Hans Irtel

Entscheidungs- und testtheoretische Grundlagen der Psychologischen Diagnostik

Universität Mannheim

1995

Vorwort

Dieses Buch ist ein Kompendium grundlegender Konzepte der Test- und Entscheidungstheorie, wie sie in der Psychodiagnostik zur Anwendung kommen. Sowohl in der Auswahl des Stoffes als auch in dessen Darstellung orientiert es sich wesentlich an dem Vorlesungszyklus „Psychologische Diagnostik“ von Prof. Dr. Jan Drösler an der Universität Regensburg. Die Niederschrift ist das Ergebnis meiner Auseinandersetzungen mit dieser Materie, die für mich als Student der Universität Regensburg in den Jahren 1975 bis 78 begannen, dann als Assistent mit Übungen zur Vorlesung von Prof. Drösler weitergeführt wurden und schließlich im Rahmen meiner Lehrbefugnis als Privatdozent im Wintersemester 1989/90 in eine eigene Vorlesung mündeten. Wenn auch die Psychodiagnostik nie im Mittelpunkt meines Forschungsinteresses stand, so konnte ich, angeregt durch die in Regensburg traditionell intensive Beschäftigung mit der psychologischen Meßtheorie, doch auch einige eigene Beiträge zur Lösung psychodiagnostischer Probleme leisten: zusammen mit Franz Schmalhofer die Entwicklung einer ordinalen Meßstruktur für psychologische Tests (Irtel & Schmalhofer, 1982) und eine Klärung und Erweiterung des Konzepts der spezifischen Objektivität psychodiagnostischer Messungen (Irtel, 1987, 1993, 1995).

Die von Prof. Drösler sehr stark an der sich damals neu entwickelnden Meßtheorie orientierte Darstellung der Psychodiagnostik hat nicht nur mein eigenes Interesse an diesem Forschungsgegenstand geweckt, sondern hat auch andere Kollegen, die damals als Studierende oder Mitarbeiter Zuhörer der Vorlesungen und Seminare waren, zu eigenen wissenschaftlichen Beiträgen angeregt: Alfred Hamerle (Hamerle, 1982; Hamerle & Tutz, 1980) und Gerhard Tutz (Tutz, 1986, 1989).

Ziel dieses Kompendiums ist es, die Grundkonzepte der Test- und Entscheidungstheorie durchschaubar zu machen. Konzeptuellen Problemen wird daher mehr Platz eingeräumt als Detailproblemen der Anwendung. Da sowohl für das Verständnis der Entscheidungs- als auch der Testtheorie erhebliche Vorkenntnisse auf dem Gebiet der Wahrscheinlichkeitstheorie unumgänglich sind, werden deren Grundbegriffe dem eigentlichen Stoff vorangestellt. Dieser Abschnitt ist zwar primär zum Nachschlagen gedacht, eignet sich aber auch zum selbständigen Durcharbeiten und wurde nicht zuletzt dem Text vorangestellt um klarzumachen, daß es sich hierbei um unverzichtbare Voraussetzungen für das Verständnis des restlichen Stoffes handelt.

Der Hauptteil besteht aus 3 voneinander weitgehend unabhängigen Abschnitten: je einer zur klassischen Testtheorie, den logistischen Testmodellen und der Entscheidungstheorie. Auch wenn die klassische Testtheorie und die logistischen Testmodelle konkurrierende Methoden psychodiagnostischer Messung sind und aus der Sicht der Meßtheorie den logistischen Modellen der Vorzug zu geben ist, werden die Grundkonzepte der klassischen Testtheorie hier abgehandelt. Dies ist unumgänglich, da Begriffe wie „Reliabilität“ und „Validität“ nicht nur in der psychologischen Forschung, sondern auch in der Praxis allgegenwärtig sind. Man kann es auch so ausdrücken: Jemand der diese Begriffe nicht kennt, kann kein Diplom im Fach Psychologie erhalten.

Historisch betrachtet ist der Ausgangpunkt dieses Textes die Auseinandersetzung mit den logistischen Testmodellen, insbeondere mit ihren vielfältigen Beziehungen zu algebraischen und probabilistischen Meßmodellen, wie sie auch in der Psychophysik angewandt werden. Hier ergaben sich unter anderem durch Fortschritte in der Theorie der Bedeutsamkeit Entwick-

lungsmöglichkeiten, die in der psychodiagnostischen Literatur längere Zeit nicht berücksichtigt wurden. In die Testtheorie wurden sie durch Raschs Konzept der spezifischen Objektivität eingeführt. Die Verbindung dieses Konzepts mit der Meßtheorie wurde allerdings erst später herausgearbeitet (Fischer, 1987, 1988; Irtel, 1987).

Das Einbeziehen des Stoffgebietes „Entscheidungstheorie“ trägt der Tatsache Rechnung, daß die praktische Anwendung psychodiagnostischer Messung in der Regel kein Selbstzweck, sondern die Grundlage für möglicherweise weitreichende, diagnostische Entscheidungen ist. Die Methoden der Entscheidungstheorie erlauben es, diese Entscheidungen auf einer rationalen und damit auch durchschaubaren Grundlage zu treffen, auch wenn ihnen ein hohes Maß an Unsicherheit anhaftet.

Wie bereits früher angedeutet, legt dieser Text ganz im Sinne der Lehrveranstaltungen von Prof. Drösler das Hauptgewicht auf die Vermittlung von Einsicht in die grundlegenden Konzepte der Psychodiagnostik. Dies geschieht nicht zuletzt deshalb, weil im Gegensatz zu den Detailfertigkeiten der Testentwicklung und Anwendung das Verständnis der Grundkonzepte diagnostischer Methoden für jede Diplom-Psychologin und jeden Diplom-Psychologen eine unverzichtbare Voraussetzung der wissenschaftlich begründeten Berufsausübung ist. Da diese Konzepte mathematisch formuliert sind, kann hier auf einen für psychologische Literatur überdurchschnittlichen Gebrauch von Mathematik nicht verzichtet werden. Schließlich sollen die Aussagen und Behauptungen der Psychodiagnostik nicht nur dargestellt, sondern auch begründet werden. Die mathematischen Methoden der Psychodiagnostik bei weitgehendem Verzicht auf Mathematik durch Verbalisieren formaler Sachverhalte zu vermitteln, scheint mir weder möglich, noch dem tieferen Verständnis förderlich zu sein. Allein zur Textvereinfachung wesentliche Fortschritte der letzten 30 Jahre psychodiagnostischer Forschung außer Acht zu lassen, wie das beispielsweise Fisseni (1990) oder Tent und Stelzl (1993) tun, ist sicher einer wissenschaftlich fundierten Berufsausübung nicht förderlich. Der Text setzt daher die mit der allgemeinen Hochschulreife verbundenen mathematischen Grundkenntnisse und zumindest teilweise auch den Stoff der statistischen Ausbildung des Grundstudiums voraus. Alles darüber hinausgehende wird im Detail dargestellt.

Die Anfänge dieses Textes gehen auf die Vorbereitung meiner ersten Vorlesung zur Psychodiagnostik im Wintersemester 1989/90 zurück. Ich danke den Studierenden der Universität Regensburg, die seitdem frühere Fassungen als Prüfungsvorbereitung durchgearbeitet und mir entweder durch ihre Verständnisfragen zahlreiche Hiweise gegeben oder nach bestandener Prüfung das Manuskript mit vielen Anmerkungen versehen zurückgebracht haben. Jürgen Heller, Bettina Laugwitz und Oliver Wilhelm danke ich für wertvolle Hinweise zum Inhalt und die Hilfe beim Korrekturlesen der Endfassung. Auch wenn, wie bereits oben erwähnt, die Auswahl des Stoffes wesentlich durch das Vorlesungskonzept von Prof. Drösler bestimmt ist, so trage ich doch allein die Verantwortung für alle verbliebenen Fehler und Unzulänglichkeiten des Textes.

Hans Irtel Mannheim, im August 1995

Inhaltsverzeichnis

1 Grundbegriffe der Wahrscheinlichkeitslehre

Selbst bei weitgehender Kontrolle aller Einflußgrößen sind die Ergebnisse psychologischer Datenerhebungen nicht mit Sicherheit vorherzusagen. Auch Wiederholungen unter genau gleichen Bedingungen führen in der Regel nicht zu den gleichen Ergebnissen. Die Analyse psychologischer Daten ist daher auf die Methoden der Wahrscheinlichkeitstheorie angewiesen. Dieses Kapitel gibt eine eher informelle Einführung in die grundlegenden Konzepte der Wahrscheinlichkeitstheorie. Sie stützt sich auf Brémaud (1988), Chung (1975), Gnedenko (1968), Goldberg (1969) und Hogg und Craig (1978). Das mathematische Standardwerk über Wahrscheinlichkeitstheorie ist Bauer (1991).

1.1 Mengen

Der Begriff der *Menge* wird nicht definiert, man regelt seinen Gebrauch durch Erläuterungen und Beispiele. Eine *Menge* ist eine Zusammenfassung von bestimmten, wohlunterschiedenen Objekten. Die Objekte heißen *Elemente* der Menge. Ist A eine Menge und a ein Element der Menge A, dann schreibt man $a \in A$, ist a nicht Element von A, schreibt man $a \notin A$. Manche Mengen können durch die Aufzählung ihrer Elemente angegeben werden; dies geschieht durch eine in geschweifte Klammern eingeschlossene Liste der Elemente: $A = \{a_1, \ldots, a_n\}$. Allgemeiner ist die Darstellung einer Menge durch definierende Bedingungen. Soll die Menge aller a mit der Eigenschaft $E(a)$ bezeichnet werden, dann schreibt man $\{a \,|\, E(a)\}$. Werden mehrere Eigenschaften benötigt, dann schreibt man $\{a \,|\, E_1(a), \ldots, E_n(a)\}$ um anzudeuten, daß alle Elemente der Menge alle Eigenschaften $E_1(a)$, …, $E_n(a)$ gleichzeitig erfüllen. Einige häufig gebrauchte Mengen erhalten besondere Bezeichnungen:

1. $\varnothing$ ist die leere Menge: $\varnothing = \{a \,|\, a \neq a\}$.
2. $\mathbb{N}$ ist die Menge der natürlichen Zahlen: $\mathbb{N} = \{1, 2, 3, \ldots\}$.
3. $\mathbb{Z}$ ist die Menge der ganzen Zahlen: $\mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \ldots\}$.
4. $\mathbb{R}$ ist die Menge der reellen Zahlen.

Definition 1.1. Seien A und B Mengen.

1. Die Mengen A und B sind *identisch*, wenn sie dieselben Elemente enthalten: $A = B$ gdw. für alle a gilt $a \in A$ gdw. $a \in B$.
2. A ist eine *Teilmenge* von B, wenn jedes Element von A auch in B ist: $A \subset B$ gdw. aus $a \in A$ folgt $a \in B$. A ist eine *echte* Teilmenge von B, wenn A eine Teilmenge von B ist und es ein Element a gibt, das zwar in B, nicht aber in A enthalten ist.
3. Die *Vereinigungsmenge* $A \cup B$ von A und B ist die Menge, die alle Elemente von A und von B enthält:
$$A \cup B = \{a \,|\, a \in A \text{ oder } a \in B\}.$$
4. Die *Durchschnittsmenge* oder der *Durchschnitt* $A \cap B$ von A und B ist die Menge, die alle Elemente enthält, die sowohl in A als auch in B sind:
$$A \cap B = \{a \,|\, a \in A \text{ und } a \in B\}.$$

5. Die *Differenzmenge* $A \setminus B$ ist die Menge aller Elemente von A, die nicht in B sind:

$$A \setminus B = \{a \,|\, a \in A \text{ und } a \notin B\}.$$

6. Das *Komplement* von A bezüglich einer Menge Ω mit $A \subset \Omega$ ist die Menge aller Elemente von Ω, die nicht in A sind:

$$\overline{A}_\Omega = \Omega \setminus A = \{a \,|\, a \in \Omega \text{ und } a \notin A\}.$$

Aus einer genauen Betrachtung der oben gegebenen Definitionen lassen sich leicht folgende Rechenregeln für die definierten Mengenoperationen ableiten (die Komplementbildung bezieht sich dabei immer auf die gleiche Obermenge Ω):

1. $A \cap B = B \cap A$, $A \cup B = B \cup A$ (Kommutativität);
2. $(A \cap B) \cap C = A \cap (B \cap C)$, $(A \cup B) \cup C = A \cup (B \cup C)$ (Assoziativität);
3. $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$, $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ (Distributivität);
4. wenn $A \subset B$, dann $\overline{B} \subset \overline{A}$;
5. $\overline{(\overline{A})} = A$;
6. $\overline{A \cup B} = \overline{A} \cap \overline{B}$, $\overline{A \cap B} = \overline{A} \cup \overline{B}$ (Regeln von De Morgan).

Definition 1.2. Das *kartesische Produkt* $A \times B$ von A und B ist die Menge aller geordneten Paare (a, b), bei denen das erste Element a aus A und das zweite Element b aus B stammt.

Man beachte, daß zwei geordnete Paare (a, b) und (a', b') genau dann identisch sind, wenn sowohl a und a', als auch b und b' identisch sind: $(a, b) = (a', b')$ gdw. $a = a'$ und $b = b'$. Statt geordneter Paare kann man allgemeiner auch geordnete n-Tupel $(a_1, \ldots, a_n)$ betrachten. Sie sind die Elemente eines n-stelligen kartesischen Produkts $A_1 \times \ldots \times A_n$:

$$A_1 \times \ldots \times A_n = \{(a_1, \ldots, a_n) \,|\, a_1 \in A_1, \ldots, a_n \in A_n\}.$$

Ist $A_1 = \ldots = A_n$, dann schreibt man für $A_1 \times \ldots \times A_n$ auch A^n.

Definition 1.3. Eine Menge R ist eine *binäre Relation*, wenn es Mengen A und B gibt, so daß R eine Teilmenge von $A \times B$ ist: $R \subset A \times B$. Wir sagen dann: R ist eine *binäre Relation auf* $A \times B$. Allgemein ist eine Menge R eine *n-stellige Relation*, wenn es Mengen $A_1, \ldots, A_n$ gibt, so daß $R \subset A_1 \times \ldots \times A_n$.

Ist R eine binäre Relation und (a, b) ein Element aus R, dann schreibt man statt $(a, b) \in R$ in der Regel kürzer aRb.

Definition 1.4. Sei A eine Menge und R eine binäre Relation auf $A \times A$.

1. R ist *reflexiv auf* A, gdw. für alle a in A gilt aRa.
2. R ist *symmetrisch auf* A, gdw. für alle a, b in A gilt: aus aRb folgt bRa.
3. R ist *transitiv auf* A, gdw. für alle a, b, c in A gilt: aus aRb und bRc folgt aRc.

Definition 1.5. Sei A eine Menge und E eine binäre Relation auf $A \times A$. Die Relation E ist eine *Äquivalenzrelation* auf A, gdw. sie auf A reflexiv, symmetrisch und transitiv ist.

Zur Bezeichnung von Äquivalenzrelationen wird häufig das Zeichen $\sim$ verwendet. Man schreibt dann $a \sim b$ wenn $(a, b) \in \sim$.

Definition 1.6. Sei A eine Menge und $\sim$ eine Äquivalenzrelation auf A. Eine Teilmenge K von A ist eine *Äquivalenzklasse* bezüglich der Relation $\sim$, wenn gilt

1. $K \neq \varnothing$;
2. wenn $a \in K$ und $b \in K$, dann ist $a \sim b$;
3. wenn $a \in K$ und $a \sim b$, dann ist $b \in K$.

Ist auf einer Menge A eine Äquivalenzrelation definiert, dann ist jedes Element a von A in genau *einer* Äquivalenzklasse enthalten. Diese wird häufig mit $[\![a]\!]$ bezeichnet: $[\![a]\!] = \{b \in K \mid a \sim b\}$. Zwei Äquivalenzklassen K und K' sind entweder identisch, oder sie haben kein Element gemeinsam.

Definition 1.7. Sei A eine Menge und $\sim$ eine Äquivalenzrelation auf A. Die Menge $A/\sim$ aller Äquivalenzklassen von A bezüglich $\sim$ ist die von $\sim$ *induzierte Zerlegung* von A.

Statt *Zerlegung* wird die Menge $A/\sim$ auch *Quotientenmenge von A bezüglich $\sim$* genannt. Die Elemente von $A/\sim$ sind Mengen. Ihre Vereinigung ist A, ihr paarweiser Durchschnitt ist leer.

Definition 1.8. Seien A und B Mengen und f eine binäre Relation auf $A \times B$. Die Relation f heißt *Abbildung von A nach B*, wenn f die folgenden Bedingungen erfüllt:

1. f ist *linkstotal:* Für alle a in A gibt es ein b in B, so daß (a, b) in f.
2. f ist *rechtseindeutig:* Sind (a, b) und (a, c) in f, dann ist $b = c$.

