angenommen, dass x, z, u_1, u_2, u_3, u_4 und v Abweichungen vom jeweiligen Mittelwert repräsentieren, multivariat normalverteilt sind und paarweise unkorreliert sind mit der Ausnahme von x und z, so sind die Diagonalelemente der Kovarianzmatrix der latenten Variablen durch

$$\begin{array}{ccc} \sigma^2_{xz} = \sigma^2_x\sigma^2_z + \sigma^2_{x,z} & \sigma^2_{u_1,u_2} = \sigma^2_{u_1}\sigma^2_{u_3}, & \sigma^2_{u_1,u_4} = \sigma^2_{u_1}\sigma^2_{u_4} \\ \sigma^2_{u_2u_3} = \sigma^2_{u_2}\sigma^2_{u_3}, & \sigma^2_{u_2u_4} = \sigma_{u_2}\sigma^2_{u_4}, & \sigma^2_{xu_3} = \sigma_{x^2}\sigma^2_{u_3} \\ \sigma^2_{zu_1} = \sigma^2_z\sigma^2_{u_1}, & \sigma_{zu_2} = \sigma^2_z\sigma_{u_2}, & \end{array} \qquad (4.9.59)$$

Hier ist $\sigma^2_{x,z}$ die Kovarianz für x und z. Unter diesen Randbedingungen können die Parameter für das nichtlineare Modell über die Verallgemeinerte KQ-Methode gefunden werden. Der Fit des Modells ist bemerkenswert gut.

Tabelle 15: Ladungen für das nichtlineare Faktorenmodell

Variable	x	z	xz	u_1	u_2	u_3	u_4	u_1u_3	u_1u_4	u_2u_3	u_2u_4	xu_3	xu_4	zu_1	zu_2
x_1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
x_2	d_1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
z_1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
z_2	0	d_2	0	0	0	0	1	0	0	0	0	0	0	0	0
x_1z_1	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0
x_1z_2	0	0	d_2	0	0	0	0	0	1	0	0	0	1	d_2	0
x_2z_1	0	0	d_1	0	0	0	0	0	0	1	0	d_1	0	0	1
x_2z_2	0	0	d_1d_2	0	0	0	0	0	0	0	1	0	d_1	0	d_2

Muthén (2002) liefert eine Übersicht über nichtlineare Faktormodelle; in Muthén und Muthén (1998 – 2001) findet man einen *user's guide* für das Programm MPlus, mit dem verschiedene Modelle auf Daten angewendet werden können.

4.9.7 Latent-Class-Modelle

Eine häufige Fragestellung in der Diagnostik ist die Zuordnung von Probanden oder Patienten zu bestimmten Klassen. Die Klassen können Krankheiten oder Untergruppen von Krankheiten sein, oder Berufsgruppen, für die ein Proband besonders geeignet ist, oder Kulturen bzw. Zeitabschnitten, denen ein Archäologe seine Funde zuordnen möchte. Gegeben sind allgemein Symptome, und das Auftreten bestimmter Gruppen von Symptomen legt eine bestimmte Zuordnung nahe. Das diagnostische Ziel ist demnach eine Klassifikation, – wobei aber zumindest am Anfang noch nicht klar ist, ob solche Klassen überhaupt existieren. Bei einer Diskriminanzanalyse sind im Allgemeinen bereits bestimmte Klassen vorgegeben und man sucht nach einer Gewichtung der Symptome, die eine optimale Zuordnung zu den betrachteten Klassen erlauben. Dabei werden aber kontinuierliche Ausprägungen der Symptome vorausgesetzt. Will man anhand von dichotomen oder ordinalen Indikatoren klassifizieren und soll u. U. erst herausgefunden werden, ob Klassen existieren, muß nach anderen Verfahren gesucht werden.

Das LC-Modell für dichotome Variable wurde von Lazarsfeld und Henry (1968) eingeführt und dann von Goodman (1974) für den Fall nominaler Variablen verallgemeinert; weitere Literaturangaben findet man in Magidson und Vermunt (2003), die insbesondere den explikatorischen Fall diskutieren. Die folgenden Betrachtungen sind an dieser Arbeit orientiert.

In einer explorativen LC-Analyse wird zunächst ein Einklassenmodell, dann ein Zweiklassenmodell etc an die Daten angepasst. Solche Modelle heißen auch *LC Cluster Modelle*. Van der Ark und Van der Heijden (1998) und Van der Heijden,

Gilula und Van der Ark (1999) haben gezeigt, dass eine LC Analyse dazu benutzt werden kann, die Anzahl der latenten Variablen zu bestimmen, die einer Menge von nominalen Variablen unterliegen. Darüber hinaus gibt es Beziehungen zur Korrespondenzanalyse und zur Faktorenanalyse.

Man kann zwischen LC Cluster Modellen und LC Faktormodellen unterscheiden. Beide Modelle können im Rahmen der log-linearen Modelle beschrieben werden.

LC Cluster Modelle Zur Illustration werden vier nominale Variane A, B, C und D betrachtet, dazu eine latente Variable X mit T Kategorien. Das log-lineare Modell ist dann

$$\log F_{ijklt} = \lambda + \lambda_t^x + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{it}^{Ax} + \lambda_{jt}^{Bx} + \lambda_{kt}^{Ck} + \lambda_{lt}^{Dx} \quad (4.9.60)$$

i, j, k, l bezeichnen Stufen von A, B, C und D, und t die Stufen der latenten Variablen, $t = 1, \ldots, T$. (4.9.60) hat die Form eines log-linearen Modells für eine 5-dimensionale Häufigkeitstabelle mit den Häufigkeiten F_{ijklt}. Es gibt Terme, die "Haupteffekte" (i) mit der latenten Variablen x repräsentieren und mit den vier "Symptomen" A bis D. Darüber hinaus gibt es "Interaktionseffekte" zwischen x und den Symptomen. Die Terme $\lambda_t^x, \lambda_i^A, \ldots, \lambda_l^D$ werden mit in die Betrachtung einbezogen, um keine Vorannahmen über die Randverteilungen machen zu müssen. Es wird angenommen, dass die Reaktionen auf $A, \ldots, D$ gegeben die t-te Stufe von x entspricht der Annahme der lokalen Unabhängigkeit, die es ermöglicht, Interaktionsterme zwischen den $A, \ldots, D$ zu vernachlässigen. Für den Fall nur einer Klasse ($T = 1$) reduziert sich das Modell auf

$$\log F_{ijkt} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D. \quad (4.9.61)$$

Es gibt hier keine Wecheselwirkungen, – dieser Fall repräsentiert die Nullhypothese H_0. H_0 ist das Modell, gegen das andere Modelle, in denen Interaktionen angenommen werden, getestet werden. Es hat

$$N_{param}(H_0) = (I-1) + (J-1) + (K-1) + (L-1) \quad (4.9.62)$$

freie, also zu schätzende Parameter. Für das Modell (4.9.60) hat man

$$N_{param}(T) = (T-1) + N_{param}(H_0) \times (1 + (T-1)) = (T-1) + N_{param}(H_0) \times T, \quad (4.9.63)$$

und die Anzahl der Freiheitsgrade für den Test zur Modellanpassung ist

$$DF_T = IJKL - N_{param}(T) - 1 = IJKL - (1 + N_{param}(H_0)) \times T \quad (4.9.64)$$

Man beginnt mit dem Grundmodell ($T = 1$), und jedesmal, wenn die Anzahl der latenten Variablen um 1 erhöht wird, erhöht sich die Anzahl verschiedener Parameter um $1 + N_{param}(H_0)$ und die Anzahl der Freiheitsgrade reduziert sich um diese Zahl.

Das Latent Class Faktorenmodell Es sei X eine latente Variable mit vier Kategorien, $X = \{1, 2, 3, 4\}$. Sie könnte erklärt werden durch Bezug auf auf zwei dichotome latente Variablen $V = \{1, 2\}$, $W = \{1, 2\}$, indem man

	$W=1$	$W=2$
$V=1$	$X=1$	$X=2$
$V=2$	$X=3$	$X=4$

(4.9.65)

Das LC Cluster-Modell der Gleichung (4.9.60) mit $T = 4$ Klassen kann nun als unbeschränktes LC Faktorenmodell mit zwei dichotomen Variablen V und W angeschrieben werden:

$$\begin{aligned} \log F_{ijklrs} &= \lambda + \lambda_r^V + \lambda_s^W + \lambda_{rs}^{VW} + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ir}^{AV} + \lambda_{jr}^{BV} \\ &= +\lambda_{kr}^{CV} + \lambda_{ir}^{DV} + \lambda_{is}^{AW} + \lambda_{irs}^{AVW} + \lambda_{jrs}^{BVW} \\ &\quad +\lambda_{krs}^{CVW} + \lambda_{irs}^{DVW} \end{aligned} \tag{4.9.66}$$

Der Term λ_x^t in (4.9.60) taucht hier nicht mehr auf, – statt dessen kann

$$\lambda_{2(r-1)+s} = \lambda_r^V + \lambda_s^W + \lambda_{rs}^{VW}$$

geschrieben werden. Die 2-Variablenterme, die X enthalten, können in der Form

$$\lambda_{i,2(r-1)+s} = \lambda_{ir}^{AV} + \lambda_{is}^{AW} + \lambda_{irs}^{AVW}$$

und

$$\lambda_{j,2(r-1)+s} = \lambda_{js}^{BW} + \lambda_{js}^{BW} + \lambda_{jrs}^{BVW}$$

ausgedrückt werden, und so fort. Man hat hier eine Reparametrisierung des ursprünglichen Modells, bei dem aber die Anzahl der freien Parameter nicht reduziert wird.

Man kann nun ein *Basis-R-Faktorenmodell* definieren: darin werden R paarweise unabhängige, dichotome latente Variablen angenommen, bei dem die Faktorladungen entsprechenden Parameter die Assoziation der latenten Variablen mit den gemessenen "Indikator"variablen[62]. Beim Basis-R-Faktorenmodell werden bestimmte Restriktionen für die latenten Variablen spezifiziert. Die erste Restriktion besteht darin, dass alle Terme die Interaktionen zwischen mehr als 2 Variablen repräsentieren gleich Null gesetzt werden, so dass

$$\lambda_{irs}^{AVW} = \lambda_{irs}^{BVW} = \lambda_{irs}^{CVW} = \lambda_{irs}^{DVW} = 0$$

resultiert. Die 2-Variablenterme nehmen dann die Form

$$\lambda_{i,2(r-1)+s}^{AX} = \lambda_{ir}^{AV} + \lambda_{is}^{AW}, \quad \lambda_{j,2(r-1)+s}^{BX} = \lambda_{jr}^{BV} + \lambda_{js}^{BW}, \ \ etc$$

[62]Der Begriff der Indikatorvariablen ist üblicherweise etwas anders definiert, ewa $\chi = 1$, wenn ein Ereignis eingetreten ist, $\chi = 0$ sonst; hier bedeutet der Ausdruck soviel wie: ein Merkmal ist vorhanden – oder nicht.

an. Die Nullsetzung der 3-Variableninteraktionen entspricht insofern der Standardfaktorenanalyse, als dann jede latente Variable einen Einfluß auf jede beobachtete Variable hat und es keine Wechselwirkungen höherer Ordnung gibt (in Gleichung (4.9.1) treten keine Terme auf, die durch Produkte der F_k definiert sind).

Das LC-Faktorenmodell ist ein Spezialfall eines LC-Cluster-Modells; die Tabelle 16 zeigt die möglichen Äquivalenzen.

Tabelle 16: Äquivalenbeziehungen zwischen LC-Cluster-Modellen und LC-Faktormodellen

LC Cluster-Modelle			LC-Faktormodelle		
Anzahl der Lat. Klassen	Anzahl der Parameter	df	Anzahl der Faktoren	Anzahl der Parameter	df
1	5	26	0	5	26
2	11	20	1	11	20
3	17	14	2	17	14
4	23	8	3	23	8
5	29	2	4	29	2

Eine Anwendung dieser Verfahren findet man in Rist, Glöckner-Rist und Demmel (2009).

4.9.8 Multiple Korrespondenzanalyse

Nicht nur dichotome Items stellen eine Schwierigkeit für faktorenanalytische Ansätze zur Abschätzung der Homogenität von Items dar. Hat man Daten von Fragebögen, bei denen die Befragungspersonen Ratings abgeben, etwa von 1 bis 7, oder -3 über 0 bis +3, so sind die Häufigkeitsverteilungen der Antworten für die einzelnen Skalen keineswegs immer in guter Näherung normalverteilt. Bei einer Skala können die Antworten nahezu gleichverteilt über die einzelnen Kategorien sein, bei einer anderen sind sie eher exponentialverteilt, bei einer dritten sind sie nahezu normalverteilt, bei wieder einer anderen kann man U-förmige Verteilungen bekommen. In welchem Ausmaß derartige Unterschiede in den Häufigkeitsverteilungen Verzerrungen der Repräsentation der Struktur der latenten Variablen implizieren, ist für einen gegebenen Datensatz oft nur sehr schwer abzusehen.

Zunächst wird allgemein die Korrespondenzanalyse, dann die Multiple Korrespondenzanalyse als Spezialfall der Itemanalyse vorgestellt.

Grundbegriffe der Korrespondenzanalyse: Hat man Kontingenztabelle mit irgendwelchen Zeilen- und Spaltenkategorien, für die keinerlei Metrik erklärt ist (weiblich – männlich, Berufe oder Studienfächer, etc), so kann man die Frage

nach der Existenz einer möglichen Beziehung zwischen den Zeilen- und Spaltenkategorien über einen χ^2-Test beantworten. Ein solcher Test sagt nichts über die Art der Abhängigkeit, sondern nur über ihre Existenz etwas aus.

Der Charme der Faktorenanalyse besteht u. a. darin, einerseits für die "Tests" – also etwa die Items – und andererseits für die Personen eine Repräsentation durch Punkte in einem Koordinatensystem zu finden, dessen Achsen inhaltlich als latente Dimensionen oder latente Variable gedeutet werden können. Die Frage ist, ob sich eine analoge Repräsentation für die Zeilen- und Spaltenvektoren einer Kontingenztabelle finden lässt.

Die Punktekonfiguration etwa der Items, wie sie durch die FA geliefert wird, erlaubt eine Interpretation in Bezug auf die Ähnlichkeit zweier Items. Je kleiner die (euklidische) Distanz zwischen zwei Punkten ist, desto ähnlicher sind die zu den Punkten korrespondierenden Items, d.h. sie erfassen die latenten Dimensionen in ähnlicher Weise. Hat man also zwei Items I_g und $I_{g'}$ und ist die latente Struktur 2-dimensional, so ist jedes Item durch zwei Koordinaten $(alpha_{g1}, \alpha_{g2})$ und $(\alpha_{g'1}, \alpha_{g's})$ charakterisiert, wobei die α-Werte die Ladungen der Items auf den beiden Achsen sind. Die euklidische Distanz zwischen den beiden Items ist

$$d(I_g, I_{g'}) = [(\alpha_{g1} - \alpha_{g'1})^2 + (\alpha_{g2} - \alpha_{g'2})^2]^{1/2}; \tag{4.9.67}$$

dies ist natürlich einfach der Satz des Pythagoras, der sich auf mehr als zwei Dimensionen verallgemeinern läßt.

Der FA-Ansatz besteht darin, die – im Allgemeinen standardisierten – Messwerte x_{ij} oder z_{ij} durch eine lineare Regressionsgleichung zu definieren, wobei die latenten Dimensionen als Prädiktoren eingesetzt werden. Bei einer Häufigkeitstabelle müssten die Häufigkeiten n_{ij} in dieser Weise erklärt werden, wollte man den FA-Ansatz direkt auf die Tabelle übertragen. Bekanntlich führt diese dirkete Übertragung in Schwierigkeiten, weshalb generalisierte lineare Modelle wie loglineare Analyse oder die Probit- bzw Logit-Analyse auf Häufigkeiten anwendet. Um latente Dimensionen für Häufigkeitstabellen zu finden, die die Abhängigkeiten zwiscen Zeilen- und Spaltenkategorien "erklären", muß also ein anderer Weg beschritten werden.

Betrachtet man die Ähnlichkeit zweier Probanden i ind i', so kann man das Profil ihrer Messwerte x_{ig} und $x_{i'g}$ über die "Tests" (Items) I_g, $g = 1, \ldots, m$ betrachten. Gilt $x_{ig} \approx x_{i'g}$ für alle I_g, so werden die Faktorwerte der beiden Probanden sehr ähnlich sein, d.h. sie werden durch nahe beieinander liegende Punkte repräsentiert. Eine analoge Aussage gilt für die Items.

Bei einer Häufigkeitstabelle kann man in analoger Weise vorgehen. Sind die Häufigkeitsverteilungen zweier Zeilenkategorien C_i und $C_{i'}$ ähnlich, so sollten die Punkte, die diese beien Kategorien repräsentieren, nahe beieinander liegen. Je unähnlicher die Verteilungen sind, desto weiter voneinander entfernt sollten die korrespondierenden Punkte sein. Man benötigt also ein Ähnlichkeits- oder Unähnlichkeitsmaß für Häufigkeitsverteilungen.

Ein solches Maß läßt sich über die χ^2-Formel finden, über die sich das Ausmaß von Abhängigkeiten in der Tabelle ausdrücken läßt. Bekantlich ist

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(n_{ij} - n_{i+}n_{+j}/N)^2}{n_{i+}n_{+j}/N}, \quad df = (m-1)(n-1) \tag{4.9.68}$$

wobei n_{ij} die Häufigkeit der Fälle in der i-ten Zeile und j-ten Spalte ist, und

$$n_{i+} = \sum_j n_{ij}, \quad n_{+j} = \sum_j n_{ij}, \quad N = \sum_i \sum_j n_{ij}$$

sind die Zeilen- bzw. Spaltensummen. N ist die Gesamtzahl der Fälle. $n_{i+}n_{+j}/N$ sind die unter H_0 erwarteten Häufigkeiten, also die Häufigkeiten, die man erwartet, wenn es keinerlei Zusammenhang zwischen Zeilen- und Spaltenkategorien gibt.