Die Menge A heißt *Definitionsbereich* und die Menge B heißt *Wertebereich* der Abbildung f. Ist M eine Teilmenge von A, dann heißt

$$f(M) = \{b \in B \mid \text{ es gibt ein } a \in M \text{ mit } (a, b) \in f\}$$

das *Bild* von M, und für eine Teilmenge N aus B heißt

$$f^{-1}(N) = \{a \in A \mid \text{ es gibt ein } b \in N \text{ mit } (a, b) \in f\}$$

das *Urbild* von N. Die Abbildung $f|M$ von M nach B heißt *Beschränkung* von f auf M. Sie unterscheidet sich von f nur durch den eingeschränkten Definitionsbereich.

Ist f eine Abbildung, dann schreibt man für $(a, b) \in f$ in der Regel $b = f(a)$. Statt *Abbildung* wird auch der synonyme Ausdruck *Funktion* benutzt. Um anzudeuten, daß eine Abbildung von A nach B definiert ist, verwendet man die Schreibweise $f\colon A \to B$. Die Zuordnung von a aus A zu $f(a)$ aus B wird durch $a \mapsto f(a)$ dargestellt.

Definition 1.9. Sei $f\colon A \to B$ eine Abbildung.

1. f ist *surjektiv*, wenn es zu jedem b in B ein a in A gibt, so daß $b = f(a)$. Surjektive Abbildungen sind also Relationen, die sowohl links- als auch rechtstotal (*bitotal*) und rechtseindeutig sind.
2. f ist *injektiv*, wenn mit a, a' in A aus $f(a) = f(a')$ folgt $a = a'$. Injektive Abbildungen werden auch *eineindeutig* genannt, da sie sowohl rechts- als auch linkseindeutig sind. Man spricht auch von *Einbettungen.*
3. f ist *bijektiv*, wenn f surjektiv und injektiv ist. Bijektive Abbildungen sind bitotale und eineindeutige Relationen, die vor allem bei der Betrachtung von strukturgleichen Systemen (Isomorphismen) eine Rolle spielen.

Definition 1.10. Sei $f\colon A \to B$ eine bijektive Abbildung. Dann besteht für jedes b in B das Urbild $f^{-1}(\{b\})$ aus einem einzigen Element, das mit $f^{-1}(b)$ bezeichnet wird. Dies definiert eine Abbildung $f^{-1}\colon B \to A$, die jedem b in B ein $f^{-1}(b)$ in A zuordnet, sie wird *Umkehrabbildung* von f genannt.

Mit Hilfe bijektiver Abbildungen kann die Mächtigkeit von Mengen definiert werden. Wir verzichten hier allerdings auf die Einführung des Begriffs der Gleichmächtigkeit und der Kardinalzahlen, da es für uns ausreicht, wenn wir zwischen endlichen und unendlichen Mengen unterscheiden können.

DEFINITION 1.11.
1. Eine Menge A ist *endlich*, wenn es eine natürliche Zahl n und eine bijektive Abbildung von A auf $\{i \mid i \in \mathbb{N}, 1 \leq i \leq n\}$ gibt.
2. Eine nichtleere Menge A ist *unendlich*, wenn sie nicht endlich ist.
3. Eine nichtleere Menge A ist *abzählbar*, wenn es eine surjektive Abbildung von $\mathbb{N}$ auf A gibt.
4. Eine nichtleere Menge A ist *überabzählbar*, wenn sie nicht abzählbar ist.

Unendliche Mengen A lassen sich auch durch die etwas kontraintuitive Eigenschaft charakterisieren, daß sie eine echte Teilmenge M besitzen und dabei gleichzeitig eine bijektive Abbildung von A nach M existiert.

1.2 Zufallsexperimente

Ein Experiment, dessen Ergebnis nicht mit Sicherheit vorhergesagt werden kann, wird als *Zufallsexperiment* bezeichnet, wenn die Menge aller möglichen Ergebnisse des Experiments bekannt ist. Diese Menge wird als *Ergebnis-* oder *Stichprobenraum* bezeichnet.

Beispiel. In einem Detektionsexperiment soll die Versuchsperson ein sehr schwaches Signal entdecken, das in ein Störgeräusch eingebettet ist. Es gibt zwei mögliche Ergebnisse: die Versuchsperson entdeckt das Signal $(+)$, oder sie entdeckt es nicht $(-)$. Der Ergebnisraum ist dann die Menge $\{+, -\}$, die diese beiden Ergebnisse enthält.

Zur wahrscheinlichkeitstheoretischen Analyse eines Experiments haben wir als erstes eine geeignete Formalisierung des möglichen Geschehens im Experiment zu geben. Diese Formalisierung besteht im wesentlichen aus einer Aufstellung aller möglichen Ereignisse und einer präzisen Formulierung der zu untersuchenden Hypothesen.

1.2.1 Der Ergebnisraum

Beispiel. Die Datenerhebung für ein Experiment bestehe darin, daß ein Ehepaar befragt wird, ob es für eine generelle Geschwindigkeitsbeschränkung auf Autobahnen sei. Wir können dann als Ergebnis notieren, wieviele „Ja"-Antworten wir bei einem Paar erhalten. Die Ergebnisse des Experiments entsprechen damit jeweils genau einem Element der Menge $\Omega_1 = \{0, 1, 2\}$.

Die Menge Ω_1 ist ein Ergebnisraum des Experiments. Sie enthält alle Ergebnisse, die möglich sind und für die Fragestellung des Experiments von Bedeutung sein könnten. Der Ergebnisraum kann auch als die Menge der möglichen Datenpunkte betrachtet werden, da seine Elemente—die Ergebnisse des Experiments—in der Regel genau diejenigen Informationen sind, die bei einem Experiment aufgezeichnet werden.

Die Definition des Ergebnisraums eines Experiments hängt nicht nur von der formalen Struktur des Experiments ab, sondern wird ganz wesentlich auch von der Art der Fragestellung bestimmt. Wir werden dies in unserem Beispiel spätestens dann bemerken, wenn wir an Hand der erhobenen Daten wissen wollen, wie Ehefrauen über das angesprochene Problem denken. Die Information, wer welche Antwort gab, ginge nämlich beim Ergebnisraum Ω_1 verloren. Um diese Frage beantworten zu können, muß der Ergebnisraum die Information enthalten, welcher Ehepartner welche Antwort gegeben hat. Wir können etwa „JN" notieren, wenn die Frau „Ja" und der Mann „Nein"

sagt, „JJ“, wenn beide „Ja“ sagen usw. Damit erhalten wir den Ergebnisraum $\Omega_2 = \{JJ, JN, NJ, NN\}$.

Wir sehen an diesen verschiedenen Ergebnisräumen, daß die Wahrscheinlichkeitslehre nicht bestimmt, wie ein Experiment zu beschreiben ist, bzw. wie die Grundbegriffe der Wahrscheinlichkeitstheorie in einem Experiment interpretiert werden müssen. Die Anwendung der Wahrscheinlichkeitslehre setzt die korrekte empirische Deutung ihrer Grundbegriffe voraus und sagt selbst nur darüber etwas, wie mit diesen Grundbegriffen umzugehen ist.

Für den ersten Grundbegriff der Wahrscheinlichkeitstheorie, den Ergebnisraum, haben wir oben zwei Beispiele kennengelernt. Auch wenn aus der Wahrscheinlichkeitslehre nicht für beliebige Experimente und Fragestellungen abgeleitet werden kann, wie der Ergebnisraum zu bestimmen ist, so lassen sich doch bestimmte Kriterien dafür angeben, ob eine Menge grundsätzlich als Ergebnisraum geeignet ist oder nicht.

Definition 1.12. Ein *Ergebnisraum* Ω eines Experiments ist die Menge aller möglichen Ergebnisse des Experiments. Für jeden Ergebnisraum Ω muß gelten:

1. Jedes Element der Menge Ω bezeichnet ein mögliches Ergebnis des Experiments.
2. Jedem Ergebnis des Experiments entspricht genau ein Element von Ω.

Wir sehen an unseren Beispielen, daß für ein Experiment mehrere Ergebnisräume definiert werden können, je nach gewünschter Fragestellung. Allgemein ist es empfehlenswert, einen möglichst differenzierten Ergebnisraum zu definieren, da dann auch mehr Fragestellungen untersucht werden können, wie dies an den Beispielen deutlich wurde.

1.2.2 Ereignisse

Obwohl es aus praktischen Gründen nicht möglich ist, im Ergebnisraum die Versuchssituation vollständig zu beschreiben, muß dieser immer in Abhängigkeit von der Fragestellung des Experiments definiert werden. Der Ergebnisraum ist die Grundmenge für wahrscheinlichkeitstheoretische Überlegungen zu einem Experiment. Alle Fragen, die anhand der Daten untersucht werden sollen, bauen auf dem Ergebnisraum auf. Wir können uns etwa für das Ereignis „Beide Ehepartner geben die gleiche Antwort“ interessieren. Dieses Ereignis tritt ein, wenn „JJ“ oder wenn „NN“ beobachtet wird. Das Ereignis „gleiche Antwort“ ist also selbst als Teilmenge von Ω_2 darstellbar. Eine mathematische Darstellung des Ereignisses „gleiche Antwort“ ist die Menge $A = \{JJ, NN\}$, eine Teilmenge von Ω_2.

Definition 1.13. Ist Ω ein Ergebnisraum, dann ist A ein *Ereignis*, wenn A eine Teilmenge von Ω ist.

Ein Ereignis ist also eine Menge und damit ein Konzept der wahrscheinlichkeitstheoretischen Betrachtungen unseres Experiments, dessen „Eintreten“ nicht direkt empirisch beobachtbar ist. Wir können aber einfach definieren:

Definition 1.14. Ist A ein Ereignis, dann sagen wir *das Ereignis A tritt ein*, wenn ein Ergebnis beobachtet wird, das Element von A ist.

In der Wahrscheinlichkeitstheorie wird also zwischen den Begriffen „Ergebnis“ und „Ereignis“ sorgfältig unterschieden. Ein Ergebnis ist beobachtbar; es ist der Ausgang des Experiments oder das aufgezeichnete Datum. „Ereignis“ ist dagegen ein abstraktes Konzept der Wahrscheinlichkeitslehre, das deshalb auch oben definiert wurde. Definiert sind Ereignisse als Mengen, und sie sind deshalb etwas Theoretisches, nicht direkt Beobachtbares. Die Elemente dieser Mengen sind Ergebnisse, die im Experiment beobachtet werden können. Das Eintreten eines Ereignisses A ist deshalb indirekt dann beobachtbar, wenn ein Ergebnis beobachtet wird, das Element von A ist.

Wir können nun auch Ereignisse definieren, die nur ein einziges Element enthalten, etwa $A = \{JJ\}$. Dieses Ereignis tritt genau dann auf, wenn das Ergebnis „JJ“ beobachtet wird.

Definition 1.15. Ein Ereignis, das nur ein einziges Ergebnis enthält, heißt *Elementarereignis*.

Auch der Ergebnisraum Ω eines Experiments ist selbst ein Ereignis. Die Besonderheit dieses Ereignisses ist, daß es bei jeder Durchführung des Experiments eintritt, da ja entsprechend der Definition jedes mögliche Ergebnis des Experiments in Ω enthalten ist.

Neben Ω gibt es noch ein weiteres—zumindest für die Wahrscheinlichkeitstheorie—wichtiges Ereignis, nämlich das Ereignis $\varnothing$, das überhaupt kein Ergebnis enthält und demnach niemals eintritt, da entsprechend den Eigenschaften von Ω in jedem Fall ein Element von Ω beobachtet wird.

Da Ereignisse als Mengen definiert sind, können die bekannten Mengenoperationen auf sie angewandt und dadurch aus gegebenen Ereignissen neue erzeugt werden. Sind etwa A und B Ereignisse, dann ist $A \cap B$ ebenfalls ein Ereignis. Man weiß dann natürlich auch, wann das Ereignis $A \cap B$ eintritt, nämlich genau dann, wenn ein Ergebnis beobachtet wird, das sowohl in A als auch in B enthalten ist. Wir können etwa A als das Ereignis „Beide Partner geben die gleiche Antwort“ betrachten und B als das Ereignis „Die Frau antwortet mit 'Ja'“. Dann ist $A \cap B$ das Ereignis „Bei einem Ehepaar mit gleicher Meinung sagt die Frau 'Ja'“. Abgekürzt erhalten wir $A = \{JJ, NN\}$, $B = \{JN, JJ\}$ und $A \cap B = \{JJ\}$

Auch die Operation $\cup$ kann zur Bildung neuer Ereignisse benutzt werden: A sei das Ereignis „Die Frau sagt 'Ja'“ und B sei „Der Mann sagt 'Ja'“. Wir erhalten $A = \{JJ, JN\}$, $B = \{JJ, NJ\}$ und $A \cup B = \{JJ, JN, NJ\}$, so daß $A \cup B$ das Ereignis „Mindestens ein Partner antwortet mit 'Ja'“ darstellt.

Die dritte Operation zur Bildung von Ereignissen ist die Komplementbildung. Stellt A das Ereignis „Beide Partner sind einer Meinung“ dar, dann ist $A = \{JJ, NN\}$, und das Komplement $\overline{A} = \{JN, NJ\}$ stellt dann das Ereignis „Die Partner sind unterschiedlicher Meinung“ dar.

Für einen endlichen Ergebnisraum Ω, der n Ergebnisse enthält, können genau 2^n Ereignisse gebildet werden, da eine Menge mit n Elementen genau 2^n verschiedene Teilmengen enthält.

Wir betrachten hier vorerst nur endliche (später auch abzählbare) Ergebnisräume. In diesem Fall ist leicht zu sehen, daß den Elementarereignissen eine besondere Bedeutung zukommt. Mit Hilfe der Mengenvereinigung ist es nämlich möglich, jedes Ereignis als Vereinigung all der Elementarereignisse zu schreiben, die genau die Elemente des zu erzeugenden Ereignisses enthalten. Das Ereignis „Mindestens ein 'Ja'“ mit $A = \{JN, NJ, JJ\}$ läßt sich etwa schreiben als $A = \{JN\} \cup \{NJ\} \cup \{JJ\}$. Diese Möglichkeit wird dann besonders interessant, wenn wir den Ereignissen Wahrscheinlichkeiten zuordnen können.

Ist nämlich dann die Wahrscheinlichkeit für jedes Elementarereignis bekannt, und ist darüber hinaus bekannt, wie man die Wahrscheinlichkeit des Ereignisses $A \cup B$ errechnet, wenn die Wahrscheinlichkeiten von A und B bekannt sind, dann kann für jedes beliebige Ereignis, das im Experiment eintreten kann, die Wahrscheinlichkeit berechnet werden.

Spätestens an dieser Stelle wird auch klar, warum so großer Wert auf den Unterschied zwischen dem Begriff *Ergebnis* („JJ“) und dem Begriff *Elementarereignis* ($\{JJ\}$) gelegt wurde. Der Grund liegt darin, daß man schreiben kann $\{JJ\} \cup \{NN\}$, aber nicht „JJ“ $\cup$ „NN“, da nur $\{JJ\}$ eine Menge ist, nicht jedoch „JJ“. Mit Elementarereignissen kann man im Sinne der Mengenlehre rechnen, mit Ergebnissen nicht.

1.2.3 Wahrscheinlichkeiten

Am Anfang dieser Einführung wurde darauf hingewiesen, daß die Wahrscheinlichkeitstheorie nicht sagt, welcher Ergebnisraum für ein bestimmtes Experiment zu definieren ist. Sie setzt jedoch gewisse Randbedingungen, die eine Menge Ω erfüllen muß, um als Ergebnisraum brauchbar zu sein. Ähnlich ist es bei der Wahrscheinlichkeit, dem zweiten wesentlichen Grundbegriff der Wahrscheinlichkeitstheorie. Die Wahrscheinlichkeitslehre sagt nichts darüber aus, wie man die Wahrscheinlichkeit eines Ereignisses in

einem Experiment erhalten kann. Sie stellt jedoch bestimmte Bedingungen, die Funktionen erfüllen müssen, um als Wahrscheinlichkeiten bestimmter Ereignisse gelten zu können. Darüber hinaus wird festgelegt, wie mit den Wahrscheinlichkeiten umzugehen ist, also wie und was mit ihnen berechnet werden kann.

Für die Definition des Begriffs „Wahrscheinlichkeit“ betrachten wir vorerst *diskrete* Ergebnisräume. Dies sind Ergebnisräume, die nur endlich oder höchstens abzählbar viele Ergebnisse enthalten.