Um den Begriff der Ähnlichkeit zweier Zeilenverteilungen zu spezifizieren, werde einmal der idealisierte Fall betrachtet, dass zwei Verteilungen identisch seien, dass also $n_{ij} = n_{i'j}$ für alle j gilt. Besteht die Tabelle nur aus diesen beiden Zeilen, so gilt für diese Tabelle $\chi^2 = 0$, wie man sich leicht überlegt. Für den allgemeinen Fall kann man sagen, dass zwei Zeilen einer Tabelle um so weniger zum Gesamt-χ^2 beitragen, je weniger sich die Häufigkeitsverteilungen der beiden Zeilen unterscheiden. Man kann demnach die Distanz zwischen zwei Zeilenkategorien über die Differenz der Beiträge der beiden Zeilen zum Gesamt-χ^2 definieren. Ein erster Ansatz wäre also, die Distanz $d(i, i')$ zwischen der i-ten und der i'-ten Zeile einfach als Differenz $\chi_i^2 - \chi_{i'}^2$ zu definieren. Das erweist sich aber wegen einer Reihe von Gründen als ungünstig; die Definition einer Distanz zwischen zwei Zeilen (und, analog dazu, zwischen zwei Spalten) ist statt dessen wie folgt:

Definition 4.9.1 *Es werden die i-te und die i'-te Zeilenkategorie betrachtet. Die durch*

$$\delta_{ii'}^2 == \sum_{j=1}^{J} \frac{1}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{i'j}}{n_{i'\cdot}} \right)^2 \tag{4.9.69}$$

definierte Größe $\delta_{ii'}^2$ heißt χ^2-Distanz *zwischen den Zeilenkategorien R_i und $R_{i'}$. Durch*

$$\delta_{jj'}^2 = \sum_{i=1}^{I} \frac{1}{n_{i\cdot}} \left(\frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij'}}{n_{\cdot j'}} \right)^2 \tag{4.9.70}$$

heißt χ^2-Distanz *für die Spaltenkategorien S_j und $S_{j'}$.*

Man kann für die Zeilen eine mittlere Häufigkeitsverteilung bestimmen und dann die χ^2-Distanz $\delta_{i\cdot}^2$ der i-ten Zeile zu dieser mittleren Zeile anschreiben; dann läßt sich zeigen, dass das Gesamt-χ^2 der Tabelle gerade durch

$$\chi^2 = \sum_{i=1}^{m} n_{i+}\delta_{i\cdot}^2 \tag{4.9.71}$$

gegeben ist. Ein analoger Ausdruck läßt sich für die Spalten herleiten.

Der Aufbau der Formeln (4.9.69) und (4.9.70) ist im Hinblick auf Eigenschaften der Repräsentation sowohl der Zeilen- als auch der Spaltenkategorien konstruiert worden, worauf hier nicht weiter eingegangen zu werden braucht. Man sieht jedenfalls, dass im Extremfall $n_{ij} = n_{i'j}$ für alle j die χ^2-Distanz $\delta^2_{ii'}$ gleich Null wird. Eine analoge Aussage gilt für irgendzwei Spalten. Je weniger sich also die Häufigkeiten in zwei Zeilen bzw. Spalten unterscheiden, desto geringer ist die entsprechende Distanz. Die Punktekonfiguration für die Zeilen und für die Spalten soll nun so bestimmt werden, dass die euklidischen Distanzen zwischen den jeweiligen Punkten, wie sie etwa in (4.9.67) erklärt wurde, den χ^2-Distanzen zwischen den entsprechenden Zeilen- bzw. Spaltenkategorien entspricht.

Um den Begriff einer Skala zu erläutern, muß noch über die Dreiecksungleichung gesprochen werden. Hat man irgend drei Punkte A, B und C, und sind $d(A, B)$, $d(A, C)$ und $d(B, C)$ Distanzen zwischen diesen Punkten, so soll stets die Ungleichung

$$d(A, C) \leq d(A, B) + d(B, C) \tag{4.9.72}$$

gelten. Drei Punkte können stets auf einer Ebene positioniert werden, sie definieren dann die Eckpunkte eines Dreiecks, es sei denn, dass Gleichheitszeichen gilt, so dass

$$d(A, C) = d(A, B) + d(B, C) \tag{4.9.73}$$

gilt. Dann liegen die drei Punkte auf einer Geraden.

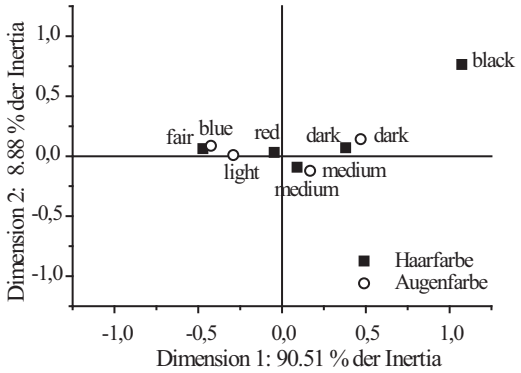
Repräsentieren nun die Punkte die Zeilen einer Tabelle und gilt für die χ^2-Distanzen zwischen ihnen die Gleichung (4.9.73), so können die Zeilen alle auf einer Geraden präsentiert werden. Es kann gezeigt werden, dass dann auch die Spaltenkategorien aller auf einer Geraden positioniert werden können. Wie bei der PCA kann man nun argumentieren, dass diese Gerade die gleiche latente Variable abbildet, und dass diese latente Variable die im Gesamt-χ^2 ausgedrückten Abhängigkeiten "erzeugt". Der Sachverhalt sei an einem Beispiel erläutert:

Beispiel 4.9.1 Das Interesse an genetischen Zusammenhängen hat früh dazu geführt, dass statistische Verfahren zur Analyse der Daten entwickelt wurden. Tocher (1908) legte eine große Studie zur Genetik der Pigmentierung schottischer Schulkinder durch, die Fischer (1940) zu ersten Überlegungen führte, Zeilen- und Spaltenkategorien auf einer gemeinsamen Skala abzubilden. Die Personen werden hier in Klassen eingeteilt: die Zeilenkategorien definieren Klassen, die durch die Augenfarbe definiert sind, die Spaltenkategorien sind Klassen, die bestimmmten Haarfarben entsprechen. Die Daten sind in der Tabelle 17 zusammengefasst worden. Der Zusammenhang zwischen Zeilen- und Spaltenkategorien ist hochsignifikant: $\chi^2 = 2466.14$ bei $df = 12$ Freiheitsgraden hat unter H_0 (kein Zusammenhang) eine Wahrscheinlichkeit von $p = .000$. Der Zusammenhang wird in Abbildung 23 dargestellt. Die erste Dimension erklärt gut 90 % des Gesamt-χ^2.

Tabelle 17: Haar- und Augenfarben schottischer Schulkinder (Tocher (1908), Maung (1941))

		Haarfarbe				
		fair	red	medium	dark	black
Augen-farbe	Blue	1368	170	1041	398	1
	Light	2577	474	2703	932	11
	Medium	1390	420	3826	1842	33
	Dark	454	255	1848	1506	112

Abbildung 23: Biplot: Die Beziehung zwischen Augen- und Haarfarbe – Fischer (1940), nach Daten von Tocher (1908)



Nur das sehr dunkle Haar ("black") scheint eine systematische Abweichung von der ersten Dimension zu erzeugen.

Die Häufigkeitsverteilungen in der Tabelle 17 weisen bereits auf eine enge Kopplung zwischen bestimmten Augen- und Haarfarben hin. Aber die Inspektion der Daten läßt noch nicht vermuten, dass der Zusammenhang im wesentlichen durch *eine* latente Dimension gestiftet wird. Man betrachte zunächst die Präsentation der Zeilenkategorien, d.h. der Augenfarben. Die Abweichungen von der x-Achse können als Resultat der unvermeidlichen Stichprobenfehler gesehen werden. Für die Distanzen zwischen den Punkten, die Augenfarben abbilden, gilt in guter Näherung die Gleichung (4.9.73). Die Augenfarben werden von "blue" über "light" und "medium" zu "dark" abgebildet: je größer der Unterschied in der Farbe, desto größer die entsprechende χ^2-Distanz: benachbarte Farben haben ähnlichere Häufigkeitsverteilungen *in Bezug auf die Kategorisierung nach der Haarfarbe* als weniger benachbarte. Die Interpretation für die Spaltenkategorien, also die Haarfarben, ist analog: je weiter zwei Haarfarben auseinander liegen, desto verschiedener sind die Häufigkeitsverteilungen bezüglich der Augenfarben. Die Skalenwerte

drücken Unterschiedlichkeit hinsichtlich der jeweiligen Farben aus.

Die Tatsache, dass sowohl für die Zeilen- als auch für die Spaltenkategorien die gleichen Skalen verwendet werden und damit sowohl die Zeilen- als auch die Spaltenkategorien in einem gemeinsamen Koordinatensystem repräsentiert werden, drückt sich im Namen 'Biplot' für diese Art der Repräsentation aus. Die Eindimensionalität (bis auf die Sonderrolle der Haarfarbe "black") deutet auf eine enge genetische Kopplung von Augen- und Haarfarbe. □

Die Frage ist nun, wie man zu den Skalenwerten kommt. So, wie man bei der Principal-Component-Analyse (PCA) eine Zerlegung der Varianz der Daten in additive Komponenten vornimmt, liegt es hier nahe, eine Zerlegung des χ^2 in additive Komponenten vorzunehmen. Der Hintergrund für diesen Ansatz ist die Tatsache, dass die Summe von unabhängigen χ^2-Verteilungen wieder eine χ^2-Verteilung ist, d.h. $\chi^2_{r+s} = \chi^2_r + \chi^2_s$. Um eine derartige Zerlegung zu erreichen, betrachtet man die Matrix der *Residuen*:

$$x_{ij} = \frac{n_{ij} - n_{i+}n_{+j}/N}{\sqrt{n_{i+}n_{+j}/N}}. \tag{4.9.74}$$

Wie der Vergleich mit (4.9.68) zeigt, gilt $\chi^2 = \sum_{i,j} x_{ij}^2$. Auf die Matrix $X = (x_{ij})$ kann nun die Singularwertzerlegung

$$X = Q\Lambda^{1/2}P' \tag{4.9.75}$$

angewendet werden. Λ ist die Matrix der Eigenwerte von X, und es gibt so viele Dimensionen, wie es hinreichend von Null verschiedene Eigenwerte gibt. Hier gelten die gleichen Regeln wie bei der üblichen PCA einer Matrix von Messwerten. Im Allgemeinen ergeben sich für 2-dimensionale Tabellen höchstens 2 relevante Eigenwerte und damit 2 Dimensionen. Die Zeilen von Q enthalten die Koordinaten der Zeilenkategorien auf den Dimensionen, die von P die der Spaltenkategorien. Diese Koordinaten werden allerdings noch so skaliert, dass die euklidischen Distanzen der Kategorien zum Ursprung des Koordinatensystems gerade den oben angegebenen χ^2-Distanzen entsprechen. Die Herleitung der Beziehungen zwischen den Koordinaten in Q und P und den χ^2-Distanzen ist ein wenig länglich und kann im Skriptum *Einführung in die Korrespondenzanalyse* gefunden werden.

Üblicherweise repräsentiert man sowohl die Zeilen- als auch die Spaltenkategorien durch Punkte im gleichen Koordinatensystem; man spricht dann von einem *Biplot*. FÜr die Interpretation wichtig ist die Tatsache, dass man wohl die Distanzen entweder zwischen den Zeilenkategorien oder zwischen den Spaltenkategorien interpretieren kann, nicht aber die Distanzen zwischen einer Zeilen- und einer Spaltenkategorie. Diese Distanzen sind nicht erkärt; interpretieren läßt sich nur der Winkel zwischen den Vektoren, die vom Ursprung des Koordinatensystems zu den verschiedenen Kategorien zeigen; Details hierzu werden im Skriptum zur Korrespondenzanalyse angegeben.

Multiple Korrespondenzanalyse: Das oben angegebene Beispiel des Zusammenhanges von Augen- und Haarfarbe entspricht in gewisser Weise bereits der Situation, die man beim Testen hat, wobei allerdings die Probanden und Probandinnen bereits in bestimmten Kategorien zusammengefasst wurden, etwa nach der Augenfarbe. Die "Testleistung" ist die Haarfarbe, die ebenfalls kategorisiert wird. Die umgkehrte Betrachtung, nach der die Probanden nach Maßgabe der Haarfarbe klassifiziert werden und die Augenfarbe die "Testleistung" ist, ist logisch gleichberechtigt.

Oft hat man aber für die Probanden noch gar keine Kategorien und man ist nur daran interessiert, welche Antwortkategorien bei einer Reihe von Testfragen angekreuzt werden und ob die Fragen in einem 1- oder mehrdimensionalen Raum repräsentiert werden können. Hat ein Item I_g etwa k_g Antwortkategorien, so kann man die Antwort einer Person durch eine 1 signalisieren, die für die Kategorie angeschrieben wird, die die Person angekreuzt hat; alle übrigen Antwortkategorien bei diesem Item werden durch eine 0 markiert. Die Datenmatrix (Indikatormatrix) Z hat dann die Gestalt Man bildet nun die Matrix $Z'Z$; dies ist

Tabelle 18: n-variate Indikatormatrix

	I_1	I_2		I_n
1	1 0 0 0	0 1 0 0 0	$\cdots$	0 1 0
2	0 1 0 0	0 0 1 0 0	$\cdots$	1 0 0
$\vdots$				
i	0 0 1 0	0 1 0 0 0	$\cdots$	0 1 0
$\vdots$				
m	1 0 0 0	0 0 0 1 0	$\cdots$	0 0 1

die *Burt-Matrix*, nach Burt (1950). Bevor auf die weitere Analyse dieser Matrix eingegangen wird, ist es sinnvoll, sich den Aufbau dieser Matrix klar zu machen.

Es werde zunächst angenommen, dass $n = 1$, und dass dieses Item gerade r Antwortalternativen hat. Jede Person kreuzt eine und nur eine der Alternativen an. Die Matrix $Z'Z$ dann eine $r \times r$-Matrix. Das Element k_{uv}, $1 \leq u, v \leq r$ ist das Skalarprodukt der u-tenZeile von Z' und der v-ten Spalte von Z, d.h.

$$k_{uv} = z_{u1}z_{v1} + \cdots + z_{um}z_{vm}.$$

Es gibt zwei Fälle: (i) $u = v$, oder (ii) $u \neq v$. Gilt (i), so sind alle Produkte gleich Null, bei denen eine Person nicht die $u = v$-te Antwortkategorie gewählt hat, und ein Produkt ist gleich 1, wenn eine Person diese Kategorie gewählt hat. Also ist k_{uu} gerade gleich der Anzahl der Personen, die die u-te Kategorie gewählt haben. Ist $u \neq v$, so müssen alle Produkte gleich Null sein, da eine Person nicht

sowohl die u-te wie auch die v-te Kategorie gewählt haben kann. Also ist $Z'Z$ eine Diagonalmatrix, deren Diagonalelemente jeweils die Anzahl der Personen enthält, die die dem Element entsprechende Kategorie gewählt haben.

Jetzt werde der Fall zweier Items mit r_1 bzw. r_2 Antwortkategorien betrachtet. Man kann $Z = [Z_1, Z_2]$ schreiben, womit angedeutet wird, dass die Matrix sich aus zwei nebeneinander geschriebenen Matrizen zusammensetzt. Die Matrix Z' hat die Form

$$Z' = \begin{bmatrix} Z_1' \\ Z_2' \end{bmatrix},$$

und $Z'Z$ ist durch

$$Z'Z = \begin{bmatrix} Z_1'Z_1 & Z_1'Z_2 \\ Z_2'Z_1 & Z_2'Z_2 \end{bmatrix} \tag{4.9.76}$$

gegeben. $Z'Z$ ist eine $(r_1 + r_2)(r_1 + r_2)$-Matrix und aus vier Teilmatrizen zusammengesetzt, wobei $Z_1'Z_1$ und $Z_2'Z_2$ die bereits besprochene Diagonalform haben, – die Diagonalelemente von $Z_2'Z_2$ enthalten die Häufigkeiten, mit denen die einzelnen Kategorien des Items I_2 angekreuzt wurden. Die Matrix $Z_1'Z_2$ ist aber keine Diagonalmatrix; das Element k_{uv} dieser Matrix ist die Häufigkeit, mit der die u-te Kategorie des ersten Items *und* die v-te Kategorie des zweiten Items angekreuzt wurde. Natürlich ist $Z'Z$ symmetrisch, insbesondere ist $(Z_1'Z_2)' = Z_2'Z_1$.

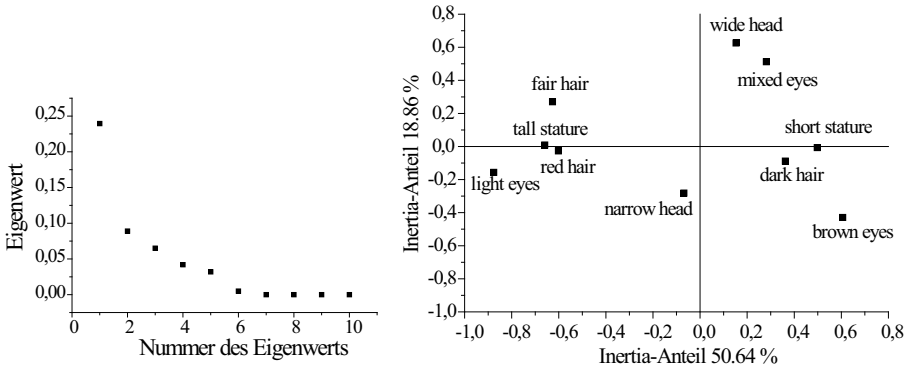
Allgemein kann $Z = [Z_1, \ldots, Z_n]$ sein, d.h. man hat den Fall von n Items, mit jeweils $r_1, r_2, \ldots, r_n$ Antwortkategorien, und $Z'Z$ besteht aus den Teilmatrizen $Z_j'Z_k$, $j, k = 1, \ldots, n$. Dies ist der allgemeine Fall einer Burt-Matrix. Die Burt-Matrix wird nun einer Korrespondenzanalyse unterzogen, d.h. es wird die Singularwertzerlegung von $Z'Z$ berechnet. Da die Burt-Matrix symmetrisch ist, gilt nun $Z'Z = Q'\Lambda^{1/2}Q$, also $P = Q$. Q enthält die noch zu re-skalierenden Koordinaten der Kategorien der einzelnen Items. Die Anzahl der Dimensionen kann man wie bei der PCA aus dem Scree-Plot der Eigenwerte abschätzen. Ein Beispiel illustriert das Vefahren.