Definition 1.16. Sei Ω ein diskreter Ergebnisraum. Ferner sei P eine Funktion, die jedem Elementarereignis $\{\omega_i\}$ mit $\omega \in \Omega$ genau eine Zahl $P\big(\{\omega_i\}\big)$ zuordnet. Die Funktionswerte $P\big(\{\omega_i\}\big)$ heißen *Wahrscheinlichkeit des Elementarereignisses* $\{\omega_i\}$ genau dann, wenn sie folgende zwei Bedingungen erfüllen:

1. Die Wahrscheinlichkeit jedes Elementarereignisses ist nicht negativ: für alle Elementarereignisse $\{\omega_i\}$ ist $P\big(\{\omega_i\}\big) \geq 0$;
2. Die Summe aller den Elementarereignissen zugeordneten Wahrscheinlichkeiten ist 1:

$$P\big(\{\omega_1\}\big) + P\big(\{\omega_2\}\big) + \ldots = 1.$$

Die Wahrscheinlichkeit beliebiger Ereignisse A wird dann in folgender Weise definiert:

Definition 1.17. Die *Wahrscheinlichkeit eines Ereignisses* A ist die Summe der Wahrscheinlichkeiten aller Elementarereignisse $\{\omega\}$, die Teilmengen von A sind:

$$P(A) = \sum_{\omega \in A} P\big(\{\omega\}\big).$$

Mit dieser Definition kann die Wahrscheinlichkeit aller möglichen Ereignisse berechnet werden. Wir mußten dazu keine Annahme über die speziellen Werte der Wahrscheinlichkeiten von Elementarereignissen machen, insbesondere brauchen diese nicht gleichwahrscheinlich sein.

1.3 Der Wahrscheinlichkeitsraum

Mit den Begriffen „Ereignis“, „Elementarereignis“ und „Wahrscheinlichkeit eines Ereignisses“ sind die drei wesentlichen Grundbegriffe der Wahrscheinlichkeitstheorie eingeführt. Wir sind dabei von diskreten Ergebnisräumen ausgegangen und haben in den Definitionen Begriffe benutzt, die vorher nicht definiert, sondern durch inhaltliche Erläuterungen eingeführt wurden, wie etwa den Begriff „Experiment“ oder „Ergebnis“. Diesen Zugang nennt man in der Mengenlehre oder der Wahrscheinlichkeitstheorie „naiv“, da er keine exakte formale Begründung der Theorie erlaubt. Deshalb soll im folgenden die axiomatische Definition der Wahrscheinlichkeit von Kolmogorov betrachtet werden. Diese ist präziser und allgemeiner als die bisher gegebenen Definitionen und stellt die wichtigste Grundlage der Wahrscheinlichkeitstheorie dar. Die Definition wird schrittweise eingeführt.

1.3.1 Die Axiomatische Definition nach Kolmogorov

Definition 1.18. Sei Ω eine Menge und $\mathfrak{A}$ eine Menge von Teilmengen von Ω, die folgende drei Bedingungen erfüllt:

1. Ω ist Element von $\mathfrak{A}$.
2. Ist A Element von $\mathfrak{A}$, dann ist auch $\overline{A}$, das Komplement von A bezüglich Ω, Element von $\mathfrak{A}$.
3. Falls jedes Element der Folge $A_1, \ldots, A_n, \ldots, n \geq 1$ in $\mathfrak{A}$ ist, dann ist auch die Vereinigung $\bigcup_{i=1}^{\infty} A_i$ in $\mathfrak{A}$.

Die Menge $\mathfrak{A}$ heißt dann *σ-Algebra* (in Ω).

Ist Ω ein Ergebnisraum, dann enthält eine σ-Algebra $\mathfrak{A}$ in Ω alle interessierenden Ereignisse. Definition 1.18 garantiert, daß jede σ-Algebra abgeschlossen gegenüber den Mengenoperationen Komplementbildung, Vereinigung und Durchschnitt ist. Sind bestimmte Teilmengen von Ω in $\mathfrak{A}$ enthalten, dann sind auch alle Mengen, die durch die Anwendung von Mengenoperationen aus diesen erzeugt werden können, in $\mathfrak{A}$ enthalten (vgl. die Übungsaufgaben am Ende dieses Abschnittes).

Jedem Ereignis, also jedem Element in $\mathfrak{A}$, soll durch die Wahrscheinlichkeitsfunktion eine Zahl $0 \leq P \leq 1$ zugeordnet werden, die die Wahrscheinlichkeit des Ereignisses angibt. Die Zuordnung soll so geartet sein, daß sie in einem bestimmten Sinn mit den Mengenoperationen verträglich ist.

Definition 1.19. Sei Ω ein Ergebnisraum und $\mathfrak{A}$ eine zu Ω gehörige σ-Algebra. Ferner sei P eine Abbildung von $\mathfrak{A}$ in die reellen Zahlen. Das Tripel $\langle \Omega, \mathfrak{A}, P \rangle$ heißt genau dann *Wahrscheinlichkeitsraum*, wenn die Funktion P die folgenden drei Axiome erfüllt:

1. $P(A) \geq 0$ für alle A in $\mathfrak{A}$.
2. $P(\Omega) = 1$.
3. Für jede Folge $A_1, \ldots, A_n, \ldots$ paarweise disjunkter Ereignisse aus $\mathfrak{A}$ gilt *σ-Additivität*:
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Aufgrund dieser Definition wird eine Wahrscheinlichkeit auch ein *normiertes, nichtnegatives, vollständig additives Maß* genannt.

Aus der Definition eines Wahrscheinlichkeitsraums lassen sich einige Eigenschaften ableiten, deren Beweise gute Übungsaufgaben bilden:

1. $\varnothing \in \mathfrak{A}$. Da Ω in $\mathfrak{A}$ ist, muß auch dessen Komplement, die leere Menge in $\mathfrak{A}$ sein.
2. Die Menge aller Teilmengen von Ω ist eine σ-Algebra in Ω.
3. $P(\varnothing) = 0$.
4. $P(A) \leq 1$.
5. $P(\overline{A}) = 1 - P(A)$.
6. Wenn $A \subset B$, dann $P(A) \leq P(B)$.
7. Für beliebige Ereignisse A, B gilt $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Die Definition 1.19 bestimmt den Rahmen für die Formalisierung eines Zufallsexperiments. Wie bereits früher bei der Darstellung des naiven Konzepts von Wahrscheinlichkeiten, wird auch durch diese Definition nicht bestimmt, welche konkreten Werte die Wahrscheinlichkeiten bestimmter Ereignisse haben. Es werden nur strukturelle Forderungen an Wahrscheinlichkeitsfunktionen und Ergebnisräume gestellt. Die empirische Interpretation des Tripels $\langle \Omega, \mathfrak{A}, P \rangle$ ist weiterhin frei, solange die strukturellen Erfordernisse erfüllt sind. Für diese Interpretation sind mehrere Möglichkeiten bekannt. Wir stellen hier nur zwei kurz vor, da sie in der Literatur auch häufig als „Definitionen“ der Wahrscheinlichkeit bezeichnet werden. Aus der Sicht von Definition 1.19 stellen sie aber nur spezielle Interpretationen des Konzepts „Wahrscheinlichkeit“ dar.

Beispiel. Wir betrachten ein Zufallsexperiment, das aus dem einmaligen Werfen von zwei Würfeln besteht. Der Ergebnisraum Ω ist die Menge aller Paare der Form $\omega = (i, j)$, wobei i und j die Ziffern 1 bis 6 sind: $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. Als σ-Algebra wird die Menge aller Teilmengen von Ω definiert. Die Wahrscheinlichkeitsfunktion sei bestimmt durch $P\big(\{\omega\}\big) = (\frac{1}{6})^2$ für jedes ω in Ω. Es ist leicht zu sehen, daß dann für alle Ereignisse A in $\mathfrak{A}$ gilt: $P(A) = (\frac{1}{6})^2 \,|\, A \,|$, wobei $|\, A \,|$ die Anzahl von Elementen in A ist. Das Ereignis „eine Summe von 3“ ist $A = \{(1, 2), (2, 1)\}$, seine Wahrscheinlichkeit ist $P(A) = \frac{1}{18}$.

1.3.2 Die Interpretation nach Laplace

Die klassische Interpretation der Wahrscheinlichkeit nach Laplace geht von den Elementarereignissen aus. Dies sind, wie bereits früher definiert, Mengen, die nur ein einziges Ergebnis enthalten und paarweise disjunkt sind. Von jedem Elementarereignis wird angenommen, daß es die gleiche Wahrscheinlichkeit wie jedes andere Elementarereignis hat. Um nun die Wahrscheinlichkeit für ein beliebiges Ereignis A zu finden, wird die Menge der Elementarereignisse in zwei disjunkte Teilmengen zerlegt: die erste Teilmenge enthält alle Ergebnisse, deren Realisierung gleichbedeutend ist mit dem Eintreten von A, die zweite Teilmenge enthält die Ergebnisse, deren Realisierung das Eintreten von $\overline{A}$ bedeutet. Die erste Teilmenge wird als die Menge der für A günstigen Fälle bezeichnet, die Anzahl ihrer Elemente sei m. Die Gesamtzahl aller möglichen Elementarereignisse sei n. Als Wahrscheinlichkeit des Ereignisses A wird dann definiert

$$P(A) = \frac{m}{n},$$

also die Anzahl der dem Ereignis A günstigen Fälle dividiert durch die Gesamtzahl aller möglichen Elementarereignisse. Es kann gezeigt werden, daß diese Definition die Bedingungen von Definition 1.19 erfüllt.

Die Nachteile des Konzepts von Laplace sind aber offensichtlich:

- Es kann nur auf endliche Mengen von Elementarereignissen angewandt werden.
- Es setzt voraus, daß alle Elementarereignisse gleichwahrscheinlich sind.

Vor allem letztere Einschränkung macht die klassische Definition nach Laplace für empirische Anwendungen unbrauchbar. Sie kann bestenfalls zur Analyse von Glücksspielen benutzt werden, wobei selbst dort „ideale“ Spielgeräte, wie etwa Würfel oder Glücksräder angenommen werden müssen. In der Regel sind empirische Elementarereignisse nicht gleichwahrscheinlich, sondern ihre Wahrscheinlichkeit hängt von den empirischen Interpretationen oder Bedingungen ab.

1.3.3 Grenzwerte relativer Häufigkeiten nach von Mises

Das Konzept der statistischen Wahrscheinlichkeitstheorie nach von Mises geht nicht von der Struktur des Ergebnisraums aus, sondern versucht eine rein empirische Begründung zu geben. In diesem System werden Wahrscheinlichkeiten als Grenzwerte relativer Häufigkeiten definiert, wenn die Anzahl der unabhängigen Beobachtungen unendlich groß wird:

$$P(A) = \lim_{n \to \infty} \frac{m}{n},$$

wobei n die Gesamtzahl der Experimente ist und m die Anzahl der Experimente, in denen das zufällige Ereignis A beobachtet wird.

Es ist klar, daß der Anspruch dieser Theorie, eine rein empirische Begründung des Wahrscheinlichkeitskonzepts zu geben, nicht eingelöst wird, da weder aus theoretischen Gründen folgt, daß der Grenzwert des Quotienten $\frac{m}{n}$ tatsächlich gegen einen festen Wert konvergiert, noch diese Annahme empirisch begründet werden kann, da sie eine unendliche Anzahl von Beobachtungen voraussetzt.

Übungsaufgaben

1. Zeigen Sie folgende Konsequenzen von Definition 1.18:
   (a) $\varnothing \in \mathfrak{A}$.
   (b) Falls jedes Element der Folge $A_1, \ldots, A_n$ in $\mathfrak{A}$ ist, dann ist auch der Durchschnitt $\bigcap_{i=1}^{n} A_i$ in $\mathfrak{A}$. Hinweis: Benutzen Sie die Formel von de Morgan:
   $$\bigcap_{i=1}^{n} A_i = \overline{\left(\bigcup_{i=1}^{n} \overline{A_i}\right)}.$$
2. Zeigen Sie, daß aus den Axiomen der Definition 1.19 folgt: $P(\overline{A}) = 1 - P(A)$. Hinweis: $\Omega = A \cup \overline{A}$.
3. Zeigen Sie, daß aus $A \subset B$ folgt, $P(A) \leq P(B)$. Hinweis: Aus $A \subset B$ folgt, daß $A \cup (B - A) = B$.
4. Zeigen Sie für beliebige Ereignisse A, B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
5. Aus $A = \varnothing$ folgt $P(A) = 0$. Gilt dies auch umgekehrt?
6. Zeigen Sie, daß der klassische Wahrscheinlichkeitsbegriff von der Wahrscheinlichkeit als Quotient von Anzahl der günstigen durch Anzahl der möglichen Fälle die Bedingungen 1 - 3 aus Definition 1.19 erfüllt.
7. Nehmen Sie an, die Grenzwerte
   $$\lim_{n \to \infty} \frac{m}{n}$$
   für die Quotienten der Anzahl m von Fällen, in denen ein Ereignis A eintritt, und der Gesamtzahl der Experimente n existieren. Zeigen Sie, daß diese Definition die Axiome 1 bis 3 von Definition 1.19 erfüllt.

1.4 Bedingte Wahrscheinlichkeiten

Beispiel. Eine Versuchsperson bekommt eine Liste von Vokabeln, beispielsweise englisch-deutsche Wortpaare, für eine Minute zum Lernen. Danach wird sie abgefragt, wobei ihr jeweils das deutsche Wort vom Versuchsleiter vorgelegt wird, und sie muß das zugehörige englische Wort sagen. Der gesamte Zyklus wird dreimal durchgeführt. Bei jedem Durchgang wird notiert, ob die Versuchsperson beim Test einen Fehler gemacht hat (F) oder nicht (R). Ein Ergebnis des Experiments könnte dann sein: „FFR“. Als Ergebnisraum kann die Menge Ω = {FFF, FFR, FRF, FRR, RFF, RFR, RRF, RRR } definiert werden. Ein Ereignis ist dann etwa „nach dem ersten Versuch keinen Fehler mehr“: A = {FRR, RRR}

Nun soll angenommen werden, daß es für jedes Elementarereignis $\{\omega\}$ dieses Experiments eine Wahrscheinlichkeit $P\big(\{\omega\}\big)$ gibt. Eine mögliche Frage ist dann etwa die nach der Wahrscheinlichkeit eines Fehlers in einem bestimmten Durchgang, da man erwartet, daß sich diese Wahrscheinlichkeit im Laufe des Experiments ändert. Darüber hinaus kann die Wahrscheinlichkeit eines Fehlers aber auch davon abhängen, was im vorhergehenden Durchgang geschah, denn wenn die Versuchsperson im zweiten Durchgang alle Vokabeln gelernt hat, dann wird im dritten ihre Fehlerwahrscheinlichkeit anders sein, als wenn sie im zweiten Durchgang auch noch Fehler hatte. Wenn nach dem zweiten Durchgang die Wahrscheinlichkeit eines Fehlers im dritten Durchgang angegeben werden soll, dann kann die Kenntnis der Antworten aus dem zweiten Durchgang diese Wahrscheinlichkeiten erheblich verändern. Das Konzept der „bedingten Wahrscheinlichkeit“ dient dazu, den Einfluß von Vorinformation auf die Bestimmung von Wahrscheinlichkeiten noch ausstehender Ereignisse zu untersuchen. Als Vorinformation dient dabei die Kenntnis von bereits eingetretenen Ereignissen.

Definition 1.20. Seien A und B Ereignisse, wobei $P(B) > 0$. Dann ist die *bedingte Wahrscheinlichkeit* $P(A \mid B)$ *von* A *unter der Bedingung* B definiert durch

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

1.4.1 Stochastische Unabhängigkeit

Die bedingte Wahrscheinlichkeit $P(A \mid B)$ gibt somit die Wahrscheinlichkeit für das Ereignis A in den Fällen an, in denen B bereits eingetreten ist. Nicht immer ist jedoch die Kenntnis bereits eingetretener Ereignisse hilfreich für die Vorhersage zukünftiger Ereignisse:

Definition 1.21. Zwei Ereignisse A und B heißen genau dann *stochastisch unabhängig*, wenn $P(A \mid B) = P(A)$. Ist der Zusammenhang klar, dann werden die beiden Ereignisse statt *stochastisch unabhängig* häufig einfach *unabhängig* genannt.

Bei unabhängigen Ereignissen A und B führt das Wissen, daß B eingetreten ist, nicht zu einer Veränderung der Wahrscheinlichkeit von A. Die Wahrscheinlichkeit, daß A und B gleichzeitig eintreten, ist dann einfach $P(A \cap B) = P(A)P(B)$ oder äquivalent dazu $P(A \mid B) = P(A \mid \overline{B})$. Die Definition der Unabhängigkeit kann auf Familien von Ereignissen ausgedehnt werden.

Definition 1.22. Sei $\mathfrak{C}$ eine Familie von Ereignissen (also eine Menge von Ereignissen). Die Familie $\mathfrak{C}$ ist eine *Familie unabhängiger Ereignisse*, wenn für alle endlichen Teilfamilien $\{A_1, \ldots, A_n\}$ von $\mathfrak{C}$ gilt

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).$$

Paarweise Unabhängigkeit (Def. 1.21) in einer Familie von Ereignissen hat nicht automatisch auch die Unabhängigkeit der gesamten Familie zur Folge. Dies zeigt Übungsaufgabe 1 am Ende dieses Abschnitts.

Ein weit verbreitetes Mißverständnis des Konzepts der stochastischen Unabhängigkeit betrifft den Zusammenhang zwischen „disjunkt“ und „unabhängig“. Disjunkte Ereignisse sind *nicht* unabhängig. Im Gegenteil, disjunkte Ereignisse können nicht unabhängig sein, denn aus der Disjunktheit folgt, daß das zweite Ereignis nicht eintreten kann, wenn das erste eingetreten ist. Wären diese Ereignisse unabhängig, dann müßte gelten

$$P(A \cap B) = P(\varnothing) = 0 = P(A)P(B),$$

eines der Ereignisse müßte daher die Wahrscheinlichkeit 0 haben. „Disjunkt“ ist also gleichbedeutend mit „unverträglich“.

1.4.2 Die Formel von Bayes

Will man von der bedingten Wahrscheinlichkeit eines Ereignisses zur unbedingten Wahrscheinlichkeit übergehen, dann kann dies durch die Betrachtung aller möglichen Bedingungen erreicht werden, unter denen das entsprechende Ereignis eintreten kann. Seien B_i Elemente einer Zerlegung von Ω, also eine Reihe paarweise disjunkter Ereignisse: $B_i \cap B_j = \varnothing$, falls $i \neq j$, mit $i, j = 1, \ldots, n$, wobei die Vereinigung aller B_i gleich Ω sei:

$$\Omega = \bigcup_{i=1}^{n} B_i.$$

Dann gilt

$$A \cap \Omega = A \cap \left(\bigcup_{i=1}^{n} B_i\right)$$

und wegen dem Distributivgesetz auch

$$A = \bigcup_{i=1}^{n} A \cap B_i.$$

Wegen der Additivität der Wahrscheinlichkeit, und da alle B_i disjunkt sind, gilt dann für die Wahrscheinlichkeit von A

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A \cap B_i) \\ &= \sum_{i=1}^{n} P(A \mid B_i) P(B_i). \end{aligned}$$

Diese Gleichung wird *Formel der totalen Wahrscheinlichkeit* genannt. Sie dient zusammen mit der Definition der bedingten Wahrscheinlichkeit zur Herleitung der „Formel von Bayes“. Aufgrund der Definition der bedingten Wahrscheinlichkeit gilt nämlich:

$$\begin{aligned} P(A \mid B) P(B) &= P(A \cap B) \\ &= P(B \mid A) P(A). \end{aligned}$$

Durch Umformen erhält man daraus

$$P(B \mid A) = \frac{P(A \mid B) P(B)}{P(A)}.$$

Wird hier das Ereignis B als „Ursache“ des beobachtbaren Ereignisses A betrachtet und kennt man die Wahrscheinlichkeit $P(A \mid B)$, dann kann damit die Wahrscheinlichkeit der „Ursache“ B bei gegebener Beobachtung A berechnet werden.

Setzt man in Gleichung (1.1) für $P(A)$ die Formel der totalen Wahrscheinlichkeit ein, dann erhält man die Formel von Bayes:

$$P(B_j \mid A) = \frac{P(A \mid B_j) P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i) P(B_i)}. \tag{1.1}$$

Die Formel von Bayes kann zur adaptiven Schätzung der Wahrscheinlichkeit von Hypothesen benutzt werden. Sei A ein Ereignis, das unter verschiedenen Bedingungen eintritt, über die es die Hypothesen $B_1, \ldots, B_n$ gibt und deren a priori Wahrscheinlichkeiten $P(B_1), \ldots, P(B_n)$ bekannt seien. Ferner seien die Wahrscheinlichkeiten des Auftretens von A bei Gültigkeit der Hypothesen B_i bekannt, also die bedingten Wahrscheinlichkeiten $P(A \mid B_i)$. Tritt nun bei einem Zufallsexperiment das Ereignis A ein, dann lassen sich mit Hilfe der Formel von Bayes die Wahrscheinlichkeiten der Hypothesen B_i neu berechnen.

1.4.3 Ereignisfolgen

Mit Hilfe eines Induktionsbeweises läßt sich folgende Beziehung zeigen:

$$P(A_1, \ldots, A_n) = P(A_1) P(A_2 \mid A_1) P(A_3 \mid A_1, A_2) \cdots P(A_n \mid A_1, \ldots, A_{n-1}).$$

In dieser Formel steht $P(A_1, \ldots, A_n)$ für $P(\bigcap_{i=1}^{n} A_i)$. Betrachtet man die A_i als eine in der Zeit auftretende Folge von Ereignissen, dann können mit dieser Formel spezielle stochastische Prozesse untersucht werden. So spricht man etwa von einem *Markov*-Prozess, wenn $P(A_i \mid A_1, \ldots, A_{i-1}) = P(A_i \mid A_{i-1})$. Jedes Ereignis A_i hängt in einem Markov-Prozess nur vom direkt vorausgehenden, nicht aber von früheren Ereignissen ab. Markov-Prozesse bilden die Grundlage mehrerer stochastischer Lernmodelle im Rahmen der Reiz-Stichproben-Theorie (Wickens, 1982).

1.4.4 Bedingte Unabhängigkeit

Definition 1.23. Sei C ein Ereignis aus dem Wahrscheinlichkeitsraum $\langle \Omega, \mathfrak{A}, P \rangle$ mit positiver Wahrscheinlichkeit. Wir definieren eine Abbildung von der σ-Algebra $\mathfrak{A}$ in $\mathbb{R}$ durch

$$P_C(A) = P(A \,|\, C).$$

Dann ist $\langle \Omega, \mathfrak{A}, P_C \rangle$ ein Wahrscheinlichkeitsraum. Zwei Ereignisse A und B, die bezüglich der so definierten Wahrscheinlichkeit P_C unabhängig sind, heißen *bedingt unabhängig bezüglich* C.

Für Ereignisse A und B, die bezüglich C bedingt unabhängig sind, gilt nach dieser Definition

$$P(A \cap B \,|\, C) = P(A \,|\, C) P(B \,|\, C).$$

Das Konzept der bedingten Unabhängigkeit ist eine wesentliche Grundlage der psychologischen Testtheorie. Dort gibt es Modelle, die annehmen, daß die richtigen Lösungen zweier Testaufgaben durch eine Person bezüglich dieser einen Person bedingt unabhängige Ereignisse sind. Diese Eigenschaft wird dort „lokal stochastische Unabhängigkeit“ genannt.

Übungsaufgaben

1. [Aus Brémaud (1988, S. 14)] Sei $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ ein Ergebnisraum und $\mathfrak{A}$ die Menge aller Teilmengen von Ω. Auf $\mathfrak{A}$ sei die Wahrscheinlichkeitsfunktion P definiert durch $P\big(\{\omega_i\}\big) = \frac{1}{4}$ für $i = 1,\ldots, 4$. Die Ereignisse A, B und C seien folgendermaßen definiert: $A = \{\omega_1, \omega_2\}$, $B = \{\omega_2, \omega_3\}$ und $C = \{\omega_1, \omega_3\}$. Zeigen Sie daß $\mathfrak{C} = \{A, B, C\}$ keine Familie unabhängiger Ereignisse ist, obwohl A, B und C paarweise unabhängig sind.

1.5 Zufallsvariablen

Die Analyse von Zufallsexperimenten mit Hilfe von Mengen und Wahrscheinlichkeiten ist verhältnismäßig umständlich, da für allgemeine Mengen nur sehr einfache Operationen zur Verfügung stehen, die das Theoretisieren erschweren. Man versucht daher Regeln zu finden, nach denen die Elemente des Ergebnisraums durch Zahlen dargestellt werden können. Diese Zuordnung von Zahlen zu Ergebnissen soll es ermöglichen, die strukturellen Eigenschaften des Wahrscheinlichkeitsraums numerisch zu analysieren. Die wesentliche Restriktion bei der Zuordnung von Zahlen zu Ergebnissen ist, daß den Mengen, die im Bereich der Zahlen gebildet werden können, Mengen von Ergebnissen—Ereignisse—entsprechen. Einer Menge von Zahlen entspricht dann das Ereignis, das alle Ergebnisse enthält, deren Zahlenwert in der Zahlenmenge enthalten ist. Aus technischen Gründen wird hier als Bildmenge $\overline{\mathbb{R}}$ üblicherweise die um $\{-\infty, +\infty\}$ erweiterte Menge der reellen Zahlen $\mathbb{R}$ benutzt: $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$.

1.5.1 Reelle Zufallsvariablen

Definition 1.24. Sei $\langle \Omega, \mathfrak{A}, P \rangle$ ein Wahrscheinlichkeitsraum. Eine Abbildung X von Ω in $\overline{\mathbb{R}}$ ist eine *Zufallsvariable* genau dann, wenn für jede reelle Zahl x die Menge $\{\omega \,|\, X(\omega) \leq x\}$ in $\mathfrak{A}$ ist.

Die Bedingung, daß unter einer Zufallsvariablen X das Urbild eines jeden Intervalls der Form $(-\infty, x] = \{y \,|\, y \leq x\}$ ein Ereignis sein muß, nennt man auch $\mathfrak{A}$-*Meßbarkeit* der Funktion X. Sie ermöglicht es, die Wahrscheinlichkeit P von der σ-Algebra $\mathfrak{A}$ auf den Wertebereich der Zufallsvariablen X zu übertragen.

Definition 1.25. Sei X eine Zufallsvariable. Die Funktion

$$F(x) = P\big(\{\omega \,|\, X(\omega) \leq x\}\big)$$

heißt *Verteilungsfunktion* der Zufallsvariablen X.

Statt $P\big(\{\omega \,|\, X(\omega) \leq x\}\big)$ wird auch häufig $P(X \leq x)$ geschrieben. Verteilungsfunktionen haben folgende allgemeine Eigenschaften:

1. Sie sind monoton steigend: wenn $x \leq x'$, dann ist $F(x) \leq F(x')$.
2. Für kleine Werte von x nähert sich $F(x)$ an den Wert 0 an:
$$\lim_{x \to -\infty} F(x) = 0.$$
3. Für große Werte von x nähert sich $F(x)$ an den Wert 1 an:
$$\lim_{x \to \infty} F(x) = 1.$$

Wegen der letzten beiden Bedingungen kann der Definitionsbereich einer Verteilungsfunktion von $\mathbb{R}$ auf $\overline{\mathbb{R}}$ ausgedehnt werden. Man setzt $F(-\infty) = 0$ und $F(\infty) = 1$.

Definition 1.26. Zufallsvariablen, die nur Werte aus $\mathbb{R}$ annehmen, werden *reelle Zufallsvariablen* genannt. Falls eine reelle Zufallsvariable eine Verteilungsfunktion F besitzt, für die eine nichtnegative Funktion f existiert, so daß gilt

$$F(x) = \int_{-\infty}^{x} f(y)\, dy,$$

dann sagt man, daß X *eine Wahrscheinlichkeitsdichte besitzt*. Die Funktion f wird *Wahrscheinlichkeitsdichte* genannt.

Es läßt sich leicht zeigen, daß aus der Definition 1.26 einer Dichtefunktion folgt:

$$P(a \leq X \leq b) = \int_{a}^{b} f(y)\, dy$$

und

$$\int_{-\infty}^{+\infty} f(y)\, dy = 1.$$

1.5.2 Diskrete Zufallselemente

Definition 1.27. Sei E eine abzählbare Menge und $\langle \Omega, \mathfrak{A}, P \rangle$ ein Wahrscheinlichkeitsraum. Eine Abbildung X von Ω in die Menge E, so daß für alle x in E die Menge $\{\omega \,|\, X(\omega) = x\}$ in $\mathfrak{A}$ ist, heißt *diskretes Zufallselement* von E. Ist E eine Teilmenge der reellen Zahlen $\mathbb{R}$, dann wird X auch *diskrete Zufallsvariable* genannt.

Auf dem Wertebereich E eines diskreten Zufallselements kann die Funktion

$$\begin{aligned} p(x) &= P\big(\{\omega \,|\, X(\omega) = x\}\big) \\ &= P(X = x) \end{aligned}$$

definiert werden, sie wird *Wahrscheinlichkeitsfunktion* des diskreten Zufallselements X genannt. Für eine Teilmenge A von E gilt damit

$$\begin{aligned} P(X \in A) &= P\big(\{\omega \,|\, X(\omega) \in A\}\big) \\ &= \sum_{x \in A} p(x). \end{aligned}$$

Für die Verteilungsfunktion $F(x)$ einer diskreten Zufalls*variablen* X gilt

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \sum_{y \leq x} p(y). \end{aligned}$$

In der Regel werden wir nur solche diskreten Zufallselemente betrachten, deren Wertebereich eine Teilmenge von $\mathbb{R}$ ist, die also diskrete Zufallsvariablen sind. Diskrete Zufallselemente mit anderen Wertebereichen werden in der psychologischen Testtheorie benutzt, um die zufällige Auswahl einer Testperson aus einer Population zu beschreiben.

1.5.3 Indikatorfunktionen

Sei $\langle \Omega, \mathfrak{A}, P \rangle$ ein Wahrscheinlichkeitsraum und A ein Ereignis daraus. Eine Indikatorfunktion 1_A ist eine Funktion von Ω in $\{0, 1\}$, definiert durch

$$1_A(\omega) = \begin{cases} 1 & \text{falls } \omega \in A \\ 0 & \text{falls } \omega \notin A \end{cases} \tag{1.2}$$

Damit ist $X = 1_A$ eine diskrete Zufallsvariable mit dem Wertebereich $E = \{0, 1\}$, die Ereignisse $\{\omega \mid X(\omega) = 1\}$ und A sind identisch. Es ist also $P(X = 1) = P(A)$ und $P(X = 0) = 1 - P(A)$.

1.5.4 Unabhängige Zufallsvariablen

Definition 1.28. Reelle Zufallsvariablen $X_1, \ldots, X_n$ werden als *(stochastisch) unabhängig* bezeichnet, wenn für alle $x_1, \ldots, x_n$ aus $\mathbb{R}$ gilt

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n).$$

Die Schreibweise $P(X_1 \leq x_1, \ldots, X_n \leq x_n)$ ist eine Abkürzung für

$$P\left(\bigcap_{i=1}^{n} \{X_i \leq x_i\}\right).$$

Falls die Zufallsvariablen X_i Dichten f_i besitzen, dann gilt

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f_1(y_1) \ldots f_n(y_n)\, dy_1 \ldots dy_n.$$

Die Dichte von $(X_1, \ldots, X_n)$ ist also das Produkt der Dichten aller X_i:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i).$$

Diskrete Zufallselemente X und Y sind genau dann unabhängig, wenn für alle x aus dem Wertebereich von X und alle y aus dem Wertebereich von Y gilt

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Das Konzept der Unabhängigkeit kann auch auf Familien von Zufallsvariablen ausgedehnt werden: Man nennt eine Familie $\mathfrak{H}$ von Zufallsvariablen unabhängig, wenn für jede endliche Teilfamilie $\{X_1, \ldots, X_n\} \subset \mathfrak{H}$ die Bedingung

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i)$$

für alle reellen Zahlen x_i, $1 \leq i \leq n$ gilt.

1.5.5 Zufallsstichproben

Eine Zufallsstichprobe erhält man, wenn in einer Population eine Folge von Zufallsexperimenten durchgeführt wird. Eine einzelne Beobachtung besteht im Registrieren eines Ergebnisses und des Wertes der damit verbundenen Zufallsvariablen. Wesentlich dabei ist, daß jedes Element der Grundgesamtheit die gleiche Wahrscheinlichkeit hat, beobachtet zu werden. Die einzelnen Beobachtungen müssen stochastisch unabhängige Ereignisse darstellen. Die Wiederholungen stellen also unabhängige Wiederholungen eines einzigen Zufallsexperiments dar.

DEFINITION 1.29. Eine *Zufallsstichprobe* vom Umfang n ist eine Folge X_1, $\ldots$, X_n stochastisch unabhängiger und identisch verteilter Zufallsvariablen.