Beispiel 4.9.2 Burt (1950) bestimmte für 100 zufällig in Liverpool ausgewählte Personen (i) die Haarfarbe (fair, red, dark), (ii) die Augenfarbe (light, medium, brown), (iii) Kopfform (narrow, wide), und (iv) die Statur (tall, short). Es ergibt sich zunächst eine 4-dimensionale Kontingenztabelle mit jeweils 3, 3, 2 und 2 Kategorien, die Burt-Matrix in Tabelle 19 findet man in Basilevsky (1994), p.532. Die Tabelle hat die Eigenwerte .2395, .0892, .0647, .0422, .0319, .0054, .000, .000, .000. Der Biplot wird in Abbildung 24 gezeigt. Die erste Dimension separiert die Personen in zwei Klassen: die erste Klasse ist durch eine "tall stature" charakterisiert; dieses Merkmal geht mit den Merkmalen "fair hair", "red hair", "light eyes" einher. Personen mit diesen Merkmalen bilden gewissermaßen eine Teilklasse der befragten Personen. Die zweite Teilklasse wird insbesondere durch die Merkmalsgruppe "short stature", "dark hair", "brown eyes" charakterisiert. Die Lage der Staturmerkmale ("tall' versus "short') ist bemerkenswert: sie liegen genau auf der ersten Achse. Die Kopfform – "wide head" versus "narrow head"

Tabelle 19: Burt-Matrix zu genetischen Abhängigkeiten (Burt, 1950); Daten aus Basilevsky (1994)

	fh	rh	dh	le	me	be	nh	wh	ts	ss
fair hair	22	0	0	14	6	2	14	8	13	9
red hair	0	15	0	8	5	2	11	4	10	5
dark hair	0	0	63	11	25	27	44	19	20	43
light eyes	14	8	11	33	0	0	27	6	29	4
mixed eyes	6	5	25	0	36	0	20	16	10	26
brown eyes	2	2	27	0	0	31	22	9	4	27
narrow head	14	11	44	27	20	22	69	0	30	39
wide head	8	4	19	6	16	9	0	31	13	18
tall stature	13	10	20	29	10	4	30	13	43	0
short stature	9	5	43	4	26	27	39	18	0	57

Abbildung 24: Eigenwerte und (Bi)Plot der Merkmale für die Daten aus Tabelle 19



scheint, gekoppelt mit den Augenfarben "mixed" versus "brown" – eine zweite Dimension aufzumachen, wobei es eine geringfügige Assoziation des "wide head" mit dem "short stature"-Pol und des "narrow head" mit dem "tall stature"-Pol der ersten Achse zu geben scheint. Der Verlauf der Eigenwerte legt nahe, dass es noch weitere Dimensionen gibt, oder aber die relativ suggestive Struktur der ersten beiden Dimensionen durch zufällige Kombinationen überlagert wird. □

Wie im Beispiel deutlich wird, liefert die Multiple Korrespondenzanalyse keine die einzelnen Personen charakterisierenden Parameter. Für testpsychologische Zwecke kann man dies als einen Mangel des Verfahrens betrachten. Andererseits kann das Verfahren Hinweise auf die Existenz von Klassen von Personen liefern.

Während die Faktorenanalyse Items durch Punkte in einem Raum geeigneter Dimensionalität repräsentiert, werden bei der Multiplen Korrespondenzanalyse die einzelnen Antwortkategorien durch Punkte abgebildet. Ein einzelnes dichotomes Item wie "stature" mit den Antwortkategorien "tall" versus "short" kann dann nicht nur einen Punkt, sondern eine ganze Dimension definieren.

In Evaluationen etwa von Lehrveranstaltungen werden z.B. Aussagen zum Dozenten, zum Stoff, oder zu Randbedingungen der Lehrveranstaltung (Hörsaal, Medienausstattung, etc) gefordert, oft auf Ratingskalen. So werden die Teilnehmer einer Veranstaltung gebeten, zu beurteilen, ob der Dozent "offen" war, d.h. ob er sich bei Fragen offen verhielt oder eher versuchte, Fragen zu unterdrücken. Das Rating kann von 1 bis 7 erfolgen; 1 bedeutet, dass das Merkmal nicht vorhanden war, 7, dass es asusgeprägt vorhanden war. Eine weitere Frage könnte sich darauf beziehen, ob der Dozent "kompetent" erschien (dises Merkmal sollte allerdings aufgeschlüsselt werden in fachliche Kompetenz, didaktische Kompetenz und soziale Kompetenz); die Ratings können wieder von 1 bis 7 vergeben werden. Weiter kann der Stoff "schwierig" sein, er kann "angemessen" sein, etc., – die Bewertung erfolgt jedes Mal in der gleichen Weise. Die Analyse der Burt-Matrix liefert dann eine Repräsentation, in der möglicherweise Kombinationen von Merkmalen mit bestimmten Skalenwerten auftreten: so könnten die Punkte für Dozent, offen (2) – der Dozent wird dann als so gut wie nicht als "offen" gesehen – mit Stoff, schwierig (6), (7) eng benachbart sein, und in einem anderen Quadranten könnten die Punkte für Dozent, offen (6), (7) und Stoff, schwierig, (1), (2) eng benachbart sein. Man könnte von einer Wechselwirkung zwischen der Beurteilung der Schwierigkeit des Stoffes und der des Dozenten sprechen. In einer PCA würde diese "Wechselwirkung" auch zum Ausdruck kommen, denn die Korrelation zwischen der Beurteilung des Dozenten als "offen" und des Stoffes als "schwierig" wäre negativ, die Punkte für (Dozent, offen) und (Stoff, schwierig) würden nicht im gleichen Quadranten liegen. Die Analyse der Burt-Matrix liefert hier einen direkteren Einblick in die Abhängigkeiten zwischen den Beurteilungen.

Greenacre (1991) hat eine tiefergehende Diskussion der Interpretation der Resultate einer Multiplen Korrespondenzanalyse vorgelegt. Eine weitergehende Diskussion der Multiplen Korrespondenzanalye und Typen von Beurteilern finde man in Hwang, Montréal, Dillon, und Takane (2006).

5 Messmodelle und Prozessmodelle

If you can't measure it, I'm not interested.
Clyde H. Coombs[63]
If you cannot measure, measure anyhow!
Frank H. Knights

[63] Clyde Hamilton Coombs (1912 – 1988), Psychologe, Afficionado der Messtheorie, Begründer des *Mathematical Psychology Program* an der University of Michigan. Frank Hyneman Knight,

Vergleicht man die KTT insbesondere mit dem Rasch-Modell, so besteht der Vorteil des Rasch-Modells sicherlich darin, dass – wenn das Modell auf die Daten passt – die Parameter spezifisch objektiv sind, d.h. die Itemparameter κ sind unabhängig von den Personenparametern θ. Generell werden bei den IRT-Theorien restriktivere Annahmen gemacht, weshalb sie dann auch mehr Informationen als die KTT über das gemessene Merkmal liefern, vorausgesetzt, der Fit ist akzeptabel. Die Bestimmung des Parameters θ_a für eine getestete Person $a \in \mathcal{P}$ ist aufwendiger als die Berechnung des Summenscores bei einem Test, der im Rahmen der KTT entwickelt wurde, ist aber wegen der Verfügbarkeit von Computern und entsprechenden Programmen kein Problem mehr.

Spezifische Objektivität: Die Frage ist aber, ob spezifische Objektivität ein notwendiges Merkmal eines Tests sein muß. Einige Autoren, etwa Fischer (1974), Fischer & Molenaar (1995) scheinen dies zu suggerieren. Die Idee geht auf Rasch selbst zurück. Andrich (2004) beschreibt die Entwicklung dieser Idee unter Bezug auf Kuhns (1961) Begriff des wissenschaftlichen Paradigmas. Entsprechend dem traditionellen Paradigma würde man ein Modell M_1 dem Modell M_2 vorziehen, wenn M_1 besser als M_2 mit den Daten übereinstimmt: "Invariably, it is models that are accepted or rejected, given the data." (Andrich, 2004, p. 8). Raschs Anspruch an Modelle geht aber über die Forderung nach einem guten "Fit" hinaus. Er postuliert, dass der Vergleich zweier Personen unabhängig von den Items aus einer Klasse von Items sein sollte, und umgekehrt sollte der Vergleich von Items bezüglich ihrer Schwierigkeit unabhängig von den getesteten Personen sein. Diese Forderung nach spezifischer Objektivität wird zunächst ohne Bezug auf ein spezielles Modell aufgestellt: *jedes* Testmodell, das zur Grundlage der Konstruktion eines Tests gewählt wird, sollte dieser Forderung genügen. Spezifische Objektivität ist, nach Rasch und anderer Autoren, eine notwendige, wenn auch keine hinreichende Bedingung für ein gutes Testmodell.

Rasch hatte zunächst mit gutem Erfolg eine Menge von Testdaten "erklärt", indem er die Poisson-Verteilung zur Modelierung der Fehlerhäufigkeiten benützte; die logistische Verteilung wurde dann nur der mathematischen Einfachheit bei gleichzeitigem guten Fit eingeführt. Rasch (1961) berichtet, dass sich die Einsicht in die Implikationen der logistischen Verteilung, insbesondere des 1PL-Modells, nur langsam einstellte. Es war Ragnar Frisch[64], den es in einer Diskussion mit Rasch verwunderte, dass je nach Vergleich von Personen oder Items sich die Items oder die Personen gewissermaßen herauskürzen. Erst durch Frischs Verwunderung wurde es Rasch klar, dass mit der Separierbarkeit der Parameter eine fundamental wichtige Klasse von Modellen gefunden worden war. Andrich (2004) ist der Ansicht, dass dieser Prozess exakt in das Muster passt, das Kuhn in seiner Analye von Paradigmenwechseln beobachtet hat: es ist die Lösung eines Problems, die neue Einsichten schafft. Rasch betrachtete nun nicht mehr mögliche Modelle, die mehr oder weniger gut an Daten angepasst werden können, sondern er fokussierte

1885 – 1972, Ökonom, Begründer der marktradikalen *Chicago School* der Ökonomie.

[64]Ragnar Frisch, 1895–1973, norwegischer Ökonom, Nobelpreis 1969

auf Modelle, von denen er annahm, dass sie die gleiche parametrische Struktur wie die Gesetze der Physik haben.

Suffiziente Statistik: Ein zweites Merkmal des Rasch-Modells ist, dass der Testscore, d.h. die ungewichtete Summe der Scores für die einzelnen Items, eine suffiziente Statistik für die Schätzung des Parameters θ_a für eine Person a ist. Suffizienz bedeutet, dass die Statistik alle Information, die in den Daten bezüglich eines zu schätzenden Parameters vorhanden ist, enthält. In Rasch (1977) liefert Rasch das, was nach Kuhn (1961) ebenfalls typisch für Paradigmenwechsel ist: Rasch wendet sich, so Andrich, philosophischen Fragen zu. Dabei geht es um die Frage, welcher Norm Forschung folgen sollte: (i) festzustellen, ob ein Modell gegebenen Daten entspricht, oder umgekehrt (ii), Daten zu finden, die einem Modell entsprechen. Der Vorrang sei der letzten Norm zu geben: die Daten müssen nach Maßgabe eines den Forderungen der Theorie der fundamentalen Messung entsprechenden Modells erhoben werden. Rasch nahm die grundlegenden Gesetze der Physik als Beispiel, insbesondere Newtons zweites Axiom, demzufolge $K = mb$ gilt, d.h. die Kraft K, die auf einen Körper wirkt, entspricht seiner Beschleunigung b, mit dem Proportionalitätsfaktor m, der die Masse des Körpers repräsentiert. Die Form dieses Gesetzes – das Produkt mb definiert die Kraft K – entpricht der multiplikativen Parametrisierung des Rasch-Modells (4.3.4) auf Seite 214: über das Produkt $\delta\sigma_g$ wird die Wahrscheinlichkeit, eine Aufgabe zu beantworten, definiert, diese Wahrscheinlichkeit entspricht der Kraft K in Newtons zweitem Axiom. Ebenso, wie sich δ und σ_g unabhängig voneinander schätzen lassen, kann man m und b bestimmen: für eine fixe Kraft K und zwei Körper mit den Massen m_1 und m_2 folgt $m_1 b_1 = m_2 b_2$, d.h.

$$\frac{m_1}{m_2} = \frac{b_2}{b_1}; \tag{5.0.77}$$

es genügt demnach, nur die Beschleunigungen b_1 und b_2 unabhängig von den Geschwindigkeiten zu bestimmen, um mit dem Quotienten b_2/b_1 ein Messungen für m_1 (in Einheiten von m_2) zu erhalten. Ramsay (1975) merkte an, dass die meisten grundlegenden Gesetze der Physik eine bemerkenswert einfache Form haben: sie könnten als Multiplikationen und Divisionen fundamentaler Messungen dargestellt werden. Eine entsprechende "rationalization of quantification may be necessary precondition to Psychology as a Quantitative Rational Science" (Krantz, Luce & Suppes (1971)).

Wie Rasch (1961), Wright (1985) und Fischer (1995) gezeigt haben, ist das Rasch-Modell – also ein 1PL–Modell auf der Basis der logistischen Funktion – das einzige Modell, das eine Separation der Parameter im Sinne der spezifischen Objektivität zuläßt. Das Rasch-Modell liefere "an operational criterion for fundamental measurement of the kind found in the physical sciences. " (Andrich (2004), p. 12) Also gäbe es nur dann sinnvolle Messungen, wenn sie im Rahmen des Rasch-Modells vorgenommen werden.

Wissenschaftstheoretische Betrachtungen: Raschs Betrachtungen sind wis-

senschaftstheoretischer Art. Die Notwendigkeit fundamentaler Messung wird durch Analogie zu den grundlegenden Gesetzen der Physik postuliert. Die Frage ist, wie weit diese Analogie getrieben werden kann. Die *a priori*-Aussage, die mit dieser Analogie gemacht wird, ist ja, dass alle fundamentalen Gesetzmäßigkeiten in der Welt eine Struktur haben, die spezifisch objektive Messungen implizieren. Das Newtonsche zweite Axiom wäre ein Beispiel, – ist es aber nicht, denn seit Einstein (1905) ist bekannt, dass dieses Axiom nur eine Approximation für den Fall kleinerer Geschwindigkeiten ist. Nach der Speziellen Relativitätstheorie gilt für die Masse m die Beziehung

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}},$$

wobei m_0 die Ruhemasse, v die Geschwindigkeit des Körpers und c die Lichtgeschwindigkeit ist. Newtons Formel geht dann über in

$$F = \frac{d}{dt}(mv) = \frac{d}{dt}\left(\frac{m_0 v}{\sqrt{1 - v^2/c^2}}\right) = \frac{m_0(\sqrt{1 - v^2/c^2} + v^2/(tc)^2)}{1 - v^2/c^2}.$$

(Beiser, 1984, p. 22–23) Die einfache Beziehung (5.0.77) mit einer von v unabhängigen Masse gilt nur dann, wenn man v konstant hält. Für fixen Wert von v kürzen sich die von v abhängigen Terme heraus und (5.0.77) wird zu

$$\frac{m_{01}}{m_{02}} = \frac{b_2}{b_1}, \quad v_1 = v_2. \tag{5.0.78}$$

Für verschiedene Beschleunigungen bei gleicher Geschwindigkeit erhält man nun einen Vergleich der Ruhemassen, – aber nicht der tatsächlichen, für $v = v_1 = v_2$ charakteristischen Massen. Das ist möglicherweise etwas anderes als das, was man mit den Messungen bezweckt hat. Ist man sich der Beziehung zwischen Masse und Geschwindigkeit nicht bewußt, so fokussiert man u. U. nur auf die Beschleunigungen und nimmt die Messungen bei verschiedenen Geschwindigkeiten vor, so dass sich eben geschwindigkeitsabhängige Messwerte ergeben, ohne dass sie als solche interpretiert würden.

Sicherlich ist es so, dass für Geschwindigkeiten v, die klein gegen die Lichtgeschwindigkeit c sind, die relativistische Formel für F nicht gebraucht wird, aber darum geht es hier gar nicht. Es geht darum, dass die einfachen Gesetze im Allgemeinen nur *ceteris paribus*-Aussagen sind, also Aussagen, die nur bei Konstanthaltung bestimmter Randbedingungen gelten, und dass daraus spätestens dann ein Problem entsteht, wenn diese Randbedingungen nicht vollständig bekannt sind. Im relativ einfachen Fall des zweiten newtonschen Axioms kennt man die Randbedingungen und kann daher ziemlich genau sagen, wie weit man mit der Anwendung dieses Axioms kommt und wann man die Einsteinsche Verallgemeinerung berücksichtigen muß. Man muß sich an dieser Stelle aber klar machen, dass die Einsteinsche Verallgemeinerung gar nicht möglich wäre, würde man sich in

der Physik an die ontologische Forderung halten, grundlegende Formeln müßten nun mal von der einfachen Form $K = mb$ sein. Anhand dieses relativ einfachen Beispiels wird deutlich, dass ein messtheoretisches Postulat kein ontologisches Postulat sein kann. Bei psychologischen "Gesetzen" ergibt sich das hier diskutierte Problem aber verschärft, weil die Randbedingungen oft gar nicht explizit bekannt sind. Die Form der Beziehungen zwischen Variablen, von denen angenommen wird, dass sie das Verhalten von arbeitslosen Jugendlichen beschreiben, kann stark von der Wohnumgebung der Jugendlichen – Essen-Süd oder Memmingen/Allgäu – abhängen. Dies ist ein triviales, weil offenkundiges Beispiel, das gleichwohl die ontologischen Implikationen der einfachen, fundamentalen Relationen in Frage stellt. Sogar bei den hochkontrollierten Untersuchungen der Psychophysik kann es zu unkontrollierten Effekten intervenierender Variablen kommen, die zu verzerrenden Urteilen bei Untersuchungen einfach erscheinender Zusammenhänge führen. Man denke etwa an Riccos Gesetz zur räumlichen Summation, demzufolge $a \times c =$Konstante gilt, wobei a die Größe der Fläche des Stimulus und c der Kontrast des Reizmusters ist. Aufmerksamkeitsfluktuationen können zu verzerrenden Schätzungen der Konstante führen, weil die Summation eben auch vom Ausmaß der Aufmerksamkeitsfokussierung abhängen kann. Um den Effekt solcher Variablen herauszufinden, müssen Messungen durchgeführt werden, – von denen nicht *a priori* gefordert werden kann, dass sie spezifisch objektiv sind. Wie es scheint, impliziert die Forderung nach spezifischer Objektivität einen logischen Zirkel. Die Forderung nach spezifischer Objektivität imponiert mit dem Status eines Axioms, dessen Gültigkeitsanspruch *jeder* Empirie vorangeht. Damit bekommt sie aber den Status einer metaphysischen Wahrheit zugeordnet, von der postuliert wird, sie stünde jedem Einsichtigen vor Augen. Vor Einstein (1905) hat man in Bezug auf das newtonsche zweite Axiom so gedacht, aber Einstein hat gezeigt, das Evidenzerlebnisse nicht Erlebnisse von Wahrheit sein müssen.