Für alle Zufallsvariablen X_i einer Zufallsstichprobe $X_1, \ldots, X_i, \ldots, X_n$ gibt es daher eine bestimmte Verteilungsfunktion F, so daß

$$P(X_i \leq x) = F(x).$$

Außerdem sind alle Zufallsvariablen X_i und X_j der Zufallsstichprobe mit $i \neq j$ stochastisch unabhängig.

1.6 Verteilungsparameter

1.6.1 Modus, Median, Quantile einer Zufallsvariablen

DEFINITION 1.30. Ist X eine reelle Zufallsvariable mit der Wahrscheinlichkeitsdichte $f(x)$ und der Verteilungsfunktion $F(x)$, dann ist

1. jeder Wert x_{Mod}, an dem $f(x)$ maximal ist, ein *Modus* oder *Modalwert* von X,
2. jeder Wert x_π mit $F(x_\pi) = \pi$ ein π-*Quantil* von X.

Ist X ein Zufallselement mit einem abzählbaren Wertebereich E und einer auf E definierten Wahrscheinlichkeitsfunktion $p(x)$, dann ist

1. jeder Wert x_{Mod}, an dem $p(x)$ maximal ist, ein *Modus* von X,
2. jeder Wert x_π mit

$$p(X \leq x_\pi) = \pi \text{ und } p(X \geq x_\pi) = 1 - \pi$$

ein π-*Quantil* von X.

Das 0.5-Quantil von X heißt *Median* von X.

1.6.2 Erwartungswert und Varianz

DEFINITION 1.31. Ist X eine reelle Zufallsvariable mit der Wahrscheinlichkeitsdichte $f(x)$, dann ist, falls der Ausdruck existiert,

$$\mathscr{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

der *Erwartungswert von* X und

$$\begin{aligned} \sigma^2(X) &= \mathscr{E}[(X - \mathscr{E}(X))^2] \\ &= \int_{-\infty}^{\infty} [x - \mathscr{E}(X)]^2 f(x)\, dx \end{aligned}$$

die *Varianz von* X. Die *Standardabweichung* von X ist $\sigma(X)$, die positive Wurzel aus der Varianz.

Definition 1.32. Sei X ein Zufallselement mit einem abzählbaren Wertebereich E und einer auf E definierten Wahrscheinlichkeitsfunktion $p(x)$. Ferner sei h eine ebenfalls auf E definierte, reellwertige Funktion, so daß

$$\sum_{x\in E} |h(x)|p(x) < \infty. \tag{1.3}$$

Dann ist

$$\mathscr{E}[h(X)] = \sum_{x\in E} h(x)p(x)$$

der Erwartungswert von $h(X)$. *Varianz* und *Standardabweichung* von X sind definiert wie in Definition 1.31.

Bei diskreten Zufallsvariablen, also diskreten Zufallselementen, deren Wertebereich eine Teilmenge von $\mathbb{R}$ ist, betrachtet man häufig den Erwartungswert von $h(x) = x$, der identischen Abbildung.

Beispiel. Wir betrachten einen Wurf mit einer idealen Münze. Dem Ergebnis $\omega_1 =$ „Kopf“ wird die Zahl $X(\omega_1) = 1$, dem Ergebnis $\omega_2 =$ „Zahl“ die Zahl $X(\omega_2) = 0$ zugeordnet. Die Wahrscheinlichkeiten seien $P\big(\{\omega_i\}\big) = 0.5$ für $i = 1, 2$. Die Wahrscheinlichkeitsfunktion von X ist dann durch $p(0) = 0.5$ und $p(1) = 0.5$ definiert. Der Erwartungswert von X ist

$$\mathscr{E}(X) = p(0)0 + p(1)1 = p(1) = 0.5.$$

Die Varianz von X ist

$$\sigma^2(X) = p(0)(0 - 0.5)^2 + p(1)(1 - 0.5)^2 = 0.25.$$

Hier einige Regeln für das Rechnen mit Erwartungswerten. Dabei werden a und b als konstante Zahlen betrachtet und X und Y als Zufallsvariablen.

1. Der Erwartungswert einer Konstanten ist die Konstante selbst:

$$\mathscr{E}(a) = a.$$

2. Der Erwartungwert ist ein linearer Operator:

$$\mathscr{E}(aX + bY) = a\,\mathscr{E}(X) + b\,\mathscr{E}(Y). \tag{1.4}$$

Für die Varianz gelten folgende Regeln, wobei a wiederum eine konstante Zahl sein soll:

1.

$$\begin{aligned} \sigma^2(a) &= 0; && (1.5)\\ \sigma^2(X + a) &= \sigma^2(X); && (1.6)\\ \sigma^2(aX) &= a^2\sigma^2(X). && (1.7) \end{aligned}$$

2. Zum Berechnen der Varianz einer Zufallsvariablen benutzt man die Formel

$$\sigma^2(X) = \mathscr{E}(X^2) - [\mathscr{E}(X)]^2. \tag{1.8}$$

Sie läßt sich folgendermaßen ableiten:

$$\begin{aligned} \sigma^2(X) &= \mathscr{E}[(X - \mathscr{E}(X))^2]\\ &= \mathscr{E}[X^2 - 2X\mathscr{E}(X) + (\mathscr{E}(X))^2]\\ &= \mathscr{E}(X^2) - \mathscr{E}[2X\mathscr{E}(X)] + \mathscr{E}[(\mathscr{E}(X))^2]\\ &= \mathscr{E}(X^2) - 2\mathscr{E}(X)\mathscr{E}(X) + [\mathscr{E}(X)]^2\\ &= \mathscr{E}(X^2) - [\mathscr{E}(X)]^2. \end{aligned}$$

Der Schritt von $\mathscr{E}[2X\mathscr{E}(X)]$ nach $2\mathscr{E}(X)\mathscr{E}(X)$ ist berechtigt, weil im ersten Ausdruck $\mathscr{E}(X)$ eine Konstante ist und daher vor den Erwartungswertoperator gezogen werden darf.

1.6.3 Kovarianz und Korrelationskoeffizient

DEFINITION 1.33. Seien X und Y zwei Zufallsvariablen. Der Erwartungswert

$$\sigma(X,Y) = \mathscr{E}[(X - \mathscr{E}(X))(Y - \mathscr{E}(Y))]$$

heißt *Kovarianz von X und Y.*

Es läßt sich zeigen, daß

$$\begin{aligned}\sigma(X,Y) &= \mathscr{E}[(X - \mathscr{E}(X))(Y - \mathscr{E}(Y))] \\ &= \mathscr{E}(XY) - \mathscr{E}(X)\mathscr{E}(Y).\end{aligned}$$

DEFINITION 1.34. Sind X und Y Zufallsvariablen, dann heißt

$$\rho(X,Y) = \frac{\sigma(X,Y)}{\sigma(X)\,\sigma(Y)}$$

der *Korrelationskoeffizient* von X und Y.

Für 3 Zufallsvariablen X, Y, Z und reelle Konstanten a, b gelten folgende Regeln:

1. $$\sigma^2(aX + bY) = a^2\sigma^2(X) + b^2\sigma^2(Y) + 2ab\,\sigma(X,Y). \tag{1.9}$$

   Diese Beziehung ergibt sich mit Gl. (1.8):

   $$\begin{aligned}\sigma^2(aX + bY) &= \mathscr{E}[(aX + bY)^2] - [\mathscr{E}(aX + bY)]^2 \\ &= \mathscr{E}[(aX)^2 + 2aXbY + (bY)^2] - [\mathscr{E}(aX) + \mathscr{E}(bY)]^2 \\ &= \mathscr{E}[(aX)^2] - [\mathscr{E}(aX)]^2 + \mathscr{E}[(bY)^2] - [\mathscr{E}(bY)]^2 \\ &\quad +2\mathscr{E}[aXbY] - 2\mathscr{E}(aX)\mathscr{E}(bY) \\ &= \sigma^2(aX) + \sigma^2(bY) + 2\sigma(aX, bY) \\ &= a^2\sigma^2(X) + b^2\sigma^2(Y) + 2ab\,\sigma(X,Y).\end{aligned}$$

   Analog hierzu läßt sich zeigen, daß

   $$\sigma^2(X - Y) = \sigma^2(X) + \sigma^2(Y) - 2ab\,\sigma(X,Y). \tag{1.10}$$

2. Sind X und Y stochastisch unabhängig, dann gilt

   $$\mathscr{E}(XY) = \mathscr{E}(X)\mathscr{E}(Y).$$

3. Sind X und Y stochastisch unabhängig, dann ist die Varianz $\sigma^2(X + Y)$ der Summe gleich der Summe der Varianzen: $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y)$. Dies gilt nicht im allgemeinen Fall, wenn $\sigma(X,Y) \neq 0$.

4. $\sigma(X,Y) = \sigma(Y,X)$.

5. Sind $X_1, \ldots, X_n$ beliebige Zufallsvariablen, dann gilt für die Varianz der Summe:

   $$\sigma^2\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma(X_i, X_j). \tag{1.11}$$

6. Für die Kovarianz der Summe $X + Y$ und der Variablen Z gilt

   $$\begin{aligned}\sigma(X + Y, Z) &= \mathscr{E}[(X + Y)Z] - \mathscr{E}(X + Y)\mathscr{E}(Z) \\ &= \mathscr{E}(XZ + YZ) - [\mathscr{E}(X)\mathscr{E}(Z) + \mathscr{E}(Y)\mathscr{E}(Z)] \\ &= \sigma(X,Z) + \sigma(Y,Z)\end{aligned} \tag{1.12}$$

1.7 Das Gesetz der großen Zahlen

Gnedenko (1968) schreibt:

> Die umfangreichen von der Menschheit angesammelten Erfahrungen zeigen, daß Erscheinungen, die eine Wahrscheinlichkeit nahe Eins besitzen, fast immer eintreten. [...] Dieser Tatbestand spielt für viele praktische Schlußfolgerungen aus der Wahrscheinlichkeitsrechnung eine grundlegende Rolle, da die erwähnte *Erfahrungstatsache* es in der Praxis gestattet, wenig wahrscheinliche Ereignisse für *praktisch unmöglich* und Ereignisse mit Wahscheinlichkeit nahe Eins für *praktisch sicher* anzunehmen. Doch kann man auf die ganz natürliche Frage, wie groß eine Wahrscheinlichkeit sein muß, damit man ein Ereignis für praktisch unmöglich halten kann, keine eindeutige Antwort geben. Das ist auch verständlich, da man im praktischen Leben die Wichtigkeit der Ereignisse berücksichtigen muß, mit denen man es zu tun bekommt (Gnedenko, 1968, S 185).

Das Gesetz der großen Zahlen besteht aus Aussagen über Ereignisse, die mit einer Wahrscheinlichkeit von 1, also „fast sicher“, eintreten. Ein Beispiel für eine solche Aussage ist folgende: Seien X_i, $i = 1, \ldots, n$ Zufallsvariablen und

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

deren arithmetisches Mittel. Wenn es eine reelle Konstante $a > 0$ gibt, so daß für beliebige $\varepsilon > 0$

$$\lim_{n \to \infty} P(|\overline{X} - a| \geq \varepsilon) = 0,$$

dann sagt man, *die Zufallsgröße* $\overline{X}$ *konvergiert in Wahrscheinlichkeit gegen* a.

1.7.1 Die Tschebyschewsche Ungleichung

Sei X ein diskretes Zufallselement mit dem Wertebereich E und h eine reelle Funktion auf E, die die Gleichung (1.3) erfüllt. Dann gilt für $a > 0$ die *Markov*sche Ungleichung:

$$P(|\, h(X)\,| \geq a) \quad \leq \quad \frac{\mathscr{E}\,[|\, h(x)\,|]}{a} \tag{1.13}$$

Begründung. Sei $C = \{x \mid |h(x)| \geq a\}$ dann ist

$$\begin{aligned} |h(X)| &= 1_C(x)|h(x)| + 1_{\overline{C}}(x)|h(x)| \\ &\geq 1_C(x)|h(x)| \\ &\geq 1_C(x)a, \end{aligned}$$

da falls $x \in C$ wegen der Definition von C gilt $|\, h(x)\,| \geq a$. Es ergibt sich

$$\begin{aligned} \mathscr{E}[|h(X)|] \geq \mathscr{E}[a\, 1_C(X)] &= a\, \mathscr{E}[1_C(X)] \\ &= a\, P(X \in C) \\ &= a\, P(|h(X)| \geq a), \end{aligned}$$

da $X \in C$ gdw. $|\, h(X)\,| \geq a$. □

Setzt man in der Markovschen Ungleichung (1.13) $h(X) = [X - \mathscr{E}(X)]^2$ und $a = \varepsilon^2$ mit $\varepsilon > 0$, so erhält man die in der psychologischen Testtheorie häufig benutzte Tschebyschewsche Ungleichung:

$$P(|X - \mathscr{E}(X)| \geq \varepsilon) \leq \frac{\sigma^2(X)}{\varepsilon^2}. \tag{1.14}$$

Es läßt sich leicht zeigen, daß sowohl die Markovsche als auch die Tschebyschewsche Ungleichung auch für stetige Zufallsvariablen gelten.

1.7.2 Der zentrale Grenzwertsatz

Der zentrale Grenzwertsatz ist einer der wichtigsten Sätze der Wahrscheinlichkeitstheorie. Er macht eine Aussage über die Summe beliebig verteilter, unabhängiger Zufallsvariablen: Sind X_i, $i = 1, \ldots, n$ unabhängig verteilte Zufallsvariablen, und ist

$$S_n = \sum_{i=1}^{n} X_i$$

deren Summe, dann ist die Zufallsvariable S_n approximativ normalverteilt mit dem Erwartungswert $\mathscr{E}(S_n)$ und der Varianz $\sigma^2(S_n)$.

Übungsaufgaben

1. Was ist der Erwartungswert einer Indikatorfunktion?
2. Zeigen Sie folgende Eigenschaften von Indikatorfunktionen:
   (a) $1_{A\cap B} = \min(1_A, 1_B) = 1_A \cdot 1_B$;
   (b) $1_{A\cup B} = \max(1_A, 1_B)$;
   (c) $1_{\overline{A}} = 1 - 1_A$, da $1_\Omega = 1$;

1.8 Mehrdimensionale Zufallsvariablen

1.8.1 Gemeinsame Wahrscheinlichkeitsfunktionen und Randwahrscheinlichkeiten

Sei $\langle \Omega, \mathfrak{A}, P\rangle$ ein Wahrscheinlichkeitsraum und X und Y Zufallselemente auf Ω mit den Wertebereichen E_X und E_Y. Die auf $E_X \times E_Y$ definierte Funktion

$$\begin{aligned} p(x,y) &= P(X = x, Y = y) \\ &= P\big(\{\omega \mid X(\omega) = x, Y(\omega) = y\}\big) \end{aligned}$$

ist die *gemeinsame Wahrscheinlichkeitsfunktion* des Zufallsvektors (X, Y). Um die Wahrscheinlichkeitsfunktion von X alleine zu erhalten, braucht man nur die gemeinsame Wahrscheinlichkeitsfunktion $p(x, y)$ über alle möglichen Werte von Y zu summieren:

$$P(X = x) = \sum_{y\in E_Y} p(x,y) = p(x,*).$$

Man nennt dann $p(x,*)$ die *Randwahrscheinlichkeitsfunktion* von X.

Im Fall stetiger Zufallsvariablen X und Y sagt man, daß diese eine *gemeinsame* Wahrscheinlichkeitsdichte f besitzen, wenn es eine Funktion f gibt, so daß

$$P(X \le x, Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,du\,dv$$

für alle reellen Zahlen x und y. Für die *Randwahrscheinlichkeiten* $P(X \le x)$ gilt dann

$$P(X \le x) = \int_{-\infty}^{x} f(u,*)\,du$$

mit

$$f(u,*) = \int_{-\infty}^{\infty} f(u,v)\,dv,$$

der *Randwahrscheinlichkeitsdichte* von X. Analoge Vereinbarungen können natürlich für Y getroffen werden.

1.8.2 Bedingte Wahrscheinlichkeitsfunktionen und bedingte Erwartungswerte

Seien X und Y diskrete Zufallselemente mit den Wertebereichen E_X und E_Y und der gemeinsamen Wahrscheinlichkeitsfunktion $f(x,y)$. $f(x)$ und $f(y)$ seien die Randwahrscheinlichkeitsfunktionen von X und Y. Sei

$$A = \{(x,y) \,|\, x = x', -\infty < y < \infty\}$$

und

$$B = \{(x,y) \,|\, -\infty < x < \infty,\, y = y'\},$$

wobei x' und y' so gewählt seien, daß

$$P(A) = P(X = x') = f(x') > 0.$$

Dann ist

$$\begin{aligned} P(B \,|\, A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{P(X = x', Y = y')}{P(X = x')} \\ &= \frac{f(x',y')}{f(x')}. \end{aligned}$$

Die Wahrscheinlichkeit, daß $Y = y'$ unter der Bedingung, daß $X = x'$, ist also gleich dem Quotienten $f(x',y')/f(x')$.