Das Rasch-Modell ist ein Latent-trait-Modell, das sich als Verteilungsfunktion einer unterliegenden Variablen ergibt, und zwar als 2PL-Modell, mit $\alpha_g \propto 1/\sigma_g$, wobei σ_g die Streuung der unterliegenden Variable, also des fluktuierenden Merkmals, ist. Wie Bartholomew (1987) gezeigt hat, ist jedes Response Function-(RF)-Modell einem Modell für unterliegende Variablen äquivalent (vergl. Anhang, Abschnitt A.3). Dem 2PL-Modell entspricht also ein Modell, in dem angenommen wird, dass das fluktuierende Merkmal logistisch verteilt ist. Umgekehrt entspricht einem solchen Modell das Response Function-(2PL)-Modell. Das Rasch-Modell entsteht daraus, wenn $\alpha_g =$ Konstante gesetzt wird, die dann in θ_a und κ_g absorbiert wird. Man wird also auf die Frage geführt, warum, um der fundamentalen Messung und der damit verbundenen Separierbarkeit willen, die Streuungen für alle Items gleich groß sein sollen. Dabei hat man hier schon eine Einschränkung, da das 2PL-Modell ja fordert, dass σ_g für alle $a \in \mathcal{P}$ gleich sein soll. Die Forderung nach fundamentaler Messung und Separierbarkeit der Parameter impliziert Aussagen, von denen nicht klar ist, warum sie für einen gegebenen Aufgabenbereich gelten sollen. Nun könnte man meinen, dass das Rasch-Modell ja gar

nicht explizit in Bezug auf ein fluktuierendes Merkmal definiert sei, der Personenparameter θ_a sei einfach eine Konstante, die die Person a charakterisiere. Andererseits repräsentiert aber die Itemfunktion irgendetwas Stochastisches, da die Itemfunktion ja nur die Wahrscheinlichkeit einer ("korrekten") Antwort liefert. Dieser stochastische Prozess, der zur Beantwortung des Items führt, wird im Rahmen des Rasch-Modells unspezifiziert gelassen. Also könnte dieser Prozess doch vielleicht so geartet sein, dass er *nicht* auf eine Forderung nach gleichen σ_g-Werten hinausliefe. Nach dem Bartholomewschen Theorem muß dann aber dieser Prozess wiederum äquivalent dem Verteilungsfunktionsmodell mit gleichen Varianzen sein, da ja beide der Itemfunktion äquivalent sind. Das hieße, dass ein Prozess, der nicht durch zufällige Veränderliche mit ungleichen Varianzen für die Aufgaben und Personen repräsentiert wird, stets einem Prozess mit entsprechenden gleichen Varianzen äquivalent ist. Die Konstruktion zweier Prozesse, die (i) einander äquivalent sind, und von denen (ii) einer durch zufällige Variable mit gleichen Varianzen für Aufgaben und Probanden und der andere keine solchen Variablen zu seiner Repräsentation benötigt, steht noch aus. Vermutlich ist die Anzahl der Elemente einer Klasse solcher Paare von Prozessen sehr gering, wenn nicht gar gleich Null. Dies hieße dann, dass das Rasch-Modell nur einen äußerst eingeschränkten Anwendungsbereich hat.

Nach Ansicht einiger Anhänger des Rasch-Modells sind solchen Betrachtungen aber nicht nur irrelevant, sondern sogar oberflächlich ("shallow"), denn für die oben durchgeführte Betrachtung, nach der das Rasch-Modell einfach ein Spezialfall eines allgemeineren (Birnbaum-)Modells ist, wird die folgende Kritik formuliert:

> This approach seems to imply that the Rasch model is just a stripped-down version of more complicated models which "must be better" because they account for more of the "presumed reality" of traditional test theory. Quite apart from Occam's razor (that entities are not multiplied beyond necessity), this interpretation is shallow in an essential way. That the Rasch model can be reached by simplifying more complicated models has nothing to do with its genesis or rationale, or with the theory of measurement.
>
> The Rasch model is not intended to fit data or to be evaluated by how well it fits any particular data set. The Rasch model is a definition of measurement derived from the universally accepted measurement requirements that:
>
> 1. The measures of objects be free of the particulars of the agents used to estimate these measures and the calibrations of agents be free of the particulars of the objects used to estimate these calibrations.
> 2. The measures of objects and calibrations of agents function according to the rules of arithmetic on a common scale so that they

can be analyzed statistically.

3. Linear combinations of measures and calibrations correspond to plausible concatenations of objects and agents.

> Introducing the superfluous parameters, a_i and c_i [korrespondierend zu α_g und γ_g], violates these requirements and disqualifies the resulting contrivances as "measurement" models. Were the values of t known prior to analysis, the 2- and 3-parameter formulations could be used to specify two, among many possible, logistic regressions, as they have been in biometric analysis for 60 years. When t [entsprechend θ] is not known and must be estimated from the same data, as is the case for measurement, then it can be demonstrated algebraically that the parameter estimates must diverge unless constrained in an arbitrary manner.
>
> In practice, arbitrary interference, such as limiting the range of a_i and c_i , does not produce sample-free item calibrations or test-free person measures - unless a_i and c_i are restricted to constants, when, of course, the formulation becomes a Rasch model and functions accordingly. This is the misunderstanding of those who claim the 2- and 3-parameter computer programs "work". In the few instances where serious 2- and 3-parameter results are reported, their quality is proportional to the extent to which a_i and c_i were kept from varying - from being influenced by the data. Indeed all important practical IRT applications reported at the 1991 AERA meeting were implemented by the Rasch model.
>
> In theory and in practice, the Rasch model is a unique measurement archetype.
>
> Shaw, F. (1991) Descriptive IRT vs. Prescriptive Rasch, ... Rasch Measurement Transactions, 5:1 p. 131
> (http://www.rasch.org/rmt/rmt51f.htm)

Das sind starke Überzeugungen, insbesondere der Punkt 1 erscheint absolut zwingend. Da Autoren wie Fischer auch der Suffizienz der Schätzungen einen ähnlich fundamentalen Status zuordnen, scheint das Rasch-Modell das einzige Modell zu sein, das wissenschaftlichen Messungen zugrunde gelegt werden kann. Die folgenden Betrachtungen relativieren allerdings diese Forderung.

Testen und das Modellieren von Prozessen: Die Repräsentation von Items und Personen durch bestimmte Parameter entspricht einem Messprozess. Wenn aber messen die grundlegende Tätigkeit der Wissenschaft schlechthin ist, ergibt sich die Frage, ob auch Prozessmodelle dem Prinzip der spezifischen Objektivität gehorchen und also einem Rasch-Modell genügen müssen. Dass die Annahme des Rasch-Modells in Konjunktion mit der Annahme einer in den Schwierigkeiten von

Aufgabenkomponenten linearen Logit-Funktion die Annahme ganz spezifischer, aber nicht explizit gemachter Abhängigkeiten zwischen den Lösungskomponenten bedeutet, ist bereits in Abschnitt 4.6.3 ausführlich diskutiert worden: eine zunächst harmlos wirkende Annahme (lineare Logit-Funktion) impliziert Annahmen über Abhängigkeiten, deren psychologische Struktur völlig unklar bleibt, die aber bei der Interpretation der Messungen mitgedacht werden müssen.

Es soll nun anhand eines von Vorberg & Schwarz (1990) diskutierten Beispiels verdeutlicht werden, dass insbesondere die Forderung nach spezifischer Objektivität keineswegs notwendig erfüllt werden kann.

Vorberg et al.s Untersuchungen wurden durch eine Arbeit Mickos (Micko, 1969) angeregt. Mickos Idee war, der Forderung Raschs nach Separierbarkeit von Personen- und Stimulusparametern zu folgen und für Reaktionszeiten, die ja auch in testpsychologischen Fragestellungen eine Rolle spielen können, ebenfalls ein Rasch-Modell anzusetzen. Er ging dabei von der multiplikativen Parametrisierung (4.3.4), Seite 214, aus. Seinem Ansatz zufolge kann die Verteilung einer Reaktionszeit bei der i-ten in der Form (Vorberg et al.s Notation)

$$F_i(\tau \leq t) = \frac{1}{1 + a(i)b(t)} \tag{5.0.79}$$

dargestellt werden[65]. Hierin ist $a(i)$ jetzt eine Funktion, die die i-te Person charakterisiert, und $b(t)$ ist eine Funktion der Zeit, die beliebig ist bis auf die Restriktion, dass F_i eine Verteilungsfunktion sein muß. Formt man (5.0.79) nach dem Produkt $a(i)b(t)$ um, so erhält man

$$a(i)b(t) = \frac{F_i(t)}{1 - F_i(t)}.$$

Also muß $a(i)b(t) \geq 0$ für alle i und t gelten. Nimmt man $a(i) > 0$ an, folgt $b(t) \to \infty$ für $F_i(t) \to 1$, und $b(t) \to 0$ für $F_i(t) \to 0$. Also muß $0 \leq b(t) \leq \infty$ gelten. Eine Verteilungsfunktion für Zeiten heißt *Rasch-repräsentierbar*, wenn sie in der Form (5.0.79) dargestellt werden kann.

Nun gibt es Modelle, die die Struktur von Reaktionszeiten abbilden. Eines geht auf Donders (1868) zurück: es wird angenommen, dass bei bestimmten komplexeren Aufgaben die Lösung im seriellen Abarbeiten von Teilaufgaben besteht, für die eine jeweils zufällig verteilte Zeit vergeht. Es ist ein Modell *serieller* Informationsverarbeitung. Die Gesamtzeit bis zur Lösung ist dann die Summe T dieser Teilzeiten, etwa

$$T = S_1 + S_2 + \cdots + S_n, \tag{5.0.80}$$

wobei die S_j die Teilzeiten sind. Die Verteilungsfunktion für T, $F(t) = P(T \leq t)$, hängt dann von den Verteilungsfunktionen für die S_j ab und davon, ob die S_j

[65] Dem dem Rasch-Modell äquivalenten Verteilungsmodell löst eine Person die g-te Aufgabe, wenn $P(\eta > \kappa_g) = \delta\sigma_g/(1+\delta\sigma_g)$. Dann ist $P(\eta \leq \kappa_g) = 1/(1+\delta\sigma_g)$; dieser Ausdruck entspricht einer Verteilungsfunktion.

stochastisch unabhängig sind oder nicht. Bei einer alternativen Klasse von Modellen wird angenommen, dass n Teilprozesse *parallel* und unabhängig voneinander ablaufen. Die Reaktionszeit ist die Zeit, die der am schnellsten ablaufende Prozess benötigt, um die Aufgabe zu lösen. Die Begriffe der seriellen bzw. parallelen Prozesse müssen noch ein wenig spezifiziert werden, worauf hier nicht weiter eingegangen werden kann. Vorberg & Schwarz' Befund ist jedenfalls, dass für eine sehr allgemeine Klasse dieser beiden Modelle keine Rasch-repräsentierbare Verteilungsfunktion gefunden werden kann.

Es handelt es sich bei den genannten Modellen um grundlegende Modelle, die für bestimmte Aufgabentypen modifiziert werden können. Der Punkt ist, dass solche Modelle offenbar nicht im Rahmen des Rasch-Modells diskutiert werden können. Fordert man also fundamentale Messungen im Rahmen des Rasch-Modells, so schließt man derartige Modelle grundsätzlich aus der Betrachtung aus. Die Implikation ist, dass die Forderung nach Rasch-Messbarkeit ausschließt, dass Prozesse entweder seriell oder parallel ablaufen können. Natürlich gibt es keinen direkten Zusammenhang zwischen der messtheoretischen Forderung und der Art, in der Informationen im Gehirn verarbeitet werden. Die messtheoretische Forderung erzwingt aber, wenn man sie denn ernst nimmt, dass bestimmte Arten von Informationsverarbeitung gar nicht existieren können. Es fällt schwer, diese Implikation nicht als abstrus zu empfinden.

Vorberg et al. stellten dann noch die Frage, ob es überhaupt eine prozessorientierte Interpretation des Mickoschen Ansatzes gibt. Diese gibt es: die entsprechenden Verteilung läßt sich als zusammengesetzte (*compound*) Verteilung des Maximums einer Stichprobe von zufälligen Veränderlichen darstellen:

$$F_j(t) = \sum_{m=1}^{\infty} P(N_j = m) P\left(\max_{k \in \{1,\ldots,M\}} C_k \leq t \right), \tag{5.0.81}$$

wobei die C_k unabhängige und identisch verteilte (iid) zufällige Veränderliche mit der Verteilungsfunktion $F_i(t)$ und die Stichprobenumfänge N_j geometrisch verteilt sind, $P(N_j = m) = c_{ij}(1 - c_{ij})^{m-1}$. Natürlich muß $c_{ij} \leq 1$ sein; dies impliziert wiederum, dass die Verteilungen als Verteilungen zufälliger Maxima von iid-Variablen[66] einer bestimmten Verteilung, etwa F_1, dargestellt werden können. Vorberg & Schwarz zeigen, dass die Rasch-repräsentierbaren Verteilungen auch als zufällige Minima dargestellt werden können:

$$F_j(t) = 1 - \sum_{m=1}^{\infty} P(N_j = m) P\left(\min_{k \in \{1,\ldots,M\}} C_k > t \right), \tag{5.0.82}$$

wobei die N_j wieder geometrisch verteilt sind. Wie die Autoren ausführen, besteht das Problem bei den Lösungen (5.0.81) und (5.0.82) allerdings darin, die

[66] iid = **i**ndependent and **i**dentically **d**istributed; die in der Statistik übliche Abkürzung für unabhängige, identisch verteilte Variablen.

Forderung nach der geometrischen Verteilung der N_j zu motivieren, – man hat im Prinzip keine Freiheiten bei der Modellierung des zur Diskussion stehenden Prozesses mehr. Vorberg & Schwarz zeigen weiter, dass Rasch-repräsentierbare Verteilungsfamilien nur eine Skalenfamilie $b(t)$ zulassen, nämlich Skalen der Form

$$b(t) = rt^s, \tag{5.0.83}$$

wobei r und s Konstante sind. Diese Implikation bezüglich der ursprünglich als nicht weiter festgelegt gedachten Funktion $b(t)$ zeigt, wie drastisch die messtheoretische Forderung

> "To construct measures, we require orderly,cooperating, noncrossing curves like Rasch curves ... This means that we must take the trouble to collect and refine data so that they serve this clearly defined purpose ..." (Wright (1997), zitiert nach Andrich (2004), p. 13).

nach einer a-priori-Wahl des Rasch-Modells die Menge der Prozesse, die mit dieser Wahl kompatibel sind einengt.

Man mag argumentieren, dass in "normalen" Tests ja gar keine Prozessmodelle betrachtet werden. Personen werden nur durch einen Parameter θ charakterisiert, damit sei man ja schon fertig. Abgesehen davon, dass die Charakterisierung durch einen konstanten Parameter bereits eine Vereinfachung darstellt, ist aber doch gar nicht gesichert, dass keine Wechselwirkung zwischen der Ausprägung der gemessenen Fähigkeit und der Schwierigkeit eines Items besteht. Ein guter Kopfrechner muß das Große Einmaleins nicht auswendig lernen, er kann womöglich in der gleichen Zeit, in der jemand, der gar nicht kopfrechnen kann, aber das Einmaleins perfekt auswendig gelernt hat, die Antwort auf eine gegebene Frage errechnen. Aber der Kopfrechner und der Lerner können auf die gleiche Anzahl gelöster bzw. korrekt beantworteter Fragen kommen und bekommen also den gleichen Fähigkeitsparameter zugeordnet, – obwohl völlig verschiedene kognitive Fähigkeiten gemessen wurden.

A Anhang

A.1 Cohens κ für $P_r \to 0$ und $P_r \to 1$

Aus (2.7.36) folgt

$$\lim_{P_r \to 0} \alpha_A \alpha_B P_r + (1-\beta_A)(1-\beta_B)(1-P_r) = (1-\beta_A)(1-\beta_B),$$

und aus (2.7.37) erhält man $\lim_{P_r \to 0}(1-\alpha_A)(1-\alpha_B)P_r + \beta_A\beta_B(1-P_r) = \beta_A\beta_B$. Also erhält man

$$\lim_{P_r \to 0} P_0 = \lim_{P_r \to 0} P(C_+) + \lim_{P_r \to 0} P(C_-) = (1-\beta_A)(1-\beta_B) + \beta_A\beta_B \qquad \text{(A.1.1)}$$

Nach (2.7.21) kann P_c in der Form

$$P_c = P(A_+)P(B_+) + P(A_-)P(B_-) \qquad \text{(A.1.2)}$$

geschrieben werden. Nach (2.7.27) ist aber $\lim_{P_r \to 0} P(A_+) = 1-\beta_A$, $\lim_{P_r \to 0} P(B_+) = 1-\beta_B$, und $\lim_{P_r \to 0} P(A_-) = 1 - \lim_{P_r \to 0} P(A_+) = \beta_A$, $\lim_{P_r \to 0} P(B_-) = 1 - \lim_{P_r \to 0} P(B_+) = \beta_B$. Also ist

$$\lim_{P_r \to 0} P_c = (1-\beta_A)(1-\beta_B) - \beta_A\beta_B. \qquad \text{(A.1.3)}$$

So erhält man schließlich (vergl. (A.1.1)

$$\lim_{P_r \to 0}(P_0 - P_c) = (1-\beta_A)(1-\beta_B) + \beta_A\beta_B - ((1-\beta_A)(1-\beta_B) + \beta_A\beta_B) = 0 \qquad \text{(A.1.4)}$$

und damit

$$\lim_{P_r \to 0} \kappa = 0. \qquad \text{(A.1.5)}$$

In völlig analoger Weise zeigt man, dass

$$\lim_{P_r \to 1} \kappa = 0. \qquad \text{(A.1.6)}$$

A.2 Erwartungswerte, Varianzen, Kovarianzen und Korrelationen

Es sei X eine zufällige Veränderliche, und es mögen n "Messungen" (Realisierungen) von X vorliegen. Dann läßt sich das arithmetische Mittel der x_i, $i = 1, \ldots, n$ berechnen:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{(A.2.1)}$$

Die x_i mögen in k Gruppen mit jeweils identischen Werten vorliegen, und zwar n_1 Werte in der ersten Gruppe, alle Messungen in dieser Gruppe haben den Wert

x_1. In der zweiten Gruppe mit n_2 Werten haben alle Messungen den Wert x_2, etc. Dann läßt sich $\bar{x}$ in der Form

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j x_j = \sum_{j=1}^{k} \frac{n_j}{n} x_j = \sum_{j=1}^{k} \hat{p}_j x_j \tag{A.2.2}$$

schreiben. Man kann $\hat{p}_j$ als eine Schätzung der Wahrscheinlichkeit p_j, dass X den Wert x_j annimmt, auffassen.

$\bar{x}$ ist der Mittelwert – das arithmetische Mittel – für eine gegebene Stichprobe von Messwerten. Man stelle sich für den Moment vor, dass man *alle* Werte von X messen könnte und dann von diesen Messungen das arithmetische Mittel berechnen würde. Dieser Mittelwert wird als *Erwartungswert* von X bezeichnet. Die Bezeichung für den Erwartungswert ist $E(X)$; in diesem Skriptum wird aber die neuere Bezeichnung $\mathbb{E}(X)$ gewählt, weil sie eindeutiger ist und vor Verwechslungen schützt; so wird gerade in der Testtheorie der Messfehler oft mit dem Buchstaben E bezeichnet.