Definition 1.35. Ist $f(x) > 0$, dann ist

$$f(y \,|\, x) = \frac{f(x,y)}{f(x)}$$

die *bedingte Wahrscheinlichkeitsfunktion des diskreten Zufallselements Y, gegeben für das diskrete Zufallselement X gilt $X = x$.* Entsprechend ist

$$f(x \,|\, y) = \frac{f(x,y)}{f(y)}$$

für $f(y) > 0$ die *bedingte Wahrscheinlichkeit des diskreten Zufallselements X gegeben $Y = y$.*

Für stetige Zufallsvariablen X und Y, die eine gemeinsame Wahrscheinlichkeitsdichte $f(x,y)$ und Randwahrscheinlichkeitsdichten $f(x)$ und $f(y)$ besitzen, werden analog zum diskreten Fall die bedingten Wahrscheinlichkeitsdichten

$$f(y \,|\, x) = \frac{f(x,y)}{f(x)}$$

für $f(x) > 0$ und

$$f(x \,|\, y) = \frac{f(x,y)}{f(y)}$$

für $f(y) > 0$ definiert. Die Funktionen $f(y \,|\, x)$ und $f(x \,|\, y)$ sind nichtnegativ, und es gilt

$$\begin{aligned} \int_{-\infty}^{\infty} f(y \,|\, x)\, dy &= \int_{-\infty}^{\infty} \frac{f(x,y)}{f(x)} dy \\ &= \frac{1}{f(x)} \int_{-\infty}^{\infty} f(x,y)\, dy \\ &= \frac{1}{f(x)} f(x) \\ &= 1. \end{aligned}$$

Damit ist gezeigt, daß $f(y \,|\, x)$ die Eigenschaften einer Wahrscheinlichkeitsdichte besitzt. Es können daher auch Wahrscheinlichkeiten und Erwartungswerte mit Hilfe von $f(y \,|\, x)$ berechnet werden. Wir erhalten etwa

$$P(Y < y \,|\, X = x) = \int_{-\infty}^{y} f(u \,|\, x)\, du.$$

Der Erwartungswert

$$\mathscr{E}(Y \mid x) = \int_{-\infty}^{\infty} y\, f(y \mid x)\, dy$$

wird als *bedingter Erwartungswert von Y, gegeben $X = x$* bezeichnet.

Im Fall diskreter Zufallsvariablen werden statt Integrationen einfach Summationen benutzt: Sind X und Y diskrete Zufallselemente mit den Wertebereichen E_X und E_Y, und ist h eine reellwertige Funktion auf E_Y, dann ist

$$\mathscr{E}(h(Y) \mid X = x) = \sum_{y \in E_Y} h(y) P(Y = y \mid X = x)$$

der bedingte Erwartungswert von $h(Y)$, gegeben $X = x$. Wegen der Formel der totalen Wahrscheinlichkeit gilt

$$P(Y = y) = \sum_{x \in E_X} P(X = x) P(Y = y \mid X = x).$$

Multipliziert man hier beide Seiten mit $h(y)$ und summiert über E_Y, erhält man

$$\sum_{y \in E_Y} h(y) P(Y = y) = \sum_{y \in E_Y} h(y) \sum_{x \in E_X} P(X = x) P(Y = y \mid X = x),$$

was äquivalent ist zu

$$\mathscr{E}[h(Y)] = \sum_{x \in E_X} P(X = x)\, \mathscr{E}[h(Y) \mid X = x].$$

Es gilt also

$$\mathscr{E}[h(Y)] = \mathscr{E}[\mathscr{E}(h(Y) \mid X)]. \tag{1.15}$$

Übungsaufgaben

1. Beweisen Sie, daß Verteilungsfunktionen monoton steigend sind.
2. X und Y seien zwei unabhängige, diskrete Zufallselemente. Zeigen Sie daß $\mathscr{E}[h(X)g(Y)] = \mathscr{E}[h(X)]\, \mathscr{E}[g(Y)]$ gilt.

1.9 Bedingte Erwartungen

Wir betrachten hier einen Wahrscheinlichkeitsraum $\langle \Omega, \mathfrak{A}, P \rangle$ mit einer Zufallsvariablen $X\colon \Omega \to \mathbb{R}$ (Def. 1.24) und einem diskreten Zufallselement $H\colon \Omega \to E_H$ (Def. 1.27). Die bedingten Erwartungswerte $\mathscr{E}(X \mid H(\omega) = h)$ sind Parameter der bedingten Verteilung von X, gegeben $H = h$. Da H ein Zufallselement, also eine Funktion auf Ω ist, kann auch der bedingte Erwartungswert als Funktion auf Ω betrachtet werden. Diese Funktion wird als *bedingte Erwartung* bezeichnet. Sie ist, da sie eine Funktion auf Ω ist, eine Zufallsvariable.

Definition 1.36. Die *bedingte Erwartung* einer Zufallsvariablen X, gegeben das diskrete Zufallselement H, ist die Zufallsvariable $T_X\colon \Omega \to \mathbb{R}$, die auf folgende Weise definiert ist: für jedes $\omega \in \Omega$ gilt

$$T_X(\omega) = t_X[H(\omega)],$$

wobei $t_X\colon E_H \to \mathbb{R}$ für alle h in E_H definiert ist durch

$$t_X(h) = \mathscr{E}(X \mid H = h).$$

Die bedingte Erwartung ist also eine Zufallsvariable, d. h. eine Abbildung von Ω in $\mathbb{R}$. Die Funktionswerte der Funktion T_X sind die bedingten Erwartungswerte $\mathscr{E}(X \mid H(\omega) = h)$.

Sei $\sim_H$ eine auf Ω definierte Äquivalenzrelation mit $\omega_1 \sim_H \omega_2$ gdw. $H(\omega_1) = H(\omega_2)$. Zwei Ergebnisse ω_1 und ω_2 des Zufallsexperiments sind

äquivalent im Sinne der Äquivalenzrelation $\sim_H$, wenn sie den gleichen Funktionswert H haben. Für die Äquivalenzklasse, in der sich ein Element ω befindet, schreibt man $[\![\omega]\!]$. Man kann dann die Zerlegung $\Omega/\sim_H$ von Ω bezüglich $\sim_H$ betrachten, also die Menge aller $\sim_H$-Äquivalenzklassen. Auf ihr ist die durch H induzierte σ-Algebra $\mathfrak{B}$, eine Teilmenge von $\mathfrak{A}$ aufgebaut. Die bedingte Erwartung T_X ist dann eine $\mathfrak{B}$-meßbare Zufallsvariable. T_X ist auf jeder Äquivalenzklasse $[\![\omega]\!] \subset \Omega$ konstant und gleich dem bedingten Erwartungswert von X, gegeben $[\![\omega]\!]$:

$$T_X(\omega) = \mathscr{E}(X \,|[\![\omega]\!]).$$

Da $\mathfrak{B}$ eine σ-Algebra über der Quotientenmenge $\Omega/\sim_H$ ist, heißt $\mathfrak{B}$-Meßbarkeit, daß T_X auf jeder $\sim_H$-Äquivalenzklasse $[\![\omega]\!]$ konstant ist. Wir werden im folgenden zur Vereinfachung der Schreibweise das Argument ω der Funktion $T_X(\omega)$ weglassen, wie das bei Zufallsvariablen üblich ist, falls der Definitionsbereich klar aus dem Kontext hervorgeht.

1.9.1 Rechenregeln für bedingte Erwartungen

1. Die bedingte Erwartung ist ein linearer Operator. Seien X und Y beliebige Zufallsvariablen, deren Erwartungswerte existieren, und a sei eine beliebige relle Zahl. Dann gilt

$$\begin{aligned} T_{X+Y}(\omega) &= \mathscr{E}(X+Y \,|[\![\omega]\!]) \\ &= \mathscr{E}(X \,|[\![\omega]\!]) + \mathscr{E}(Y \,|[\![\omega]\!]) \\ &= T_X(\omega) + T_Y(\omega) \qquad (1.16) \\ T_{aX}(\omega) &= \mathscr{E}(aX \,|[\![\omega]\!]) \\ &= a\,\mathscr{E}(X \,|[\![\omega]\!]) \\ &= a\,T_X(\omega). \qquad (1.17) \end{aligned}$$

2. Seien X und Y Zufallsvariablen, so daß $\mathscr{E}(X)$ und $\mathscr{E}(XY)$ existieren und Y auf jedem Element von $\Omega/\sim_H$ konstant ist. Dann gilt

$$\begin{aligned} T_{XY}(\omega) &= \mathscr{E}(XY \,|[\![\omega]\!]) \\ &= Y(\omega)\,\mathscr{E}(X \,|[\![\omega]\!]) \\ &\qquad\qquad \text{da } Y \text{ auf } [\![\omega]\!] \text{ konstant} \\ &= Y\,T_X(\omega). \qquad (1.18) \end{aligned}$$

   Als Spezialfälle hiervon ergeben sich

$$\begin{aligned} T_Y &= Y \qquad (1.19) \\ T_{T_X} &= T_X. \qquad (1.20) \end{aligned}$$

3. Wegen $\mathscr{E}(T_X \,|\,\Omega) = \mathscr{E}(X)$ gilt

$$\mathscr{E}(T_X) = \mathscr{E}(X). \qquad (1.21)$$

4. Aus (1.16) und (1.21) folgt

$$T_{X-\mathscr{E}(X)} = T_X - \mathscr{E}(T_X). \qquad (1.22)$$

2 Klassische Testtheorie

Der klassischen Testtheorie liegt die Vorstellung zugrunde, daß sich der bei einer Testanwendung beobachtete Punktewert, der *Beobachtungswert*, additiv aus einem „wahren“ Meßwert und einem Fehlerwert zusammensetzt. Dabei werden alle unsystematischen Anteile des Beobachtungswertes dem Fehler und alle systematischen Anteile dem Meßwert zugerechnet. Der Meßwert könnte in einfacher Weise dann hinreichend genau bestimmt werden, wenn der Meßvorgang beliebig oft unabhängig wiederholbar wäre, denn die wiederholten Beobachtungen würden eine Trennung von systematischen und unsystematischen Anteilen erlauben. Bei physikalischen Meßverfahren wird in vergleichbaren Situationen der Meßwert einfach durch das arithmetische Mittel wiederholter Beobachtungen geschätzt. In der psychologischen Diagnostik sind aber unabhängige Meßwiederholungen mit dem gleichen Verfahren in der Regel nicht möglich. Es bedarf daher einiger statistisch-theoretischer Überlegungen, um geeignete Meßverfahren zu entwickeln und den bei der Messung auftretenden Fehler abschätzen zu können.

Die Datenerhebung im Rahmen der Testkonstruktion wird als Zufallsexperiment aufgefaßt. Eine Besonderheit dieses Zufallsexperiments besteht darin, daß es in zwei Teilen durchgeführt wird. Der erste Teil besteht in der zufälligen Auswahl einer Person a aus der Population von Personen, auf die der Test anwendbar sein soll, und der zweite Teil besteht in der Bestimmung der Punktezahl, die die ausgewählte Person erreicht. Dieser zweite Teil ist ein Zufallsexperiment, weil die zu messende Eigenschaft nicht direkt beobachtbar ist, sondern indirekt über die Beantwortung von Testaufgaben bestimmt werden muß. Es ist aber anzunehmen, daß die Anzahl der gelösten Testaufgaben nur ein Indikator für die zugrundeliegende Fähigkeit ist, ein Indikator, der fehlerbehaftet und möglicherweise unzuverlässig ist. Ziel der psychologischen Testtheorie ist es, den Zusammenhang zwischen der zu untersuchenden Fähigkeit und dem beobachtbaren Indikator aufzuklären. Als Meßproblem formuliert kann man auch sagen, das Ziel besteht darin, Möglichkeiten und Methoden für die Definition der zu messenden Größe aufzuzeigen. Das inzwischen leider etwas veraltete Standardwerk zur psychologischen Testtheorie ist Lord und Novick (1968). Eine gute Zusammenfassung der wesentlichen Elemente der klassischen Testtheorie nach Lord und Novick gibt Fischer (1974, S. 1–145). Wir benutzen hier die einfachere und auf das wesentliche beschränkte Darstellung von Zimmerman (1975, 1976) und Zimmerman und Williams (1977). Diese Formulierung der klassischen Testtheorie liegt den meisten neueren Darstellungen zugrunde (Tack, 1980; Steyer, 1989; Knoche, 1990; Steyer & Eid, 1993).

2.1 Die Grundannahmen der klassischen Testtheorie

2.1.1 Beobachtungswert und Personenparameter

Für die Beschreibung des Zufallsgeschehens bei einer Testanwendung benutzen wir einen Wahrscheinlichkeitsraum $\langle \Omega, \mathfrak{A}, P \rangle$ (Def. 1.19). Der Ergebnisraum Ω enthält als Elemente ω alle möglichen Ergebnisse der Testentwicklung und Anwendung. Unter „Testentwicklung“ verstehen wir hier nicht nur die Bestimmung eines einzelnen Testwertes, sondern die gesamte Datenerhebung, die mit der Konstruktion und Validierung eines Testverfahrens verbunden ist. Dazu gehört sowohl die Bestimmung von Testwerten für mehrere äquivalente Testformen, als auch die Validierung eines Tests bezüglich externer Kriterien. Jedes Ergebnis $\omega \in \Omega$ muß angeben, von welcher Person

die Daten stammen und welche Daten bei dieser Person beobachtet wurden. ω ist also mehrdimensional. Eine Komponente beschreibt die Person, und eine oder mehrere weitere Komponenten beschreiben mögliche Testergebnisse der Person.

Die Punktezahl, die eine Person bei einem Test k erreicht, wird als Zufallsvariable $X_k\colon \Omega \to \overline{\mathbb{R}}$ betrachtet (Def. 1.24). Sie wird *Beobachtungswert* genannt. Für ein Testverfahren werden immer mehrere Beobachtungswerte benötigt, andernfalls ist eine Bestimmung der Meßgenauigkeit und der Validität eines Testverfahrens nicht möglich. Wir haben es also in der Regel mit einer Menge $\{X_k \mid k \in I\}$ von Beobachtungswerten zu tun, wobei I einfach eine endliche Indexmenge ist. Zur Vereinfachung der Schreibweise werden wir immer dann, wenn nur über einen oder zwei Beobachtungswerte gesprochen wird, die Bezeichnungen X und Y benutzen. Ein bestimmter Wert $X(\omega)$ gibt das Datum an, das von der in ω kodierten Person bei der Messung X erzeugt wurde, also etwa die Punktezahl der Person bei einem einzelnen Test. Da in ω neben der Person auch noch alle anderen Komponenten eines Ergebnisses kodiert sind, wird eine Funktion H benötigt, die angibt, welche Person in einem Ergebnis ω kodiert ist. Ist A die (abzählbare) Menge aller Personen, dann ist $H\colon \Omega \to A$ ein diskretes Zufallselement (Def. 1.27). $H(\omega) = a \in A$ bedeutet, daß beim Ergebnis ω die Person a zur Datenerhebung ausgewählt wurde.

Definition 2.1. Sei $\langle \Omega, \mathfrak{A}, P\rangle$ ein Wahrscheinlichkeitsraum, I sei eine endliche Indexmenge, $X_k\colon \Omega \to \overline{\mathbb{R}}$, $k \in I$ seien Zufallsvariable mit endlichem Erwartungswert und endlicher Varianz, A sei eine abzählbare Menge und $H\colon \Omega \to A$ ein diskretes Zufallselement. Dann ist $\langle \Omega, \mathfrak{A}, P, \{X_k \mid k \in I\}, H\rangle$ ein *System psychometrischer Daten*.

Wie bereits oben angedeutet, wird in der klassischen Testtheorie nicht der Beobachtungswert X_k einer Person a als Parameter für die zu diagnostizierende Fähigkeit benutzt, da dieser ja fehlerbehaftet ist. Als Personenparameter[1] wird stattdessen der bedingte Erwartungswert von X_k für die Person a definiert. Da aufgrund der Struktur des Zufallsexperiments die beobachtete Person bei der Testkonstruktion selbst Ergebnis eines Zufallsexperiments ist, ist auch der Parameter der Person a eine Zufallsvariable.