Zufällige Veränderliche können diskret oder kontinuierlich bzw. stetig sein. X ist *diskret*, wenn die möglichen Realisierungen (Messungen) durchnummeriert werden können, also in der Form X_1, X_2, $X_3, \ldots$ angeschrieben werden können. Beispiele ergeben sich aus dem Münzwurf: $X = 1$, wenn etwa der Kopf oben liegt, $X = 0$, wenn die Zahl oben liegt, oder umgekehrt. Jedenfalls kann X nur zwei Werte anehmen. Beim Würfeln kann X die Werte $X_1 = 1$, $X_2 = 2, \ldots X_6 = 6$ annehmen. Die mögliche Anzahl von X-Werten kann auch unendlich sein, etwa die Anzahl der Versuche, eine bestimmte Aufgabe zu lösen. X ist *stetig*, wenn X alle Werte aus einem Intervall der reellen Zahlen annehmen kann, wobei das Intervall auch $(-\infty, \infty)$ sein kann. Ist X zB eine Reaktionszeit, so sind die möglichen Werte alle Zahlen aus dem Intervall $(0, \infty)$. Für den Erwartungswert von X ergibt sich dann die Definition

$$\mathbb{E}(X) = \begin{cases} \sum_j p_j x_j, & X \text{ ist diskret} \\ \\ \int_{-\infty}^{\infty} x f(x) dx, & X \text{ ist stetig} \end{cases} \tag{A.2.3}$$

Da man kaum jemals *alle* möglichen Realisierungen von X "messen" kann, scheint der Erwartungswert zunächst garnicht angebbar zu sein. Andererseits enthält die Definition (A.2.3) die Wahrscheinlichkeiten p_j bzw. die Dichtefunktion f. Damit wird $\mathbb{E}(X)$ berechenbar, wenn man bestimmte Annahmen über die p_j bzw. über f macht. Der Erwartungswert hängt also nicht nur vom Wertebereich der Variablen X ab, sondern auch vom Wahrscheinlichkeitsmaß, das X zugeordnet wird.

Beispiele: Bei einer Münze liege der Kopf mit der Wahrscheinlichkeit p oben, und die Zahl mit der Wahrscheinlichkeit $q = 1 - p$. Die zufällige Veränderliche nehme die Werte $X = 1$ (Kopf) und $X = 0$ (Zahl) an. Der Erwartungswert ist dann

$$\mathbb{E}(X) = p \cdot 1 + q \cdot 0 = p \cdot 1 = p. \tag{A.2.4}$$

Bei dieser Definition von X kann $\mathbb{E}(X)$ also keinen Wert annehmen, den X bei einem beliebigen Münzwurf annimmt. Hat man nun eine konkrete Folge von Resultaten eines Münzwurfexperiments vorliegen, etwa 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, so kann man das arithmetische Mittel der Werte berechnen: nach (A.2.2) ist $\bar{x} = (6/10) \cdot 1 + (4/10) \cdot 0 = 6/10 = \hat{p}$. $\hat{p}$ ist die relative Häufigkeit, mit der "Kopf" oben lag und ist damit eine Schätzung der Wahrscheinlichkeit p. Es steht einem natürlich frei, X anders zu definieren, etwa $X = +1$, wenn nach einem Wurf "Kopf" oben liegt, und $X = -1$, wenn "Zahl" oben liegt. Der Erwartungswert ist in diesem Fall

$$\mathbb{E}(X) = p \cdot 1 + q \cdot (-1) = p - q = p - 1 + p = 2p - 1. \tag{A.2.5}$$

Für die übliche Annahme $p = 1/2$ erhält man den Erwartungswert $\mathbb{E}(X) = 0$. Logisch ist diese Definition von X gleichberechtigt mit der ersten, die aber den praktischen Vorteil hat, dass $\mathbb{E}(X)$ einer Schätzung von p entwpricht. Man beachte, dass der logischen Möglichkeit, dass die Münze auf dem Rand stehen bleibt, in diesen beiden Beispielen die Wahrscheinlichkeit 0 zugeordnet wurde; diese Zuordnung ist zwar plausibel, weil die Anzahl der Male, bei denen eine Münze auf dem Rand stehen bleibt, verschwindend gering ist, die Annahme einer Wahrscheinlichkeit von Null ist aber keineswegs logisch zwingend.

Beim Würfel hat man $X_j = j$, $j = 1, \ldots, 6$. Der Erwartungswert ist dann

$$\mathbb{E}(X) = p_1 \cdot 1 + p_2 \cdot 2 + \cdots + p_6 \cdot 6, \tag{A.2.6}$$

wobei p_j die Wahrscheinlichkeit für die Augenzahl j ist und natürlich $\sum_j p_j = 1$ gelten muß. Nimmt man an, dass $p_j = p$ für alle j gilt, so folgt $p = 1/6$ und man hat

$$\mathbb{E}(X) = \sum_{j=1}^{6} p_j j = p \sum_{j=1}^{6} j = \frac{1}{6} \frac{6(6+1}{2} = 3.5 \tag{A.2.7}$$

Wieder ist $\mathbb{E}(X)$ nicht gleich einer der möglichen Realisierungen von X.

Abschließend werde noch der Fall einer unendlich großen Anzahl von Werten von X betrachtet. X sei die Anzahl der Versuche bis zum ersten Versuch, bei dem eine schwierige Aufgabe gelöst wurde. Die Wahrscheinlichkeit, dass bei einem einzelnen Versuch die Aufgabe gelöst wird, sei $p < 1$, und es werde die (vereinfachende) Annahme gemacht, dass alle Lösungsversuche voneinander (stochastisch) unabhängig sind und p konstant bleibt. Es muß geklärt werden, was mit der Ausdrucksweise " bis zum ersten Versuch" gemeint ist: X kann die Anzahl der Versuche sein, bei denen die Aufgabe nicht gelöst wurde, der erste Versuch, bei dem die Aufgabe gelöst wurde, ist dann der $(X + 1)$-te Versuch. Oder man hat in $X - 1$ Versuchen die Aufgabe nicht gelöst und löst sie beim X-ten Versuch. Ist x die Anzahl der Mißerfolge, so tritt der Erfolg beim $(x + 1)$-ten Erfolg ein und man hat

$$\mathbb{E}(X) = \sum_{j=1}^{\infty} j \cdot P(X = j) = \sum_{j=1}^{\infty} j(1-p)^j = \frac{1-p}{p}. \tag{A.2.8}$$

Setzt man x gleich der Gesamtzahl der Versuche, so ergibt sich der Erwartungswert

$$\mathbb{E}(X) = \frac{1}{p.} \tag{A.2.9}$$

Es sei nun X die Zeit, die eine Person benötigt, um eine Aufgabe zu lösen, und die Zeit sei exponentialverteilt, d.h. $f(x) = \lambda \exp(-\lambda x)$. Der Erwartungswert ist dann

$$\mathbb{E}(X) = \lambda \int_0^\infty x \exp(-\lambda x) dx = \frac{1}{\lambda}. \tag{A.2.10}$$

Es seien $X_1, \ldots, X_n$ zufällige Veränderliche. Die folgenden Aussagen ergeben sich direkt aus der Definition des Erwartungswerts:

$$\begin{aligned} \mathbb{E}(a) &= a, \ a \text{ eine Konstante} && \text{(A.2.11)} \\ \mathbb{E}(X_1 + \cdots + X_n) &= \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n), && \text{(A.2.12)} \\ \mathbb{E}(aX + b) &= a\mathbb{E}(X) + b, \ a \text{ und } b \text{ Konstanten} && \text{(A.2.13)} \end{aligned}$$

Momente einer Verteilung: Das k-te Moment einer zufälligen Veränderlichen X ist gleich dem Erwartungswert der k-tem Potenz von X:

$$\mathbb{M}_k = \mathbb{E}(X^k). \tag{A.2.14}$$

Die Varianz: Die Varianz einer zufälligen Veränderlichen ist der Erwartungswert der "Abweichungsquadrate" $(X - \mathbb{E}(X))^2$:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \mathbb{E}^2(X), \tag{A.2.15}$$

d.h. $\mathbb{V}(X)$ ergibt sich als Differenz des 2-ten Moments von X und der 2-ten Potenz – also dem Quadrat – des Erwartungswertes.

Die Kovarianz: X und Y seien zwei zufällige Veränderliche mit den Erwartungswerten $\mathbb{E}(X)$ und $E(Y)$. Die Kovarianz $\mathbb{K}(X, Y)$ von X und Y ist der Erwartungswert der Produkte $(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))$,

$$\mathbb{K}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{A.2.16}$$

$\mathbb{K}(X, Y)$ ergibt sich also als Differenz des Erwartungswerts der Produkte XY und dem Produkt der Erwartungswerte $E(X)$ und $\mathbb{E}(Y)$. Sie die Variablen X und Y stochastisch unabhängig, gilt $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ und man erhält $\mathbb{K}(X, Y) = 0$. Für den Spezialfall $X = Y$ erhält man offenbar

$$\mathbb{K}(X, Y) = \mathbb{E}[X - \mathbb{E}(X)]^2 = \mathbb{V}(X). \tag{A.2.17}$$

Lineare Transformationen und Summen: Es sei $Y = aX + b$ eine lineare Transformation von X. Für den Erwartungswert von Y erhält man nach (A.2.13) $\mathbb{E}(Y) = a\mathbb{E}(X) + b$, und für die Varianz von Y ergibt sich

$$\mathbb{V}(Y) = \mathbb{E}[aX + b - a\mathbb{E}(X) - b]^2 = a^2\mathbb{E}[X - \mathbb{E}(X)]^2 = a^2\mathbb{V}(X). \qquad \text{(A.2.18)}$$

Nun sei $Z = X+Y$, wobei X und Y zufällige Variablen mit den Erwartungswerten $\mathbb{E}(X)$ und $\mathbb{E}(Y)$ und den Varianzen $\mathbb{V}(X)$ und $\mathbb{V}(Y)$ sowie der Kovarianz $\mathbb{K}(X, Y)$. Für die Varianz $\mathbb{V}(Z)$ von Z ergibt sich nun

$$\begin{aligned} \mathbb{V}(Z) &= \mathbb{V}(X + Y) = \mathbb{E}[X + Y - \mathbb{E}(X) - \mathbb{E}(Y)]^2 = \\ &= \mathbb{E}[(X - \mathbb{E}(X))^2 + (Y - \mathbb{E}(Y))^2 - 2(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \end{aligned}$$

und nach (A.2.12)

$$\begin{aligned} \mathbb{V}(X + Y) &= \mathbb{E}[(X - \mathbb{E}(X))^2] + \mathbb{E}[(Y - \mathbb{E}(Y))^2] - 2\mathbb{K}(X, Y) \\ &= \mathbb{V}(X) + \mathbb{V}(Y) - 2\mathbb{K}(X, Y). \qquad \text{(A.2.19)} \end{aligned}$$

Offenbar gilt

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) \qquad \text{(A.2.20)}$$

dann und nur dann, wenn X und Y stochastisch unabhängig sind, wenn also $\mathbb{K}(X, Y) = 0$ gilt.

Es sei $U = aX + b$, $V = cY + d$. Die Kovarianz $\mathbb{K}(U, V)$ läßt sich über die Kovarianz $K(X, y)$ von X und Y ausdrücken. Es ist ja

$$\begin{aligned} \mathbb{K}(U, v) &= \mathbb{E}[(aX + b)(cY + d)] - \mathbb{E}(aX + b)\mathbb{E}(cY + d)] \\ &= ac\mathbb{K}(X, Y), \qquad \text{(A.2.21)} \end{aligned}$$

wie man durch Ausmultiplizieren und Berücksichtigung der Sätze über Erwartungswerte leicht verifiziert.

A.3 Äquivalenz der Definitionen von Itemfunktionen

Gegeben seien n dichotome Items mit den Indikatorfunktionen $X_j = \{0, 1\}$. Die folgenden Wahrscheinlichkeiten werden für den Äquivalenznachweis benötigt:

$$\left.\begin{array}{ll} P(X_j = 1), & j = 1, \ldots, n \\ P(X_i = 1, X_J = 1), & i, j = 1, \ldots, n,\ i \neq j \\ P(X_i = 1, X_j = 1, X_k = 1), & i, j, k = 1, \ldots, n,\ \ i \neq j \neq k, \text{etc} \\ \vdots & \\ P(X_1 = 1, X_2 = 1, \ldots, X_n = 1), & \end{array}\right\} \qquad \text{(A.3.1)}$$

Diese Wahrscheinlichkeiten entsprechen den insgesamt 2^n möglichen Antwortmustern (response patterns):

$$\binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = (1 + 1)^n = 2^n.$$

Es wird nun angenommen, dass für jede der binären Variablen X_j eine "unterliegende" (underlying) oder latente Variable ξ_j existiert derart, dass

$$X_j = \begin{cases} 1, & \xi_j \leq \tau_j \\ 0, & \text{sonst} \end{cases} \tag{A.3.2}$$

Die ξ_j können zu einem Vektor $\vec{\xi} = (\xi_1, \ldots, \xi_n)'$ zusammengefasst werden, für den eine faktorenanalytische Repräsentation der Form

$$\vec{\xi} = \vec{\mu} + A\vec{z} + \vec{e} \tag{A.3.3}$$

angesetzt werden kann; A ist eine $n \times r$-Matrix mit Faktorladungen α_{jk}, $j = 1, \ldots, n$, $k = 1, \ldots, r$, d.h. es kann eine latente Struktur mit $r \geq 1$ Dimensionen angenommen werden (dies kann stets angenommen werden und schränkt also die Allgemeinheit der Aussage nicht ein). $\vec{z}$ ist ein Vektor mit Faktorenscores, und $\vec{e}$ ist der übliche Fehlervektor.

Es sei S irgendeine nicht-leere Teilmenge der Items $I_1, \ldots, I_n$. Für gegebenen, fix gehaltenen Vektor $\vec{z}$ gilt nun lokale oder bedingte Unabhängigkeit für die X_j, $j \in S$

$$P\left(\cap_{j\in S} X_j = 1\right) = \prod_{j\in S} P(X_j = 1|\vec{z}).$$

Die unbedingte Wahrscheinlichkeit, dass alle $X_j = 1$ für $j \in S$, erhält man durch Integration über alle $\vec{z}$, d.h.

$$P\left(\cap_{j\in S} X_j = 1\right) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j\in S} P(X_j = 1|\vec{z}) h(\vec{z}) d\vec{z} \tag{A.3.4}$$

Hierin ist $h(\vec{z})$ die -apriori-Verteilung der Komponenten von $\vec{z}$, und $P(X_j = 1|\vec{z})$ ist die Itemfunktion.

Nun gilt

$$\begin{aligned} P\left(\cap_{j\in S} X_j = 1\right) &= P\left(\cap_{j\in S} \xi_j \leq \tau_j|\vec{z}\right) && \text{(A.3.5)} \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P\left(\cap_{j\in S} \xi \leq \tau_j|\vec{z}\right) h(\vec{z}) d\vec{z} && \text{(A.3.6)} \\ &= \int_{-\infty}^{\infty} \prod_{j\in S} P(\xi_j \leq \tau_j|\vec{z}) h(\vec{z}) d\vec{z}. && \text{(A.3.7)} \end{aligned}$$

Die komponenten von ξ sind, für fixen Vektor $\vec{z}$, unabhängig, weil ja die Fehler in $\vec{e}$ unabhängig sind. Wegen (A.3.3) gilt weiter

$$\begin{aligned} P(\xi_j \leq \tau_j|z) &= P(\mu_j + e_j + \sum_{k=1}^{r} \alpha_{jk} z_j \leq \tau_j|\vec{z}) \\ &= P\left[\frac{e_j}{\sigma_j} \leq \frac{\tau_j - \mu_j - \sum_{k=1}^{r} \alpha_{jk} z_j}{\sigma_j} \middle| \vec{z}\right] \\ &= R\left[\frac{\tau_j - \mu_j - \sum_{k=1}^{r} \alpha_{jk} z_j}{\sigma_j}\right]. && \text{(A.3.8)} \end{aligned}$$

Hierin ist $R[\cdot]$ die Verteilungsfunktion von e_j/σ_j. Für die Itemfunktion (response model) gilt andererseits

$$P(X_j = 1|\vec{z}) = G\left(\lambda_{j0} + \sum_{k=1}^{r} \lambda_{jk}\right). \tag{A.3.9}$$

Die Modelle sind äquivalent wenn (i) $G \equiv R$ und (ii)

$$\lambda_{j0} = \frac{\tau_j - \mu_j}{\sigma_j}, \text{ und } \lambda_{jk} = \frac{\alpha_{jk}}{\sigma_j}, j = 1, \ldots, n$$

Für jedes Response- (Itemfunktion) Modell läßt sich ein UV- (underlaying variable) Modell finden und umgekehrt. Insbesondere hat man

$$G(u) = \frac{1}{1 + e^{-u}} = R(e_j/\sigma_j). \tag{A.3.10}$$

Links steht die Verteilungsfunktion für eine logistisch verteilte Variable mit Erwartungwert 0 und Varianz $\pi^2/3$, und deshalb ist auf der rechten Seite e_j logistisch verteilt mit Erwartungswert 0 und Varianze $\sigma_j^2\pi^2/3$.

Implikation: Aus der Äquivalenz folgt, dass die Parameter τ_j, μ_j und α_{jk} nicht individuell schätzbar sind. In der Likelihoodfunktion sind nur die Parameter

$$\beta_j = \frac{\tau_j - \mu_j}{\sigma_j}$$

schätzbar. □

A.4 Cronbachs α: Beweis von Satz 3.2.1

Es ist

$$\rho_{xx'} = \frac{\mathbb{V}(\tau)}{\mathbb{V}(X)}, \quad \tau = \sum_j \tau_j \tag{A.4.1}$$

Zut Vereinfachung der Notation werde $\sigma_j^2 = \mathbb{V}(\tau_j)$, $\sigma_{ij} = \mathbb{K}(\tau_i, \tau_j)$ gesetzt. Es gilt

$$\mathbb{V}(\tau) = \sum_j \sigma_j^2 + \sum_{i \neq j} \sigma_{ij} \tag{A.4.2}$$

$$\mathbb{V}(X) = \sum_j \mathbb{V}(Y_j) + \sum_{i \neq j} \sigma_{ij} \tag{A.4.3}$$

wobei von der Identität (**??**) (Seite **??**) Gebrauch gemacht wurde, die hier die Form $\mathbb{K}(Y_i, Y_j) = \mathbb{K}(\tau_i, \tau_j) = \sigma_{ij}$ annimmt. Nun gilt allgemein die Ungleichung[67]

$$2\sigma_{ij} \leq \sigma_i^2 + \sigma_j^2, \quad i \neq j. \tag{A.4.4}$$

[67]Für beliebige Zahlen x_i und y_i gilt $\sum_i (x_i - y_i)^2 = \sum_i x_i^2 + \sum_i y_i^2 - 2\sum_i x_i y_i \geq 0$, so dass $\sum_i x_i^2 + \sum_i y_i^2 \geq 2\sum_i x_i y_i$. Haben die x_i und y_i den Mittelwert Null, so entspricht diese Ungleichung der Ungleichung $\sigma_i^2 + \sigma_j^2 \geq 2\sigma_{ij}$, σ_{ij} die Kovarianz der x_i und y_j.