Definition 2.2. Sei $\langle \Omega, \mathfrak{A}, P, \{X_k \mid k \in I\}, H\rangle$ ein System psychometrischer Daten. Ferner sei $\mathfrak{B} \subseteq \mathfrak{A}$ die σ-Algebra der durch H induzierten Zerlegung von Ω. Der *Personenparameter T_{X_k} zum Beobachtungswert X_k* ist dann die $\mathfrak{B}$-meßbare[2] Zufallsvariable $T_{X_k}\colon \Omega \to \overline{\mathbb{R}}$, so daß für alle $B \in \mathfrak{B}$ gilt

$$\mathscr{E}(T_{X_k} \mid B) = \mathscr{E}(X_k \mid B). \tag{2.1}$$

Man nennt eine so definierte Zufallsvariable auch die *bedingte Erwartung von X_k, gegeben das diskrete Zufallselement H*.[3]

Die Zufallsvariable T_{X_k} ist auf dem gleichen Ergebnisraum Ω definiert wie X_k. Sie ist darüberhinaus mit der σ-Algebra verträglich, die aus $\mathfrak{A}$ entsteht, wenn man statt von Ω von der durch H induzierten Zerlegung Ω/H von Ω ausgeht. Die Elemente von Ω/H sind Teilmengen von Ω, auf denen die Funktion H konstant ist. Oder anders ausgedrückt, die Elemente von Ω/H sind Mengen von Ergebnissen ω, die von der gleichen Person $H(\omega) = a$ stammen. Die bedingte Erwartung T_{X_k} ist auf jeder dieser Mengen konstant,

[1] In der englischsprachigen Literatur (Lord & Novick, 1968) wird die Zufallsvariable X_k als „observerd score“ und die bedingte Erwartung T_{X_k} als „true score“ bezeichnet. Wir übersetzen diese Ausdrücke mit „Beobachtungswert“ und „Personenparameter“, auch wenn in der deutschsprachigen Literatur dafür häufig die Ausdrücke „Rohwert“ und „wahrer Wert“ gebraucht werden.

[2] Zur $\mathfrak{B}$-Meßbarkeit vgl. Def. 1.24 und 1.36.

[3] Zur Vereinfachung der Darstellung begnügen wir uns hier in einigen Fällen mit einer mathematisch unvollständigen Ausdrucksweise, wie sie auch von Zimmerman (1976) benutzt wird. Eine mathematisch vollständige Darstellung findet man bei Zimmerman (1975), Zimmerman und Williams (1977) und Knoche (1990).

und zwar gleich dem bedingten Erwartungswert von X_k beschränkt auf diese Menge. Es gilt also

$$T_{X_k}(\omega) = \mathscr{E}(X_k \mid H^{-1}(a)),$$

wobei $H^{-1}(a) = \{\omega \mid \omega \in \Omega, H(\omega) = a\}$ die Menge aller Ergebnisse der Person a ist. Mit Hilfe der Indikatorfunktion 1_B: $\Omega \to \{0,1\}$ (Def. (1.2)) kann man Gleichung (2.1) folgendermaßen ausdrücken: für alle $B \in \mathfrak{B}$ gilt

$$\mathscr{E}(1_B T_{X_k}) = \mathscr{E}(1_B X_k).$$

Aus Gl. (2.1) folgt auch, daß nicht nur für jede einzelne Person, sondern sogar in jeder Teilpopulation der Erwartungswert des Personenparameters gleich dem Erwartungswert des Beobachtungswertes ist.

Der Unterschied zwischen Personenparameter und Beobachtungswert wird als Meßfehler betrachtet.

Definition 2.3. Sei $\langle \Omega, \mathfrak{A}, P, \{X_k \mid k \in I\}, H\rangle$ ein System psychometrischer Daten. Der *Fehlerwert E_{X_k} zum Beobachtungswert X_k* ist die Zufallsvariable E_{X_k}: $\Omega \to \overline{\mathbb{R}}$, für die gilt

$$E_{X_k} = X_k - T_{X_k}. \tag{2.2}$$

2.1.2 Meßfehler und Reliabilität

Die Definition des Personenparameters und des Fehlerwertes zeigt das Konzept der klassischen Testtheorie: Den Beobachtungsdaten liegt ein Parmeter zugrunde, der additiv von einer Fehlerstreuung überlagert ist. Aus Gleichung (2.2) folgt ein linearer Zusammenhang zwischen dem Personenparameter und dem Beobachtungswert, wie er in Gleichung (2.10) zum Ausdruck kommt. Darin besteht ein wesentlicher Unterschied zwischen der klassischen Testtheorie und der Testtheorie, die auf logistischen Testmodellen aufbaut, wie etwa der von Rasch (1960) und Birnbaum (1968). Das folgende Theorem faßt die wichtigsten Aussagen der klassischen Testtheorie über den Zusammenhang zwischen Personenparametern und Fehlerwerten zusammen.

Theorem 2.1. *Sei $\langle \Omega, \mathfrak{A}, P, \{X, Y\}, H\rangle$ ein System psychometrischer Daten mit den Beobachtungswerten X und Y. T_X und T_Y seien die zugehörigen Personenparameter (Def. 2.2), E_X und E_Y die entsprechenden Fehlerwerte (Def. 2.3) und B ein Element der σ-Algebra $\mathfrak{B}$ der durch H induzierten Zerlegung von Ω (Def. 2.2). Dann gelten folgende Aussagen:*

1. *Der Beobachtungswert ist gleich der Summe aus Personenparameter und Fehlerwert:*
$$X = T_X + E_X. \tag{2.3}$$

2. *Der Erwartungswert des Fehlers ist in jeder Teilpopulation 0:*
$$\mathscr{E}(E_X \mid B) = 0. \tag{2.4}$$

3. *Die Korrelation zwischen Fehlerwert und Personenparameter zum Beobachtungswert X ist 0:*
$$\rho(E_X, T_X) = 0. \tag{2.5}$$

4. *Die Korrelation zwischen Fehlerwert zum Beobachtungswert X und Personenparameter zum Beobachtungswert Y ist 0:*
$$\rho(E_X, T_Y) = 0. \tag{2.6}$$

5. *Die Varianz der Beobachtungswerte ist gleich der Summe der Varianzen von Personenparameter und Fehlerwert:*
$$\sigma^2(X) = \sigma^2(T_X) + \sigma^2(E_X). \tag{2.7}$$

Begründungen

1. Gleichung (2.3) folgt unmittelbar aus Definition 2.3.

2. Für jedes $B \in \mathfrak{B}$ gilt

$$\begin{aligned} \mathscr{E}(E_X \mid B) &= \mathscr{E}(X - T_X \mid B) \\ &= \mathscr{E}(X \mid B) - \mathscr{E}(T_X \mid B) \\ &= 0. \qquad \text{wegen (2.1)} \end{aligned}$$

   Aus $\mathscr{E}(E_X \mid B) = 0$ für alle $B \in \mathfrak{B}$ ergeben sich noch einige weitere interessante Konsequenzen:

   (a) Für jede einzelne Person a ist der Erwartungswert des Meßfehlers 0:

$$\mathscr{E}(E_X \mid H = a) = 0. \tag{2.8}$$

   (b) Der Erwartungswert des Fehlers bei gegebenem Personenparameter ist ebenfalls 0:

$$\mathscr{E}(E_X \mid T_X = \tau) = 0. \tag{2.9}$$

   (c) Die Regression des Beobachtungswertes auf den Personenparameter ist eine Gerade durch (0,0) mit dem Anstieg 1:

$$\begin{aligned} \mathscr{E}(X \mid T_X = \tau) &= \mathscr{E}(T_X + E_X \mid T_X = \tau) \\ &= \mathscr{E}(T_X \mid T_X = \tau) + \mathscr{E}(E_X \mid T_X = \tau) \\ &= \tau. \end{aligned} \tag{2.10}$$

3. Gleichung (2.5) folgt aus (2.11) für den Fall $X = Y$.

4. Gleichung (2.6) folgt, da die Kovarianz zwischen E_X und T_Y den Wert 0 hat:

$$\begin{aligned} \sigma(E_X, T_Y) &= \mathscr{E}(E_X T_Y) - \mathscr{E}(E_X)\mathscr{E}(T_Y) \\ &= \mathscr{E}(E_X T_Y) \quad \text{wegen (2.4) ist } \mathscr{E}(E_X) = 0 \\ &= \mathscr{E}(\mathscr{E}(E_X T_Y \mid B)) \qquad \text{wegen (1.15)} \\ &= \mathscr{E}(T_Y \mathscr{E}(E_X \mid B)) \qquad \text{wegen (1.18)} \\ &= 0. \end{aligned} \tag{2.11}$$

5. Folgt sofort aus $\sigma(T_X, E_X) = 0$:

$$\begin{aligned} \sigma^2(X) &= \sigma^2(T_X) + \sigma^2(E_X) + 2\,\sigma(T_X, E_X) \\ &\qquad\qquad \text{wegen (1.10)} \\ &= \sigma^2(T_X) + \sigma^2(E_X). \end{aligned}$$

Eine Folgerung aus den Definitionen 2.2 und 2.3 ist also, daß die Fehlerwerte eines Tests nicht mit den Personenparametern zusammenhängen. Dies ist einfach eine Folge davon, daß durch diese Definitionen eben alles, was an einem Beobachtungswert systematisch mit der Personen zusammenhängt als Personenparameter und alles was nicht systematisch mit der Person zusammenhängt als Fehler festgelegt wird. Zur Beurteilung der Meßgenauigkeit eines Tests ist es notwendig, den Anteil des Fehlers am Beobachtungswert abzuschätzen. In der Testtheorie wird dazu der Anteil der Varianz des Personenparameters an der Gesamtvarianz benutzt.

Definition 2.4. Die *Reliabilität* $Rel(X)$ *eines Beobachtungswertes* X ist das Verhältnis der Varianz des Personenparameters zur Varianz des Beobachtungswertes:

$$Rel(X) = \frac{\sigma^2(T_X)}{\sigma^2(X)}. \tag{2.12}$$

Es ist leicht zu zeigen, daß die Reliablität das Quadrat der Korrelation zwischen Personenparameter und Beobachtungswert ist:

$$\begin{aligned}\rho^2(X,T_X) &= \frac{\sigma^2(X,T_X)}{\sigma^2(X)\sigma^2(T_X)} \\ &= \frac{[\sigma^2(T_X)]^2}{\sigma^2(X)\sigma^2(T_X)} \qquad \text{wegen (2.17)} \\ &= \frac{\sigma^2(T_X)}{\sigma^2(X)} \\ &= Rel(X). \qquad (2.13)\end{aligned}$$

Die Reliabilität gibt den Anteil der Personenparametervarianz an der Gesamtvarianz an. Die Wurzel aus der Reliabilität wird als *Reliabilitätsindex* bezeichnet. Mit Hilfe der Reliabilität wird der Meßfehler beurteilt. Geht man von der additiven Zerlegung der Gesamtvarianz nach (2.7) aus, und dividiert diese Gleichung durch $\sigma^2(X)$ so ergibt sich

$$1 = \frac{\sigma^2(T_X)}{\sigma^2(X)} + \frac{\sigma^2(E_X)}{\sigma^2(X)}.$$

Daraus erhält man

$$\sigma(E_X) = \sigma(X)\sqrt{1 - Rel(X)}. \qquad (2.14)$$

Wir werden später sehen, wie dieser Ausdruck zur Abschätzung des individuellen Meßfehlers und zur Berechnung eines Konfidenzintervalls für den Personenparameter benutzt werden kann.

2.1.3 Der Zusammenhang zwischen Beobachtungs- und Fehlerwerten verschiedener Tests

In Theorem 2.1 wird festgestellt, daß die Korrelation zwischen Meßfehler und Personenparameter 0 ist. Dies gilt sowohl innerhalb eines Tests (2.5) als auch zwischen zwei Tests (2.6). Daraus folgt aber nicht, daß die Korrelation der Meßfehler zweier verschiedener Tests ebenfalls 0 ist.

Wir betrachten zuerst den Zusammenhang zwischen dem Beobachtungswert X und dem Personenparameter im Test Y. Für beliebige Zufallsvariable X, Y mit endlicher Varianz gilt:

$$\begin{aligned}\mathscr{E}(XT_Y) &= \mathscr{E}(T_{XT_Y}) \quad \text{wegen (1.21) und (1.18)} \\ &= \mathscr{E}(T_XT_Y). \qquad (2.15)\end{aligned}$$

Die Kovarianz eines Beobachtungswertes X mit einem Personenparameter T_Y ist gleich der Kovarianz der zugehörigen Personenparameter:

$$\begin{aligned}\sigma(X,T_Y) &= \mathscr{E}(XT_Y) - \mathscr{E}(X)\mathscr{E}(T_Y) \\ &= \mathscr{E}(T_XT_Y) - \mathscr{E}(T_X)\mathscr{E}(T_Y) \\ &\qquad\qquad \text{wegen (2.15) und (1.21)} \\ &= \sigma(T_X,T_Y). \qquad (2.16)\end{aligned}$$

Für $X = Y$ folgt daraus

$$\sigma(X,T_X) = \sigma^2(T_X). \qquad (2.17)$$

Für die Kovarianz zwischen zwei Beobachtungswerten erhalten wir schließlich

$$\begin{aligned}\sigma(X,Y) &= \sigma(T_X + E_X, T_Y + E_Y) \\ &= \sigma(T_X,T_Y) + \sigma(T_X,E_Y) + \sigma(E_X,T_Y) + \sigma(E_X,E_Y) \\ &\qquad\qquad \text{wegen (2.11)} \\ &= \sigma(T_X,T_Y) + \sigma(E_X,E_Y).\end{aligned}$$

Die Kovarianz $\sigma(E_X, E_Y)$ zwischen den Meßfehlern in verschiedenen Tests wird im allgemeinen nicht 0 sein. In der älteren Literatur zur klassischen Testtheorie wird dies zusätzlich zu den Bedingungen des Theorems 2.1 gefordert. Es wird angenommen, daß die Fehlerwerte unterschiedlicher Tests nicht miteinander korreliert sind: für $X \neq Y$ gilt

$$\rho(E_X, E_Y) = 0. \tag{2.18}$$

Diese Bedingung ist aber nicht aus den Definitionen 2.2 und 2.3 abzuleiten, sondern muß zusätzlich gefordert werden. Dies könnte dadurch geschehen, daß man für die Zufallsvariablen X und Y lokale Unkorreliertheit nach Bedingung (2.19) fordert. Diese Bedingung wird in der Literatur auch *lineare, experimentelle Unabhängigkeit* genannt. Sie hat zur Folge, daß die Kovarianz der Zufallsvariablen X und Y für festes $H(\omega) = a$ verschwindet. Diese stellt eine abgeschwächte Form der *lokalen stochastischen Unabhängigkeit* dar.

Die Gleichungen (2.3), (2.4), (2.5) und (2.6) werden zusammen mit (2.18) in der älteren Literatur als *Grundannahmen* der klassischen Testtheorie bezeichnet, da sie deren formalen Kern enthalten, aus dem alle anderen Folgerungen ableitbar sind. Wie Theorem 2.1 zeigt, sind aber außer (2.18) bereits diese Grundannahmen aus den Definitionen 2.2 und 2.3 und den Annahmen über den vorliegenden Wahrscheinlichkeitsprozeß herzuleiten.

2.2 Abschätzung des Meßfehlers

2.2.1 Parallele Messungen

Wie bereits früher angedeutet, sind in der Psychodiagnostik Meßwiederholungen im Sinne einer wiederholten Anwendung des gleichen Meßinstruments in der Regel nicht möglich. Da ganz ohne Meßwiederholungen eine Abschätzung des Meßfehlers aber nicht möglich ist, wird in der klassischen Testtheorie ein Konzept eingeführt, das Meßwiederholungen in einem Sinne erlaubt, der für die Abschätzung des Fehlers ausreichend ist. Für die Meßwiederholung wesentlich ist, daß der systematische Anteil am Beobachtungswert jeder einzelnen Messung und die für die statistischen Überlegungen wesentlichen Parameter der Fehlerverteilung gleich bleiben. Diese Bedingungen werden im Konzept der parallelen Messung erfaßt. In der folgenden Definition betrachten wir die Restriktion einer Zufallsvariablen auf eine Teilmenge von Ω: $X \mid H = a$ ist die Beschränkung von X auf die Ergebnisse ω, für die $H(\omega) = a$.

Definition 2.5. Sei $\langle \Omega, \mathfrak{A}, P, \{X, Y\}, H\rangle$ ein System psychometrischer Daten.