Dann gilt auch

$$2\sum_{i\neq j}\sigma_{ij} \leq \sum_{i\neq j}(\sigma_i^2+\sigma_j^2). \tag{A.4.5}$$

Es ist

$$\begin{aligned}\sum_i\sum_j(\sigma_i^2+\sigma_j^2) &= \sum_{i=j}(\sigma_i^2+\sigma_j^2)+\sum_{i\neq j}(\sigma_i^2+\sigma_j^2), \text{ d.h.}\\ 2n\sum_i\sigma_i^2 &= 2\sum_i\sigma_i^2+\sum_{i\neq j}(\sigma_i^2+\sigma_j^2),\end{aligned} \tag{A.4.6}$$

denn $\sum_i\sum_j(\sigma_i^2+\sigma_j^2) = n\sum_i\sigma_i^2 + n\sum_j\sigma_j^2 = 2n\sum_i\sigma_i^2$, weil ja $\sum_i\sigma_i^2 = \sum_j\sigma_j^2$, und $\sum_{i=j}(\sigma_i^2+\sigma_j^2) = \sum_i(\sigma_i^2+\sigma_i^2) = 2\sum_i\sigma_i^2$. Aus (A.4.6) und (A.4.5) folgt dann weiter

$$2(n-1)\sum_j\sigma_j^2 = \sum_{i\neq j}(\sigma_i^2+\sigma_j^2) \geq 2\sum_{i\neq j}\sigma_{ij}, \tag{A.4.7}$$

d.h. aber

$$\sum_j\sigma_j^2 \geq \frac{1}{n-1}\sum_{i\neq j}\sigma_{ij} \tag{A.4.8}$$

Nach (A.4.2) gilt

$$\mathbb{V}(\tau) = \sum_j\sigma_j^2 + \sum_{i\neq j}\sigma_{ij} \geq \frac{1}{n-1}\sum_{i\neq j}\sigma_{ij} + \sum_{i\neq j}\sigma_{ij} = \frac{n}{n-1}\sum_{i\neq j}\sigma_{ij}.$$

Aus (A.4.3) erhält man $\mathbb{V}(X) - \sum_j\mathbb{V}(Y_j) = \sum_{i\neq j}\sigma_{ij}$, so dass

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

und dies ist (3.2.46). □

B Theorie des Messens

B.1 Grundbegriffe der Messtheorie

v. Helmhotz bemerkte, dass elementare physikalische Größen wie Masse und Länge die gleiche Struktur haben wie die positiven reellen Zahlen ($\mathbb{R}_+$), zusammen mit der Ordnungsrelation $\geq$ und der Addition + (Narens & Luce, 1986). Misst man etwa die Länge von Objekten und ist die Länge des Objekts b größer als die des Objekts a, so kann man diesen Sachverhalt durch $a \prec b$ ausdrücken und Maßzahlen $x = \phi(a)$, $y = \phi(b)$ schreiben, wobei $x, y \in \mathbb{R}^+$ und $x < y$ gilt. s ist eine Regel, nach der reele Zahlen Objekten zugeordnet werden derart, dass die Relation $\prec b$ durch $x < y$ widergespiegelt wird. Sind die Objekte gleich lang, so kann man $a \sim b$ schreiben, und $x = \phi(a) = \phi(b) = y$. Analoge Betrachtungen

gelten für die Masse der Objekte. Man kann auch zwei Objekte aneinander legen und die Länge bzw. Masse dieser Kombination dieser Objekte bestimmen; man spricht dann von einer *Verkettung* (concatentation) $a \oplus b$, der der Messwert $x + y$ entspricht. Ist M die Menge der zu messenden Objekte und wird ein bestimmtes Merkmal betrachtet, das gemessen werden soll, so kann man diese Menge durch $\mathbb{M} = \{M, \preceq, \oplus\}$ repräsentieren; $\mathbb{M}$ ist ein *empirisches Relativ* oder eine *qualitative Struktur*, der eine bestimmte numerische Struktur $X = \{\mathbb{R}^*, \geq, +\}$ entspricht. Die Idee des Messens ist nun, dass der Struktur der Objekte (bezüglich des gemessenen Merkmals) eine bestimmte Struktur in den reellen Zahlen entsprechen soll. Dazu muß die Funktion ϕ geeignet definiert werden. Der Verkettung $\oplus$ soll etwa die Operation der Addition, $+$, entsprechen, und es soll gelten

$$\phi(a \oplus b) = \phi(a) + \phi(b). \tag{B.1.1}$$

Wenn $\mathbb{M}$ die Objektmenge ist und $\mathbb{R}^+$ die "Bild"menge (die Zahlen $\phi(a)$ sind "Bilder" der Objekte $a \in \mathbb{M}$), so kann man sagen, dass die durch $\oplus$ definierte Struktur in $\mathbb{M}$ auf die Bildmenge übertragen wird; man spricht von einem *Homomorphismus*. Allgemein sind Homomorphismen strukturerhaltende Abbildungen von Mengen in Mengen. Bei Längen und Massen hat man sicher die Bedingungen (i) $a \preceq b$ genau dann, wenn $\phi(a) \geq \phi(b)$, und (ii) wenn (B.1.1) gilt. Die Menge der Homomorphismen für die Merkmale Masse und Länge ist nicht eindeutig; hat man zwei Abbildungen ϕ_1 und ϕ_2, so kann man allerdings zeigen, dass sich ϕ_1 und ϕ_2 nur durch einen Faktor unterscheiden, – man ist frei, eine Einheit zu wählen. Die Abbildung ϕ definiert eine Skala, und Skalen mit verschiedenen Einheiten können durch Wahl eines geeigneten Faktors aufeinander bezogen werden.

Extensive Quantitäten: Ein zentraler Begriff in der Messtheorie, der erläutert werden sollte. Der Mathematiker Hölder[68] führte ein Axiom ein, mit dem bereits Archimedes den Begriff des Koninuums klarer definieren wollte, wie es heißt als eine Reaktion auf die Paradoxien, die Zenon aufgestellt hatte[69]. Es sei $\mathbb{M}$ eine Menge von Stäben, die der Länge nach geordnet werden können, $a \preceq b$ mit $a, b \in \mathbb{M}$ soll heißen, dass der Stab b mindestens so groß wie a ist. Die Verkettung von a und b, also $a \oplus b$, soll bedeuten, dass man a und b so aneinander legt, dass sie einen neuen Stab bilden. Für a finde man nun einen Stab a_1 mit $a \sim a_1$, wobei $\sim$ bedeutet, dass a und a_1 gleichlang sind. Dann soll ein weiterer Stab a_2 gefunden werden mit $a \oplus a_1 \sim a_2$, etc. Es entsteht eine *Standardfolge* $a_1, a_2, \ldots, a_n$ *auf der Basis* a. Der Punkt dabei ist, dass jede begrenzte Teilfolge einer Standardfolge endlich ist. Hölder nahm nun an, dass $\mathcal{M} =< \mathbb{M}, \preceq, \oplus >$ durch fünf Eigenschaften charakterisiert wird.

1. **Schwache Ordnung:** Die Relation $\preceq$ ist transitiv, d.h. aus $a \preceq b$ und $b \preceq c$

[68] Otto Ludwig Hölder (1859 - 1937)

[69] Nach Zenon führt die Idee, eine zetiliche oer räumliche Ausdehnung aufteilen zu wollen, in die Irre. Bekannt ist die Paradoxie, derzufolge Achilles eine Schildkröte nicht einholen kann, oder ein Pfeil nicht von A nach B fliegen kann.

folgt $a \preceq c$, für $a, b, c \in \mathbb{M}$. Weiter ist $\preceq$ verbunden, d.h. es gilt entweder $a \preceq b$ oder $b \preceq a$.

2. **Monotonie:** Die Verkettung von Objekten ist ordnungsbewahrend, d.h. für alle $a, b, c, d \in \mathbb{M}$ gilt, wenn $a \preceq c$ und $b \preceq d$, dann auch $a \oplus c \preceq b \oplus d$.

3. **Lösbarkeit:** Für alle $a, b \in \mathbb{M}$ und $a \prec b$ existiert ein $c \in \mathbb{M}$ derart, dass $a \prec b \oplus c$.

   Mit diesem Axiom wird – zusammen mit den anderen Axiomen – angenommen, dass es beliebig kleine Ausdehnungen gibt. Denn $a \prec b$ heißt zwar, dass sich a und b unterscheiden /a ist größer als b), aber sie können sich auch nur beliebig wenig unterscheiden. Dies heißt wiederum, dass man noch eine beliebig kleine Größe zu b addieren kann und a ummer noch größer ist als die Verkettung von b und c.

4. **Positivität:** Verkettet man Objekte, so ergibt sich stets ein Objekt, das größer ist als jedes der Objekte allein. Also für $a, b \in \mathbb{M}$ gilt sowohl $a \oplus b \prec a$ als auch $a \oplus b \prec b$.

5. **Assoziativität:** Verkettet man drei oder mehr Objekte, so kommt es nicht auf die Ordnung, in der verkettet wird, an. Für alle $a, b, c \in \mathbb{M}$ soll demnach $a \oplus (b \oplus c) = (a \oplus b) \oplus c$ gelten.

Eine (Archimedische) Struktur $\mathcal{M} =< \mathbb{M}, \preceq, \oplus >$, die den Axiomen 1 – 5 genügt, heißt *extensiv* und kann in die Menge $\mathbb{R}^+$ der reellen Zahlen abgebildet werden, mit "+" als Operation, die der Verkettung entspricht. Die Axiome 1, 2, 4 und 5 sind notwendig, damit die Repräsentation in der Menge der reellen Zahlen additiv ist. Das Axiom 3 heißt *strukturell*, weil es den Fokus auf Strukturen mit einer additiven Eigenschaft begrenzt.

Fundamentale Messung und Skalen: Campbell (1920) argumentierte, dass elementare Messungen notwendig extensive Messungen seien; die korrespondierenden Skalen nannte er 'fundamental', im Unterschied zu den 'abgeleiteten' (derived) Messungen, die solche Skalen voraussetzen. Masse, Länge, Temperatur etc seien fundamentale Messungen, während etwa die Messung eines Impulses p, definiert als Produkt $m\,v$ der Masse m und Geschwindigkeit v eines Körpers, eine abgeleitete Messung ist, weil der Impuls eben aus den fundamentalen Größen m und v zusammengesetzt ist.

Aus Diskussionen über die Möglichkeit fundamentaler Messungen in der Psychologie ergab sich der Begriff des Skalentyps, wie er von Stevens (1946) und (1951) eingeführt wurde. Skalen werden demnach über die Menge von zulässigen Transformationen definiert. Bekanntlich kann man eine Temperatur in Celsius oder in Fahrenheit messen; die Maßzahlen sind dann zwar verschieden, bedeuten aber dasselbe und können durch eine lineare (besser: affine) Transformation der Form $y = bx + a$ ineinander überführt werden. Man kann die Länge einer

Strecke in Metern, Kilometern, Zentimetern und Milllimetern messen; die Maßzahlen bedeuten wieder dasselbe und können durch Multiplikation mit einem geeignet gewählten Faktor ineinander überführt werden. Man erhält etwa die folgenden Skalentypen, definiert durch die korrespondierenden Transformationen: im Allgemeinen werden nur die Typen Nominal-, Ordinal-, Intervall-, Verhältnis-

Tabelle 20: Skalentypen; $\mathbb{R}$ Menge der reellen Zahlen, $\mathbb{R}^+$ Menge der positiven reellen Zahlen, $\mathbb{N}$ Menge der natürlichen Zahlen, $\mathbb{N}^+$ Menge der positiven natürlichen Zahlen

Skala	Transformation	Spezifikation
Nominal	$x \to f(x)$	$x \in \mathbb{R}$, f beliebig, aber 1-deutig
Ordinal	$x \to f(x)$	f streng monoton
Log-Intervall	$x \to sx^p$	$x, s, p \in \mathbb{R}^+$
Intervall	$x \to bx + a$	$x, a, b \in \mathbb{R}$
Diskret-Intervall	$x \to k^n x + s$	$k > 0$, $n \in \mathbb{N}$, $x, s \in \mathbb{R}$
Verhältnis (Ratio)	$x \to a\,x$	$x, a \in \mathbb{R}$
Diskret-Verhältnis	$x \to k^n$	$k > 0$ fix, $n \in \mathbb{N}^+$, $x \in \mathbb{R}$
Absolut	$x \to x$	$x \in \mathbb{R}$

und Absolutskala betrachtet, die übrigen Typen ergeben sich aus grundsätzlichen Betrachtungen, auf die hier nicht im Detail eingegangen werden muß (vergl. Narens & Luce (1986), Luce & Narens (1987)). Stevens hat keine Begründung gegeben, warum nur diese Skalentypen betrachteter werden sollen, und in der Tat sind weitere Skalentypen inzwischen diskutiert worden. In der Physik gilt, dass dimensional invariant bei einer Veränderung der Maßeinheiten sein sollen; diese Annahme liegt auch dem Tevenschen Skalenbegriff zugrunde: die Spezifikation zulässiger Transformationen soll klarstellen, was eine Messung als sinnvoll auszeichnet. Wie Luce & Narens (1986) zu Recht herausstellen, ist die Frage, was denn allgemein eine *sinnvolle* wissenschaftliche Aussage charakterisiert, keineswegs eindeutig beantwortet (vergl. hierzu auch Cartwright (1983)).

B.2 Verbundmessung (Conjoint Measurement, CM)

Fundamental für fundamentale Messungen ist der Begriff der Verkettung (concatenation), der in der Physik zunächst problemlos erscheint (s. die Beispiele für Länge, Gewicht, etc); in der Tat ist der Begriff der fundamentalen Messung für Messungen im Bereich der Physik entwickelt worden. Für psychologische Variable wie Intelligenz, Lautheit, Nützlichkeit (utility → Entscheidungstheorie) etc ist dagegen nicht klar, wie der Begriff der Verkettung eindeutig eingeführt werden soll. Nun ist es so, dass viele Größen von bestimmten Komponenten abhängen. In der Physik etwa hängt der Begriff der Kraft F von denen der Masse m und der Besdchleunigung b eines Körpers ab, der Begriff des Impulses p ebenfalls von

dem der Masse sowie von dem der Geschwindigkeit v. Die Frage ist, in welcher Form F von m und b abhängt, und wie p von m und v abhängt. Seit Newton gilt $F = mb$ und $p = mv$, aber ob man gleichermaßen behaupten kann, die Intelligenz I sei ebenfalls von der räumlichen Vorstellungskraft ρ und der rechnerischen Fähigkeit ν in der Form $I = \rho\nu$ bestimmt, ist keineswegs klar und muß empirisch entschieden werden. Man kann so vorgehen, dass man die Veränderungen in einer der Variablen betrachtet, etwa ρ, wenn ν verändert wird derart, dass I konstant bleibt. Man erhält dann *Indifferenzkurven*, aus denen eventuell Information über das Zusammenspiel von ρ und ν abgeleitet werden kann.

Hier werden also die Ausprägungen von Merkmalen zusammen mit den Ausprägungen anderer Variablen betrachtet. Man spricht von *Verbundmessung* oder *conjoint measurement*. Die Betrachtung ist in gewisser Weise ganzheitlich (holistisch), und so ist auch der Ausdruck *conjoint* entstanden: er ist eine Zusammenziehung von "*con*sidered *joint*ly". Im Prinzip ist das Vorgehen bei dieser Art von Messung so, dass Objekte hinsichtlich einer bestimmten Menge $\mathbb{M}$ von Merkmalen beurteilt und nach Maßgabe eines bestimmten Merkmales z.B. in eine Rangordnung gebracht werden. Diese Ordnung hängt davon ab, in welcher Weise die verschiedenen Merkmale die Elemente in $\mathbb{M}$ definieren. Das Ziel ist, diese Merkmalskomposition zu identifizieren und dabei die Merkmale auf Skalen zu repräsentieren. So kann man physikalische Gegenstände nach Maßgabe ihrer Masse und ihres Volumens rangordnen und daraus Skalen für die Masse einerseits und das Volumen andererseits gewinnen, oder man für Paare von Lautstärken von Sinustönen eine Rangordnung erstellen lassen, wobei eine Lautstärke von der Amplitude einerseits und der Frequenz des Tones andererseits bestimmt wird. Aus der Rangordnung lassen sich dann Skalen für die psychische Repräsentation der Amplituden und der Frequenzen gewinnen. Wichtig ist dabei wieder der *trade-off* zwischen den Merkmalen einer Komposition von Merkmalen, der die Indifferenzkurve bestimmt. Man kann überprüfen, bei welchen Variationen der Ausprägungen der Merkmale in der Komposition der Ort in der Rangordnung oder das Ausmaß an Präferenz konstant bleibt. Bei Anwendungen des Verfahrens etwa in der Marktforschung wird die praktische Bedeutung dieses Sachverhalts deutlich: man kann den Preis einer Ware (Auto, Fleckenentfernungsmittel, Handy, etc) vorgeben und dann bestimmte Ausstattungsmerkmale so variieren, dass der Preis konstant bleibt. Auf diese Weise läßt sich dann bestimmen, mit welcher Kombination von Merkmalen ein Handy für gegebenen Preis maximale Attraktivität für prospektive Käufer hat. In einem ganz anderen Zusammenhang hat man etwa die Aufgabe, die Elemente einer medizinischen Therapie so zu kombinieren, dass die Kosten bei konstanter Wirksamkeit minimal werden. Im einfachsten Falle wirken die Merkmale dann additiv zusammen, und man spricht von *additiver Verbundmessung*.

Tversky (1967) nennt als Beispiel das Merkmal *performance* p, das nach der Hullschen Theorie aus den drei Komponenten a = *learning*, b = *incentive* und c = *drive* multiplikativ zusammengesetzt ist. Man betrachte den Fall, dass a, b, c

und p unabhängig voneinander bestimmt werden können. Die Frage ist dann, wie a, b und c zusammenwirken, um p zu bestimmen. Genügen die Messungen dieser Variablen bestimmten Kriterien, könnte die Varianzanalyse Informationen über die Art der Kombination liefern. Allerdings sind oft keine unabhängigen Messungen verfügbar; statt dessen liegen nur Rangordnungen bezüglich der gemeinsamen Effekte der a, b und c auf p vor. Das Messproblem sowie die Frage nach der Kombination sollen dann möglichst simultan gelöst bzw. beantwortet werden, wobei letztere die nach der Überprüfung einer Hypothese über die Kombination sein kann.

Gegeben sei etwa ein $A \times B \times C$ faktorielles Design, und (a, b, c) bezeichne eine bestimmte Zelle dieses Designs, mit p als entsprechender abhängiger Variable. Ist die Regel *additiv*, so werden Skalenwerte $u(a)$, $v(b)$ und $w(c)$ gesucht derart, dass $p = u(a) + v(b) + w(c)$. Die additive (Kompositions-) Regel ist sicherlich ein Spezialfall; allgemeiner sind Polynome, etwa $u(a)(v(b) + w(c)$, oder $u(a)v(b)w(c)$, etc. Die additive Regel ist ein Spezialfall eines Polynoms. Allgemein habe man die "Faktoren" – also unabhängige Variablen – $A, B, \ldots, K$, und $(a, b, \ldots, k)$ sei eine bestimmte Kombination dieser Stufen, durch die eine abhängige Variable p bestimmt wird. Die $(a, b, \ldots, k)$ (bezüglich p) können ranggeordnet werden. Die Menge der $(a, b, \ldots, k)$ heißen *Datenstruktur*. Eine bestimmte Kombination $(a, b, \ldots, k)$ heißt *Datenelement*. Eine Kompositionsregel, derzufolge die unabhängigen Variablen durch ein Polynom definiert sind, definiert ein *polynomiales Messmodell*. Allgemein läßt sich ein solches Modell wie folgt charakterisieren (Tversky (1967), p. 2):

Definition B.2.1 *Eine Datenstruktur D genügt einem* polynomialen Messmodell *M, wenn eine auf D definierte reellwertige Funktion f existiert und reellwertige Funktionen f_A, $f_B, \ldots, f_K$ auf den Faktoren A, $B, \ldots, K$ existieren derart, dass für jedes Datenelement $(a, b, \ldots, k)$*

1. *$f(a, b, \ldots, k) = M(f_A(a), f_B(b), \ldots, f_K(k))$ gilt, wobei M ein Polynom (polynomiale Funktion) ist, das durch Summen, Differenzen und Produkten der $f_A, \ldots, f_K$ definiert ist,*
2. *für alle $x = (a, b, \ldots, k)$, $x' = (a', b', \ldots, k')$ gilt:*

$$\begin{aligned} x >_o x' \quad &\textit{impliziert} \quad f(x) > f(x') \\ x =_o x' \quad &\textit{impliziert} \quad f(x) = f(x'). \end{aligned} \tag{B.2.1}$$

Dabei bezeichnen $>_o$ und $=_o$ die Ordnung, die in den Daten beobachtet wird.

Anmerkungen: $f(a, b, \ldots, k)$ entspricht der im Beispiel genannten Performance p. M hat im additiven Fall die Form $f(a, b, \ldots, k) = f_A(a) + f_B(b) + \cdots + f_K(k)$. In alternativen Formen können z.B. Produkte der $f_A(a), f_B(b), \ldots$ auftreten. Die f_A, f_B, etc sind zu bestimmende Skalenwerte. Die Bedingung 2 für x und x' drückt

aus, dass f die Ordung zwischen den Datenelementen $(a, b, \ldots, k)$ widerspiegeln muß. □

Für $x = f(a, b, \ldots .k)$ kann ein numerischer Wert angegeben werden. Hat x nur eine ordinale Bedeutung, so ist D ordinal. Hat x eine absolute Bedeutung, – etwa als Preis für eine Ware, oder Ausprägung eines Merkmals – so heißt die Datenstruktur *numerisch*. Eine numerische Datenstruktur D_g ist eine Datenstruktur D zusammen mit einer reellwertigen Funktion g, die jedem Datenelement $(a, b, \ldots, k)$ einen eindeutigen, reellen Zahlenwert zuordnet. g ersetzt die Funktion f. Entsprechen die Daten einem numerischen Modell, so entsprechen sie auch einem ordinalen Modell. Die Umkehrung gilt nicht: ein ordinales Modell entspricht nicht notwendig auch einem numerischen Modell.

Sich ein Messmodell auszudenken, ist eine Sache, es zu überprüfen eine andere. Die Überprüfbarkeit des Modells hängt davon ab, ob das zu dem Modell korrespondierende System polynomialer Gleichungen und Ungleichungen lösbar ist.

Systeme, die durch ordinale Datenstrukturen erzeugt werden, heißen *homogen*, solche, die durch numerische Datenstrukturen erzeugt werden, heißen *inhomogen*. Inhomogene Datenstrukturen enthalten numerische Konstante. Ordinale Datenstrukturen repräsentieren *fundamentale Messungen*, während numerische Datenstrukturen *abgeleitete Messungen* sind. Bei fundamentalen Messungen werden Zahlen nur durch das Messmodell eingeführt, während bei abgeleiteten Messungen numerische Messwerte auf der Basis schon vorher zugeordneter Zahlen bestimmt werden.

Beispiel B.2.1 Subjektiver erwarteter Nutzen: (Savage, 1954) Gegeben sei ein Spiel x mit möglichen Ergebnissen (outcomes) $o_1, \ldots, o_n$. $u(o_i)$ sei der Nutzen von o_i. Die Ergebnisse hängen von zufälligen Ereignissen $e_1, \ldots, e_n$ ab; das Ereignis e_i impliziert also das Ergebnis o_i. x' sei ein Spiel mit den Ergebnissen $o'_1, \ldots, o'_n$, die von den zufälligen Ereignissen $e'_1, \ldots, e'_n$ abhängen, und $u(o'_i)$ sei wieder der Nutzen von o'_i. $\phi(e_i)$ und $\phi(e'_i)$ seien die Wahrscheinlichkeiten von e_i bzw. e'_i. Das Modell des erwarteten subjektiven Nutzens besagt, dass

$$x >_o x' \text{ genau dann, wenn } \sum_{i=1}^{n} u(o_i)\phi(e_i) > \sum_{i=1}^{n} u(o'_i)\phi(e'_i), \tag{B.2.2}$$

wobei $\sum_{i=1}^{n} u(o_i)\phi(e_i)$ und $\sum_{i=1}^{n} u(o'_i)\phi(e'_i)$ die erwarteten Nutzen der beiden Spiele sind.

Beispiel B.2.2 Hulls und Spencers Performance-Modelle: Es sei p ein Maß für Performanz, das in einem $H \times D \times K$ untersucht wird, – H repräsentiert Lernen, D "Drive" (Antrieb), K Motivation. Es sei $x = (h, d, k)$ und $x' = (h', d', k')$ ein beliebiges Paar von Kombinationen von H, D und K. Das Hullsche Modell postuliert

$$x >_0 x' \text{ genau dann, wenn } f_H(h)f_D(d)F_k(k) > f_H(h')f_D(d')f_K(k'). \tag{B.2.3}$$

Darin sind f_H, f_D und f_K Skalen für Lernen, Antrieb und Motivation. Das Messmoell erscheint als multiplikativ. Da aber die Ordnung $>_o$ monoton mit dem Logarithmus ist und

$$\log[f_H(h)f_D(d)F_k(k)] = \log f_H(h) + \log f_D(d) + \log f_K(k),$$

(analog für $f_H(h')f_D(d')f_K(k')$), ist das Modell äquivalent einem additiven Messmodell, vorausgesetzt die f_H etc -Werte sind alle positiv.

Das Modell von Spence postuliert dagegen

$$x >_o x' \text{ genau dann, wenn } f_H(h)(f_D(d) + f_K(k)) > f_H(h')(f_D(d') + f_K(k')). \tag{B.2.4}$$

Um die Modelle (B.2.3) und (B.2.4) miteinander zu vergleichen, müssen alle Variablen H, H un dK simultan variiert werden, – eine ausführliche Diskussion dieser beiden Modelle findet man in Krantz & Tversky (1966). □

Beispiel B.2.3 Bradley-Terry-Luce-Modell: Es seien x und y zwei "Objekte" und es soll zwischen beiden nach Maßgabe der Präferenz gewählt werden. Es wird angenommen, dass eine Nutzenfunktion v existiert; dann wird postuliert, dass die Wahrscheinlichkeit der Wahl von x gemäß

$$p(x, y) = \frac{v(x)}{v(x) + v(y)} \tag{B.2.5}$$

gegeben ist. Dann ist $p(y, x) = v(y)/(v(x) + v(y))$, so dass

$$\frac{p(x, y)}{p(Y, x)} = \frac{v(x)}{v(y)} \tag{B.2.6}$$

unter der Voraussetzung $p(x, y) \neq 0$, $p(y, x) \neq 0$. Dies ist das *Bradley-Terry-Luce* (BTL-Modell). Das BTL-Modell kann als ploynomiales Messmodell angeschrieben werden, denn aus (B.2.6) folgt

$$\log p(x, y) - \log p(y, x) = v'(x) - v'(y), \quad v' = \log v, \tag{B.2.7}$$

also kann man

$$p(x, y) > p(w, z) \text{ genau dann, wenn } u(x) - u(y) > u(w) - u(z) \tag{B.2.8}$$

schreiben. Da links eine Relation zwischen den Zahlen $p(x, y)$ und $p(w, z)$ steht, hat man es hier mit einem numerischen Messmodell zu tun. □

Die ausführliche messtheoretische Diskussion des polynomialen CM-Modells wird in Tversky (1967) gegeben, worauf hier nicht weiter eingegangen werden kann.

Literatur

[1] Adorno, T.W., Frenkel-Brunswik, E., Levinson, D. J., Sanford, R.N.: The Authoritarian Personality. Harper und Brothers, New York 1950.

[2] Agresti, A. (1988) A model for agreement between ratings on an ordinal scale. *Biometrics*, 44, 539 –548

[3] Agresti, A. (1992) Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research*, 1, 201 – 218

[4] Andrich, D. (1978) Relationship between the Thurstone and the Rasch approaches to item scaling. *Applied psychological Measurement* 2, 449-460

[5] Andersen, S. (1997) Re: Bayesian estimation of desease prevalence and the parameters of diagnostic tests in the absence of a gild standard. *American Journal of Epidemiology*, 145, 250 – 291

[6] Andrich, D. (2004) Controversy and the Rasch model. A characteristic of incompatible paradigms? *Medical Care*, 42, 1, 1 - 7

[7] Arkes, H. R., Wortmann, R. L., Saville, P. D., Harkness, A. R. (1981) Hindsight bias among physicians weighing the likelihood of diagnoses. *Journal of Applied Psychology*, 66(2), 252-254

[8] Ayres, I.: Super crunchers. London 2007

[9] Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D. (1999) Beyond Kappa: a review of interrater agreement measures. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 27 (1), 3 – 23

[10] Bartholomew, D.J.: Latent Variable Models and Factor Analysis. London 1987

[11] Bartko, J.J. (1966) The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 19, 3-11

[12] Bartko, J.J. (1976) On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 83, 762-765

[13] Bartolucci, F. (2007) A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, 72, 2, 141-157

[14] Bar-Hillel, M. (1980) The base-rate fallacy in probability judgments. *Acta Psychologica*, 44, 211-233

[15] Basilevsky, A.: Statistical factor analysis and related methods. Theory and applications. John Wiley & Sons, New York, 1994

[16] Beiser, A.: Concepts of modern physics. Auckland etc, 1984

[17] Bennet, E.M., Alpert, R., Goldstein, A.C. (1954) Communications through limited response questioning. *The Public Opinion Quarterly*, 18(3), 303 – 308

[18] Blinkhorn, S. (1994) Willow, titwillow, titwillow!, *Nature*, 372, 417–419

[19] Boldt, R.F. (1966) Study of linearity and homoscedasticity of test scores in the cance range. *Research Bulletin* 66-43. Princeton, N.J.: Educational Testing Service.

[20] Bock, R. D., Wood, R. (1971) Test theory. *Annual Revies of PSychology*, 22, 193 – 224

[21] Bollen, K.A. (2002) Latent variables in psychology and sociology. *Annual Review of Psychology*, 53, 605 – 634

[22] Borkenau, P, Ostendorf, F. (192) The big five as states: how useful is the five factor moel to descriobe intraindividual variations over time? *Journal of Research in Personality*,32, 202–221

[23] Borsboom, D., Mellenbergh, G. J., and van Heerden. J. (2003) The Theoretical Status of Latent Variables. *Psychometrika*, 110. 203–219

[24] Burt, C. (1950) The factorial analysis of qualitative data. *British Journal of Psychology, (Statistical Section)* 3, 166–185

[25] Busemeyer, J.R., Jones, L. E. (1983) Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, 93 (3), 549 - 562

[26] Byrt, T., Bishop, J., Carlin, B. (1993) Bias, prevalence, and Kappa. *Journal of Clinical Epidemiology*, 46 (5), 423 – 429

[27] Campbell, N. R. Physics: the elements. Cambridge, 1920. (Reprinted [1957] as Foundations of Science, New M>rk: Dover)

[28] Cervone, D. (1977) Social-cognitive mechanisms and personality coherence: Self-knowledge, situational beliefs, and cross-situational coherence in perceived self-efficacy. *Psychological Science*, 8, 43–50

[29] Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 378 – 382

[30] Coughlin, S.S., Pickle, L. W., Goodman, M.T., Wilkens, L.R. (1992) The logistic modeling of interobserver agreement. *Journal of Clinical Epidemiology*, 45, 1237–1241

[31] Christoffersson, A. (1975) Factor analyis of dichotomized variables. *Psychometrika*, 40, 5-32

[32] Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, n: The dependability of Behavioural Measurements: theory of Generalizability for Scores and Profiles. New York 1972

[33] Cohen, J. (1960) A coefficient of agreement fornominal scales.*Educational and Psychological Measurement*, 20, 1960, 37-46.

[34] Cronbach, L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–234

[35] Courant, R.: Vorlesungen über Differential- und Integralrechnung. Erster Band – Funktionen einer Veränderlichen. Springer-Verlag, Berlin, Göttignen, Heidelberg 1955

[36] Cronbach, L.J., Meehl, P.E. (1955) Construc validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

[37] Dawes, R., Faust, D., Meehl, P.E. (1989) Clinical versus Actuarial Judgment. *Science* 243, 1668 – 1673

[38] Dawes, R.M.: House of cards. Psychology and psychotherapy built on myth. The Free Press, New York 1994

[39] Dawes, R.M.: Everyday Irrationality. How pseudo-scientists, lunatics, and the rest of us fail to think rationally. Westview Press 2001

[40] Devlin, B., Fienberg, S.E., Resnick, D. P., Roeder, K.: Intelligence, Genes, Success. Scientists respond to the *Bell Curve*. New York 1977

[41] Dilthey, W.: Ideen über eine beschreibende und zergliedernde Psychologie (1904), in: Dilthey, W.: Die Philosophie des Lebens. Aus seinen Schriften ausgewählt von Herman Nohl, Göttingen 1961

[42] Drösler, J. (1986) Figurale Wahrnehmung doch additiv? *Zeitschrift für experimentelle und angewandte Psychologie*, 33, 351–359

[43] Egan, J.P.: Signl Detection Theor and ROC Analysis. New York, 1975

[44] Ellis, J.L., van den Wollenberg, A.L. (1993) Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. *Psychometrika*, 58, 417–429

[45] Embretson (Whitley), S.: (1984) A general trait model for response processes. *Psychometrika*, 49, 175-186

[46] Enø, C., Georgiadis, M.P., Johnson, W.O. (2006) Estimation of sensitivity an specificity of diagnostic tests and desease prevalence when the true desease state is unknown. *Preventice Veterinary Medicine*, 45, 61 – 81

[47] Feinstein, A.R., Cicchetti, D., (1990) High agreement but low Kappa I, the problems of two paradoxes. *Journal of Clinical Epidemiology* 43 (6), 543 – 549

[48] Feldman, L. A. (1995) Valence focus and arousal focus: INdividual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, 69, 153–166

[49] Feller, W.: Introduction to probability theory and its applications. John Wiley & Sons, Inc., New York 1968

[50] Ferrando, P. J. (2004) Person reliability in personality measurement: an item response theory analysis. *Applied Psychological Measurement*, 28(2), 126–140

[51] Ferrando, P.J. (2007) A Pearson-Type-VII item response model for assessing person fluctuation. *Psychometrika*, 72(1), 25–41

[52] Fischer, G. H. (1973) The linear logistic test model as an instrument in educational research. *Acta Psychologica*,36, 207–220

[53] Fischer, G. H.: Einführung in die Theorie psychologischer Tests. Bern 1974

[54] Fischer, G.H., Molenaar, I.W.: Rasch models – Foundations, recent developments, and applications. New York, Springer 1995

[55] Fischer, G. (1995). Derivations of the Rasch model. In G. Fischer & I.W. Molenaar (Eds.), Rasch models: Foundations, recent developments, and applications (pp. 15-38). New York: Springer.

[56] Fischer, G.H., Formann, A.K. (1982) Some applications of logistic latent trait models withlinear constraints on the paramters. *Applied Psychological Measurement*,6, 397–416

[57] Fisher, R.A. (1922) On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222 , 309 – 368

[58] Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics*, 10, 422-429.

[59] Fleiss, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–382

[60] Glymour, C.: Social statistics and genuine inquiry: reflections on *The Bell Curve*. In: Devlin et al.: Intelligence, Genes, and Success. New York 1977

[61] Goldberg, L.R. (1960) Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 23, 483-496

[62] Gould, S. J.: Der fasch vermessene Mensch. Suhrkamp, Frankfurt/Main 1988

[63] Graham, P. (1995) Modelling covariate effects in observer agreement studies: The case of nominal scale agreement. *Statistical Methods* 14, 299 – 310

[64] Green, B. F. (1956) A method of scalogram analysis using summary statistics. *Psychometrika*, 21, 79–88

[65] Greenacre, M.: Theory and Applications of Correspondence Analysis. London 1984

[66] Greenacre, M. (1991) Interpreting Multiple Correspondence Analysis. *Applied Stochastic Modelling and Data Analysis*, 7, 195–210

[67] Gulliksen, H. (1950) Theory of mental tests. New York: John Wiley and Sons, 1950.

[68] Guttman, L. (1946) The Test-Retest Reliability of Qualitative Data. *Psychometrika*, 11(2), 81 – 95

[69] Guttman, L. (1950) The basis for scalogram analysis. In: Stouffer et al. Measurement and Prediction. The American Soldier Vol. IV. New York: Wiley

[70] Gwet, K. (2002) Inter-rater reliability: Dependency on trait prevalance and marginal homogeneity. *Series: Statistical Methods for Inter-Rater Reliability Assessment*, 2, 1 – 9

[71] Fahrmeir, L., Hamerle, A., Tutz, G.: Multivariate Verfahren der Statistik. Berlin, New York 1996

[72] Herrnstein, R. J., Murray, C.: The Bell Curve. Intelligence and Class Structure in American Life. New York 1994

[73] Heuser, H.: Lehrbuch der Analysis. Teil 2. Stuttgart 2002

[74] Hölder, O. (1901) Die Axiome der Quantitat und die Lehre vom Mass *Sächsische Akademie Wissenschaften zu Leipzig, Mathematisch-PhysischeKlasse*, 53,1-64.

[75] Holland, P. W. (1990) On the sampling theory foundations of item response theory models. *Psychometrika*, 55, 577–601

[76] Holland, P.W., Hoskens, M. (2003) Classical test theory as a first-order response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika* ,68 (1), 123–149

[77] Hornke, L.F., and Habon, M.W. (1986) Rule-Based Item Bank Construction and Evaluation Within the Linear Logistic Framework. *Applied Psychological Measurement* , 10, 369–380

[78] Hui, S.L. Walter, S.D.(1980) Estimating the errorrates of diagnostic tests. *Biometrics* , 36, 167 – 171

[79] Hui, S. L., Zhou, X. H. (1998) Evaluation of diagnostic tests without gold standard. *Statistical Methods in Medical Research* , 7, 354 – 370

[80] Hwang. H., Montréal, H., Dillon, W. R., Takane, Y. (2006) An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika* ,71, 161–171

[81] Iverson, G., Falmagne, J.-C. (1985) Statistical issues in measurement. *Mathematical Social Sciences* ,10, 131–153

[82] Jöreskog, K. G. (1970) A general method for the analysis of covariance structures. *Biometrika* , 57 (2) 239 – 251

[83] Jöreskog, K. G. (1971) Statistical analysis of sets of congeneric tests. *Psychometrika* , 36(2), 109 – 133

[84] Kahneman, D., Tversky, A. (1973) On the psychology of prediction. *Psychological Review* , 80, 237-251

[85] Kaminsky, L. J.: Science and Politics of IQ. Lawrence Erlbaum Associates Inc 1974

[86] Kendall, M.G., Stuart, A.: The Advanced Theory of Statistics, Vol. 1, Distribution Theory, London, Griffin 1969

[87] Kendall, M.G., Stuart, A.: The Advanced Theory of Statistics, Vol. 2., London, Griffin 1973

[88] Kenny, D. A., Judd, C. M. (1984) Estimating the nonlinear interactive effects of latent variables. *Psychological Bulletin* , 96 (1), 201 – 210

[89] Klauer, K.C., Sydow, H. (1992) Interindividuelle Unterschiede in der Lernfähi9gkeit: zur Analyse von Lernprozessen bei Kurzzeitlerntests. *Zeitschrift für Differentielle und Diagnostische Psychologie* , 3, 175–190

[90] Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A.: Foundatons of measurement. Vol. I. New York 1971

[91] Kuhn, T. S.: the structure of scientific revolutions. 2nd ed. The University of Chicago Press, Chicago 1970

[92] Kuhn, T. S. (1961) The function of measurement in modern physical science. *Isis*,52, 161–190

[93] Landis, J.R., Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data *Biometrics*, 33, (1), 159–174

[94] Lipsitz, S. R., Parzen, M., Fitzmaurice, G. M., Klar, N. (2003) A two-stage logistic regression model foranalyzing inter-rater agreement. *Psychometrika*, 68 (2), 289–298

[95] Lord, F. M. (1958): Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika* 23, 291-296

[96] Lord, F. M., Novick, M.R.: Statistical Theories of Mental Tests. Reading (Mass.) 1968

[97] Luce, R.D., Narens L. (1987) Measurement scales on the continuum. *Science*, 236, 1527 – 1532

[98] Lucke, J. F. (2005) The α and the ω of congeneric test theory: an extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement*,29, 65 - 81

[99] Lumsden, J. (1976) Test theory. *Annual Revies of Psychology*, 27, 251 – 280

[100] Lumsden, J. (1977) Person Reliability. *Applied Psychological Measurement*, 1 (4), 477–482

[101] Macmillan, N.A., Creelman, C.D.: Detectin theor – user's guide. Mahwah, New jersey, 2005

[102] Magidson, J., Vermunt, J. K. (2002) Latent class models for clustering: a comparison wiht K-means. *Canadian Journal of Marketing Research*, 20, 37 – 44

[103] Martin-Löf, P. (1973) Statistika modeller: Anteckningar från seminarier Iasåret 1969-1971, utarbetade av Rolf Sundbeerg. Obetydligt ändrat nytrick. October 1973. Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

[104] Masters, G. N.: (1982) A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174

[105] McClish, D., Quade, D. (1983) Improving estimates of prevalence by repeated testing. *Biometrics*, 41, 81 – 89

[106] McDonald, R.D., Ahlawat, K. S. (1974) Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 27, 82 – 99

[107] Meehl, P. E.: Clinical versus statistical prediction. Minnneapolis 1963 (5. Auflage, erste Auflage 1954)

[108] Mereditch, W. (1993) Measurement invarian ce, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543

[109] Micko, H. C. (1969) A psychological scale for reaction time measurement. *Acta Psychologica*, 30 Attention and Performance II (W.G. Koster, ed.) 324–335

[110] Mischel, W. Shoda, Y. (1998) Reconciling processing dynamics and personality dispositions. *Annual Review of Psychology*,49, 229–258

[111] Molenaar, P.C.M., Huizenga, H.M., Nesselroade, J.R.: (2003) The relationship between the structure of interindividual and intraindividual variability: A theoretical and empirical vindication of developmental systems theory. In: U.M. Staudinger, U. Lindenberger (Eds.), Understanding human development: Dialogues with lifespan psychology (pp. 339–360). Dordrecht: Kluwer Academic Publishers.

[112] Muraki, E., Carlson, J.E. (1997) Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, 19, 73-90

[113] Muthé, B. O. (2002) Beyond SEM: General latent variable modelling. *Behaviormetrika*, 29 (1), 81 – 117

[114] Narens, L., Luce R.D. (1986) Measuremen:The theory of numerical assignments. *Psychological Bulletin*, 99, 166-180

[115] Nelder, J.A., Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society, A*, 135, 370–384

[116] Novick, M.R., Lewis, C. (1967) Coefficient alpha ant the reliability of measurements. *Psychometrika*, 32 (1), 1 – 13

[117] punglia, R.S., D'Amico, A.V., Roehl, K.A.,kuntz, K.M. (2005) Impact of age, benign prostatic hyperplasia, and cancer on prostate-specific antigen level. *Cancer*, 106 (7), 1507 – 1513

[118] Ramsey, J. (1975) Review of *Foundation of Measerement*,Vol. I, by D.H. Krantz, R. D. Luce, P. Suppes and A. Tversky. *Psychometrika*„ 40, 257–262

[119] Rasch, G.: Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press, (1960/1980)

[120] Rasch, G.: On general laws and the meaning of measurement in psychology. Berkeley, University of California Press 1961

[121] Rasch, G. (1977) On specific objectivity: an attempt at formalising the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy* 14, 58–94

[122] Rindler, S.E. (1979) Pittfalls in assessing test speededness. *Journal of Educational Measurement* 16, 261–270

[123] Rist, F., Glöckner-Rist, A., Demmel, R. (2009) The Alcohol Use Disorders Identification Test revisited: Establishing its structure using nonlinear factor analysis and identifying subgroups of respondents using latent class analysis. *Drug and Alcohol Dependence*, 100, 71 – 82

[124] Rost, J.: Lehrbuch Testtheorie – Testkonstruktion. Bern 2004

[125] Samejima, F. (1974) Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, 39, 111-121

[126] Samejima, F. (1977a) Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, 42 (2), 193 – 198

[127] Sarbin, T. R. (1943) A Contribution to the Study of Actuarial and Individual Methods of Prediction. *The American Journal of Sociology*, 48, 593-602

[128] Scheiblechner, H.H. (1972) Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift für Experimentelle und Angewandte Psychologie*, 19, 476–506.

[129] Scott, W.A. (1955) Reliability of Content Analysis: The case of nominal scale coding. *The Public Opinion Quarterly*

[130] Shrout, P.E., Fleiss, J.L. (1979) Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 86(2), 420-428

[131] Sijtsma, K., Junker, B. W. (2006) Item response theory: past performance, present developments, and future expectations. *Behaviormetrika*, 33, 75–102

[132] Spearman, C. (1910) Correlation calculated from faulty data. *British Journal of Psychology*,3, 271 –195

[133] Suppes, P.M., Jerman and D. Brian: Computer-assisted instruction: Standford's 1965-66 Arithmetic Program. New York, London: Academic Press 1968

[134] Suppes, P., Zanotti, M. (1981) When are probabilistic explanations possible? *Synthese*, 48, 191–199

[135] Suppes, P., Zinnes, J.L. Basic Measurement Theory. In: Luce, R. D., Bush, R. R., Galanter, E.: Handbood of Mathematical Psychology, Vol. I. New York 1963

[136] Uebersax,J. (1987) Diversity of decision making modelsand the measurement of interrater-agreement. *Psychological Bulletin*, 101 (1), 146–146

[137] Uebersax,J. (1990) Latent class analysis of diagnostic agreement. *Statistics in Medicine*, 9, 559–572

[138] Takane, Y., de Leeuw, J. (1987) On the relationship between item response theory and factor analyssis of discretized variables. *Psychometrika*, 52 (3), 393 – 408

[139] Tanner, M.A., Young, M.A. (1985) Modelling agreement among raters. *Journal of the American Statistical Association*, 80, 175 – 180

[140] Tocher, J.F. (1908) Pigmentation Survey of School Children in Scotland. *Biometrika*, Vol. 6, No. 2/3 (Sep., 1908), pp. 129-235

[141] Tversky, A. (1967) A general theory of polynomial conjoint measurement. *Journal of Mathematical Psychology*, 4, 1–20

[142] Uebersax, J., Grove, W. (1990) Latent class analysis of diagnostic agreement. *Statistics in Medicine*, 9, 559 – 572

[143] Vorberg, D., Schwarz, W. (1990) Rasch-representable reaction time distributions. *Psychometrika*, 55, 617-632

[144] Walter, S.D., Irvig, L. M. (1988) Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *Journal ofClinical Epidemiology*, 41, 923 – 317

[145] Warrens, M. J. (2008) On association coefficients for 2×2-tables and properties that do not depend on the marginal distributions. *Psychometrika*, 73, 777–789

[146] Wason, P.C. (1960) On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129-140.

[147] Weiss, D. J. (1973) The stratified computerized ability test. Research Report (73-3) Minneapolis: University of Minnesota, Department of Psychology,Psychometric Methods Program (NTIS No. AD-768376)

[148] Westlund, K.B., Kurland, L.T. (1953) Studies on multple sclerosis in Winnipeg, Manitoba, and New Orleans, Louisana. *American Journal of Hygiene*, 57, 380–396

[149] Wirtz, M., Caspar, F.: Beurteilerübereintstimmung und Beurteilerreliabilität. Göttingen 2002

[150] Whiteley, S.E. (1980) Multicomponent latent trait models for ability tests. *Psychometrika*, 45, 479-494

[151] Wood, R. (1978) Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology*, 31, 27–32

[152] Wright, B. D. (1997) A history of social science measurement. *Educational Measure Issues Pract.*,16, 33–45

[153] Zimmermann, D.W. (1975) Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, 40(3), 395 – 412

Index

κ-Koeffizient, gewichtet, 107
d', Detektierbarkeit, 90

Antwortfunktion, 31
Antwortkodierung, 18
Antwortmuster, 192
Approximationssatz (Weierstraß), 233
Ashenfelters Formel, 13
asymptotisch normalverteilt, 222
Asymptotische Normalität, 219
Attenuierung, 186
Attenuierungsformeln, 190

base rate, 12
base rate fallacy, 11
Basis, 247
Basis-R-Faktorenmodell, 261
Bayes
    Satz von, 63
between-subject Modell, 71
Bias, 79, 98
    konfirmatorischer, 9
Bias-Index, 99
biasfrei, 80
binomial, compound, 205
Binomialverteilung, verallgemeinerte, 205
Biplot, 127, 267
Birnbaum-Modell, 212
Bradley-Terry-Luce-Modell, 295
Bradley-Terry-Modell, 214

Chi-Quadrat-Distanz, 264
chicken sexers, 144
Clustering, hierarchisches, 239
conjoint measurement, 292
Cronbachs α, 160
cut off, 93

Datenstruktur, 217, 293
Datenstrukturen
    homogene, 294
    inhomogene, 294
discriminal dispersion, 239
Discriminal Process, 57
Diskriminanzfunktion, 64

Effizienz
    relative, 81
Eichstichprobe, 27
Exponentialfamlie, 210
extensive Struktur, 290

Faktorisierung, 84
Faktorisierungstheorem
    Fishers, 83

Gültigkeitskoeffizient, 185
Gültigkeitsparadox, 187
gain, 190
Gauß-Verteilung, trunkierte, 91
Gesamttestwert, 22
Gewichtungsformel (item-weighting formula, 20
Goldstandard, 99, 102
    imperfekter, 111
graded response model, 240
Grundquote, 61
Grundquoten, 12, 39
Grundquotentäuschung, 11

Heterogenität, 15
homogene Items, 22
Homogenität, 15, 22
    Korrelation zwischen Items, 177
    lokale, 69
Homomorphismus, 289
Homoskedastizität, 194

Indifferenzkurven, 292
Indikatorvariable, 20
interne Konsistenz, 170
Interpretation
    konstrukivistscher Ansatz, 67
    operationalistischer Ansatz, 67

Intraklassenkorrelationskoeffizient (IKK), 120
ipsativ, 225
IQ-Transformation, 26
Item Characteristic Function, 31
Item Response Function, 31
item scoring rule, 20
Itemfunktion, 31
Items, dichotome, 31

Kendall-Summe, 116
Klassifikation, 61
Kodierung
    dichotome, 19
    polynome, 19
kongenerisch, 150
kongenerischer Test, 201
Konsistenz, interne, 166
Konstrukt, 16
Konstruktvalidität, 16
Korrelation
    tetrachorische, 252
    Vierfelder (ϕ-), 248
Kuder-Richardson-20, 167
Kuder-Richardson-21, 167

latent trait, 31
latente Klassen, 61
latente Variable, 31, 186
Law of comparative Judgment, 58
LC Cluster Modelle, 259
LCJ, 58
Likelihood, 12, 74
Likelihood-Funktion, 218
Likelihood-Quotient, 92
linear unabhängig, 246
linear-logistisches Modell, 231
Link-Funktion, 211
log-Likelihood, 74
logistische Verteilung, 43
logistische/kategoriale Regression, 211
logistisches Modell, 36, 43
Logit, 215
Lognormalverteilung, 91
Lokale stochastische Unabhängigkeit, 38
lokale Unabhängigkeit, 139

manifeste Variable, 31
marginal likelihood, 230
Maximum-Likelihood, 217
Meßtheorien, 33
Messfehlertheorie, 141
Messungen
    (τ)-äquivalente, 149
    abgeleitete, 294
    fundamentale, 273, 294
Messwiederholungen, 144
Modell
    between subject, 70
    dreiparametrisches logistisches (3PL-Modell), 212
    einparametriges logistisches (1PL-Modell), 212
    within subject, 70
    zweiparametriges logistisches (2PL-Modell), 212
Mortons Dämon, 10
multiple choice, 20

negative predictive value, 88
Normal Ogive Model, 35
normativ, 225
Normierung
    z-, 25
NPV, 88
nuisance parameter, 210

Objektivität, 14
    spezifische, 36, 216
odds ratio, 92

Parallele Itemfunktionen, 215
parallele Tests, 149
Parametrisierung
    multiplikative, 214, 278
    subtraktive, 214
part-whole-correction, 183
partial credit scoring model, 240
pay-off, 94

person characteristic function, 49
person reliability, 47
Personenfunktion, 49
positive predictive value, 88
Power, 95
PPV, 88
Prävalenz, 87
Präzisionskoeffizient, 196
Probit, 209
propensity, 68
propensity distribution, 68
Prophezeiung
    sich selbst erfüllende, 10
Prozentrang, 24
Prozentrangnorm, 25

qualitative Merkmale, 193
quantitative Merkmale, 193

random sampling of examinees, 69
Randverteilung, homogéne, 96
Rasch-Homogenität, 213
Rasch-Modell, 36
    mehrdimensionales, 225
Rasch-repräsentierbar, 278
Ratingformat, 18
Receiver-Operating-Characteristic, 90
Relativ, empirisches, 289
Reliabilität, 15, 147
    Retest-, 169
    Split-Half, 155
replicate measurements, 68
Replikationen, 143
Repräsentativität, 11
Reproduzierbarkeitsfehler, 55
Reproduzierbarkeitskoeffizienten, 55
Residuen, 267
Response Function, 31
ROC-Analyse, 88
ROC-Kurve, 90
Rohwerteverteilung, 23

Satz von deMoivre-Laplace, 203
Schätzer, 79
Schwierigkeit, 36, 45, 178
    kategorienspezifische, 225
Schwierigkeitsfaktoren, 251
Scores
    ipsative, 225
    normative, 225
scoring coefficient, 240
scoring functions, 145
Sensitivität, 87, 109
Signalentdeckung, 88
Signierung, 18
Singularwertzerlegung, 65
Singularwertzerlegung, SVD, 125
Skalen, 26
Skalogramm, 53
Spearman-Brown-Formel, 157
Spearman-Brown-Formel für Test doppelter Länge, 156
Spearman-Brown-Formel, allgemein, 158
Speededness-Quotient, 196
spezifisch objektiv, 36
Spezifität, 87, 109
Split-Half-Reliabilität, 169
Stabilitätskoeffizient, 198
Stabilitätsschätzung, 198
StandardMessfehler, 175
Standardschätzfehler, 176
Standardvorhersagefehler, 176
Statistik, 79
    effiziente, 81
    erschöpfende, 81
    erwartungstreue, 79
    konsistente, 80
    suffiziente, 81, 219
stepped-up reliability, 197
stochastisch unabhängig, 38
stochastisches Subjekt, 41, 69
Struktur
    numerische, 289
    qualitative, 289
Swell, 48

Tchebyscheffsche Ungleichung, 26
test characteristic functions, 145
trade-off, 292

Transformation
    T-, 26
    IQ-, 26
    Stanine-, 26
Tremors, 48
Trend
    developmental, 47
Trennschärfe, 15, 36, 45
Trennschärfeindex, 182

Uebereinstimmungsmatrix, 96
Unabhängigkeit
    bedingte, 98
    lokale, 38
underlying variable, 34
utility, 291

valide ("gültig"), 9
Validität, 15, 185
    faktorielle, 189
    Inhalts, 188
    kurrente, 188
    logische, 188
    Konstrukt, 188
    Kriteriums-, 188
    prädiktive, 188
Vektor, 192
Verbundmessung, 292
Verfügbarkeit (availability), 12
Verkettung (concatenation), 289
Verteilung
    logistische, 45

wahrer Wert, 141
    platonische Definition, 144
    operationale Definition, 144
Wahrscheinlichkeit
    a posteriori, 63
    a priori, 62
    modale, 104
Wahrscheinlichkeitselement, 218
Wahrscheinlichkeitsverteilung, 40
Weibull-Funktion, 202
Wettquotient, 215