1. Die Zufallsvariablen X und Y heißen *lokal unkorreliert*, wenn für alle $a \in H(\Omega)$

$$\sigma(X \mid H = a, Y \mid H = a) = 0. \tag{2.19}$$

2. Die Zufallsvariablen X und Y heißen *parallel*, wenn sie lokal unkorreliert sind und für alle $a \in H(\Omega)$ folgende Bedingungen gelten:

$$\mathscr{E}(X \mid H = a) = \mathscr{E}(Y \mid H = a) \tag{2.20}$$
$$\sigma^2(X \mid H = a) = \sigma^2(Y \mid H = a) \tag{2.21}$$

Eine unmittelbare Folge von (2.20) und (2.21) ist, daß für parallele Tests $\mathscr{E}(X) = \mathscr{E}(Y)$ und $\sigma^2(X) = \sigma^2(Y)$. Erwartungswert und Varianz paralleler Tests sind gleich. Dies gilt sogar für jede Teilpopulation. Insbesondere heißt dies auch, daß bei jeder einzelnen Person der gleiche Personenparameter und wegen (2.21) auch die gleiche lokale Fehlervarianz vorliegen muß, denn $\sigma^2(X \mid H = a) = \sigma^2(E_X)$. Es wird nicht verlangt, daß bei jeder Person der gleiche Meßfehler vorliegt, denn dann müßten ja die Beobachtungswerte

übereinstimmen. Aber es wird verlangt, daß bei jeder Person der Personenparameter und die Varianz des Meßfehlers in beiden Tests gleich sind. Man kann also sagen, daß parallele Messungen das gleiche messen, und zwar mit der gleichen Genauigkeit.

Empirisch kann man Paralleltests bis zu einem gewissen Grad daran erkennen, daß die Bedingungen (2.20) und (2.21) in jeder Teilpopulation gelten. Eine hinreichende empirische Bestätigung der Parallelität zweier Messungen ist allerdings mit Hilfe dieser Bedingungen nicht möglich, da weder die Erwartungswerte $\mathscr{E}(X)$ noch die Varianzen $\sigma^2(X)$ in hinreichend kleinen Teilpopulationen bestimmt werden können, insbesondere nicht in Teilpopulationen mit nur einer einzigen Person.

2.2.2 Paralleltests und die empirische Bestimmung der Reliabilität

Aus der lokalen Unkorreliertheit paralleler Messungen X und Y in Bedingung (2.19) folgt, daß die Kovarianz der Fehler E_X und E_Y verschwindet, denn es ist $\sigma(X \mid H = a, Y \mid H = a) = \sigma(E_X, E_Y)$. Damit ist wegen Gleichung (2.18) und Bedingung (2.20) die Kovarianz zwischen parallelen Tests gleich der Varianz der Personenparameter:

$$\sigma(X, Y) = \sigma(T_X, T_Y) = \sigma^2(T_X). \tag{2.22}$$

Für die Korrelation paralleler Tests X und Y erhält man daher

$$\begin{aligned} \rho(X, Y) &= \frac{\sigma(X, Y)}{\sqrt{\sigma^2(X)\sigma^2(Y)}} \\ &= \frac{\sigma^2(T_X)}{\sigma^2(X)} \qquad (2.23) \\ &= \mathit{Rel}(X). \qquad (2.24) \end{aligned}$$

Die Reliabilität eines Tests kann also mit Hilfe der Korrelation zwischen parallelen Tests berechnet werden. Damit ist die mit Hilfe der nicht beobachtbaren Varianz $\sigma^2(T_X)$ definierte Reliabilität durch beobachtbare Größen ausgedrückt und auf diese Weise empirisch bestimmbar, falls es gelingt parallele Tests zu konstruieren.

Aus Gleichung (2.24) und den vorausgehenden Überlegungen ergeben sich weitere empirische Prüfmöglichkeiten der Parallelität einer Reihe von Tests $X_1, \ldots, X_n$:

1. Sind Tests $X_1, \ldots, X_n$ parallel, dann sind ihre Interkorrelationen gleich:
$$\rho(X_1, X_2) = \rho(X_1, X_3) = \ldots = \rho(X_{n-1}, X_n). \tag{2.25}$$
2. Sind Tests $X_1, \ldots, X_n$ parallel und ist ein Test Y mit $X_1, \ldots, X_n$ lokal unkorreliert, dann sind alle Korrelationen zwischen den Tests $X_1, \ldots, X_n$ und Y gleich:
$$\rho(X_1, Y) = \ldots = \rho(X_n, Y). \tag{2.26}$$

2.2.3 Ein Konfidenzintervall für den Personenparameter

Für die Standardabweichung des Meßfehlers, den *Standardmeßfehler* ergab sich in Gleichung (2.14)

$$\sigma(E_X) = \sigma(X)\sqrt{1 - \mathit{Rel}(X)}.$$

Üblicherweise wird in der klassischen Testtheorie angenommen, daß der Standardmeßfehler für alle Personen gleich ist. Der Standardmeßfehler kann dann zur Berechnung eines Konfidenzintervalls für den Personenparameter T_X beim Beobachtungswert X benutzt werden. Geht man von der Tschebyschewschen Ungleichung (1.14) aus, ist dafür keine weitere Annahme über die Verteilung von E_X nötig:

$$P\left(|X - T_X| \;\leq\; (1/\sqrt{\alpha})\,\sigma(E_X)\right) \;\geq\; 1 - \alpha \tag{2.27}$$

definiert ein Konfidenzintervall für den Personenparameter T_X zum Niveau $1-\alpha$:

$$[x - (1/\sqrt{\alpha})\sigma(E_X),\ x + (1/\sqrt{\alpha})\sigma(E_X)],$$

Beispiel. Für den *Test für medizinische Studiengänge* (TMS) gibt Stumpf (1988) eine Reliabilität von $Rel(X) = 0.91$, einen Gesamtmittelwert von $\overline{X} = 104.17$ und eine Standardabweichung des Beobachtungswertes von $\sigma(X) = 21.87$ an. Daraus ergibt sich ein Standardmeßfehler von

$$\begin{aligned}\sigma(E_X) &= 21.87\sqrt{1-0.91} \\ &= 6.56.\end{aligned}$$

Bei einem Beobachtungswert von $x = 112$ ergibt sich dann für ein Konfidenzintervall zum Niveau 0.75

$$\begin{aligned}&[112 - (1/\sqrt{0.25}) \cdot 6.56,\ 112 + (1/\sqrt{0.25}) \cdot 6.56] \\ &= [98.88, 125.12] \qquad (2.28)\end{aligned}$$

Bei einem Test mit einem Standardmeßfehler von $\sigma(E_X) = 6.56$ ist also in mindestens 75 % der Fälle der Personenparameter T_X zwischen 98.88 und 125.12, falls der Beobachtungswert X den Wert 112 hat. Das Konfidenzintervall zum Niveau 0.75 hat eine Breite von 26.24 Rohwertpunkten, ist also sehr groß. Dies liegt daran, daß bei der Verwendung der Tschebyschewschen Ungleichung keine Annahme über die Verteilungsfunktion der Fehlerwerte gemacht wird.

Beispiel. Der verhältnismäßig große Wert für das Konfidenzintervall in Gleichung (2.28) ist unter anderem darauf zurückzuführen, daß bei Anwendung der Tschebyschewschen Ungleichung keine Annahme über die Verteilung der Fehler gemacht wird. Nimmt man dagegen an, daß die Fehlerwerte normalverteilt sind, dann ergibt sich ein Konfidenzintervall aus der Gleichung

$$P\left(|X - T_X| \ \leq\ z(1-\alpha/2)\,\sigma(E_X)\right) \ =\ 1-\alpha.$$

Hier ist $z(1-\alpha/2)$ durch die Standardnormalverteilung definiert:

$$P(Z \leq z(1-\alpha/2)) = 1 - \alpha/2.$$

Für ein Konfidenzintervall zum Niveau $(1-\alpha) = 0.75$ erhält man

$$\begin{aligned}&[x - z(1-\alpha/2)\sigma(E_X),\ x + z(1-\alpha/2)\sigma(E_X)] \\ &= [112 - 0.809 \cdot 6.56,\ 112 + 0.809 \cdot 6.56] \\ &= [106.69,\ 117.31].\end{aligned}$$

Die Breite dieses Intervalls ist 10.62 Rohwertpunkte, also erheblich geringer als bei der Anwendung der Tschebyschewschen Ungleichung.

2.2.4 Eine Regressionsschätzung des Personenparameters

Im allgemeinen wird der Personenparameter T_X einer getesteten Person durch den Beobachtungswert X geschätzt. Man kann durch das Einbeziehen der Reliabilität und des Populationsmittelwertes die Güte der Schätzung verbessern. Dazu wird die Regression $\mathscr{E}(T_X \,|\, X = x)$ des Personenparameters auf den Beobachtungswert betrachtet. Nimmt man an, daß die Regression linear ist, also die allgemeine Form

$$\mathscr{E}(T_X \,|\, X = x) = \alpha + \beta\, x$$

hat, so muß für α und β gelten:

$$\begin{aligned}\beta &= \frac{\sigma(T_X, X)}{\sigma^2(X)} \\ &= \frac{\sigma^2(T_X)}{\sigma^2(X)} \\ &= Rel(X) \\ \alpha &= \mathscr{E}(T_X) - \beta\mathscr{E}(X) \\ &= \mathscr{E}(X)[1 - Rel(X)].\end{aligned}$$

Für die Regressionsgleichung ergibt sich

$$\mathscr{E}(T_X \mid X = x) = Rel(X)\, x + [1 - Rel(X)]\, \mathscr{E}(X). \tag{2.29}$$

Die Schätzung des Personenparameters mit Hilfe der Regressionsgleichung (2.29) ist also eine konvexe Mischung zwischen dem beobachteten Datum x und dem Erwartungswert von X in der Population. Je größer die Reliabilität umso größer ist das Gewicht des aktuellen Datums x. Bei geringer Reliabilität wird dagegen dem Erwartungswert mehr Gewicht beigemessen. Der Vorteil dieser Schätzung zeigt sich dann, wenn man den Fehler betrachtet, der dabei auftritt. Bei der einfachen Vorgehensweise X als Schätzwert für T_X zu verwenden, ist ja der Fehler $E_X = X - T_X$. Bei der Regressionsschätzung ist der Fehler

$$\begin{aligned} E_X^* &= \mathscr{E}(T_X \mid X = x) - T_X \\ &= [Rel(X)x + (1 - Rel(X))\mathscr{E}(X)] - T_X. \end{aligned}$$

Für die Varianz von E_X^* erhält man dann unter Berücksichtigung des Sachverhalts, daß $[1 - Rel(X)]\mathscr{E}(X)$ eine Konstante ist,

$$\begin{aligned} \sigma^2(E_X^*) &= \sigma^2(Rel(X)X + [1 - Rel(X)]\mathscr{E}(X) - T_X) \\ &= \sigma^2(Rel(X)X - T_X) \\ &= \sigma^2(Rel(X)X) + \sigma^2(T_X) - 2\sigma(Rel(X)X, T_X) \\ &= Rel^2(X)\sigma^2(X) + \sigma^2(T_X) - 2Rel(X)\sigma(X, T_X) \\ &= \Big(\frac{\sigma^2(T_X)}{\sigma^2(X)}\Big)^2 \sigma^2(X) + \sigma^2(T_X) - 2Rel(X)\sigma^2(T_X) \\ &= \sigma^2(T_X)Rel(X) + \sigma^2(T_X) - 2Rel(X)\sigma^2(T_X) \\ &= \sigma^2(T_X)(1 - Rel(X)). \end{aligned}$$

Für die Standardabweichung des Fehlers ergibt sich in diesem Fall

$$\sigma(E_X^*) = \sigma(T_X)\sqrt{1 - Rel(X)}. \tag{2.30}$$

Dieser Ausdruck wird *Standardschätzfehler* genannt. Da

$$\sigma(T_X) = \sqrt{Rel(X)}\sigma(X),$$

ist der Standardschätzfehler immer kleiner als der Standardmeßfehler $\sigma(E_X)$ von Gleichung (2.14). Wie man leicht nachprüfen kann, gilt

$$\sigma(E_X^*) = \sqrt{Rel(X)}\, \sigma(E_X).$$

Beispiel. Für das obige Beispiel des TMS ergibt sich ein Standardschätzfehler von

$$\begin{aligned} \sigma(E_X^*) &= \sqrt{0.91} \cdot 6.56 \\ &= 6.26. \end{aligned}$$

Der Standardschätzfehler und der Standardmeßfehler unterscheiden sich hier kaum, weil die Reliabilität des Tests verhältnismäßig hoch ist. Als Schätzwert für den Personenparameter ergibt sich bei einem Beobachtungswert von $x = 112$:

$$\begin{aligned} \hat{T}_X &= 0.91 \cdot 112 \;+\; 0.09 \cdot 104.17 \\ &= 111.3. \end{aligned}$$

Für das Konfidenzintervall zum Niveau 0.75 erhält man

$$[111.3 - (1/\sqrt{0.25}) \cdot 6.26,\; 111.3 + (1/\sqrt{0.25}) \cdot 6.26] \;=\; [98.78, 123.82].$$

Es hat eine Breite von 25.04 Rohwertpunkten.

2.3 Validität

Die Reliabilität kann als ein Maß der internen Konsistenz eines Tests aufgefaßt werden. Sie beschreibt unabhängig von der inhaltlichen Bedeutung seine Meßgenauigkeit. Eine hohe Reliabilität ist daher immer eine der wichtigsten Zielvorgaben bei der Entwicklung eines Tests. Für die praktische Brauchbarkeit eines Tests ist eine hohe Meßgenauigkeit allerdings nicht hinreichend. Mindestens genauso wichtig ist hierfür die Eignung des Tests, andere, vom Test unabhängige, Verhaltensdaten vorhersagen zu können. Häufig werden Testergebnisse benutzt, um andere quantitative Daten, die ebenfalls als Messungen aufgefaßt werden können, vorherzusagen. Die Güte solcher Vorhersagen wird in der klassischen Testtheorie mit Hilfe der Korrelation zwischen den beiden Messungen beschrieben.

Definition 2.6. Sei $\langle \Omega, \mathfrak{A}, P, \{X, Y\}, H\rangle$ ein System psychometrischer Daten. Der *Validitätskoeffizient* einer Messung X bezüglich einer zweiten Messung Y ist die Korrelation zwischen X und Y:

$$\rho(X,Y) = \frac{\sigma(X,Y)}{\sigma(X)\sigma(Y)}. \tag{2.31}$$

Der Validitätskoeffizient definiert die Semantik eines Tests, da er angibt, für welche anderen Messungen der Test Vorhersagen erlaubt. Damit ist auch klar, daß man nicht von „der Validität“ eines Tests sprechen kann, da ein Test viele unterschiedliche Validitätskoeffizienten haben kann, was eben nur heißt, daß er verschiedene andere Messungen unterschiedlich gut vorherzusagen erlaubt.

Wir werden im folgenden sehen, daß zwischen der Reliabilität eines Tests und seinen Validitätskoeffizienten bestimmte Zusammenhänge bestehen. Dies ist nicht weiter verwunderlich, denn es ist kaum zu erwarten, daß ein schlechtes Meßinstrument gute Prognosen ermöglichen kann. Die empirische Bestimmung eines Validitätskoeffizienten ist häufig mit einem nicht unerheblichen Aufwand verbunden. Da der Validitätskoeffizient für eine bestimmte Anwendung aber das wichtigste Erfolgskriterium ist, bzw. alle anderen Erfolgskriterien von ihm abhängen (vgl. Abschnitt 4.4), werden wir in den folgenden Abschnitten einige Methoden betrachten, die Schätzung von Validität und Reliabilität zu verbessern.

2.3.1 Minderung der Validität durch Meßfehler

Die Korrelation zwischen zwei Beobachtungswerten, wie sie bei der Bestimmung der Validität betrachtet wird, hängt einerseits von der Korrelation der beiden Personenparameter und andererseits davon ab, wie genau diese bestimmt werden können. Dies zeigt folgende Überlegung, bei der Gleichung (2.18) und die Beziehung $\sigma^2(X) = \sigma^2(T_X)/Rel(X)$ benutzt werden.

$$\begin{aligned}
\rho(X,Y) &= \frac{\sigma(X,Y)}{\sqrt{\sigma^2(X)\sigma^2(Y)}} \\
&= \frac{\sigma(T_X,T_Y)+\sigma(E_X,E_Y)}{\sqrt{\sigma^2(T_X)/Rel(X)\,\sigma^2(T_Y)/Rel(Y)}} \\
&= \sqrt{Rel(X)Rel(Y)}\,\frac{\sigma(T_X,T_Y)+\sigma(E_X,E_Y)}{\sqrt{\sigma^2(T_X)\sigma^2(T_Y)}} \\
&= \sqrt{Rel(X)Rel(Y)}\Big(\rho(T_X,T_Y)+\frac{\sigma(E_X,E_Y)}{\sqrt{\sigma^2(T_X)\sigma^2(T_Y)}}\Big).
\end{aligned}$$

Für die Korrelation der Personenparameter T_X und T_Y erhält man hieraus nach Einsetzen von $\sigma^2(E_X) = \sigma^2(X)(1 - Rel(X))$ und Vereinfachen

$$\begin{aligned}
\rho(T_X,T_Y) &= \frac{\rho(X,Y)}{\sqrt{Rel(X)Rel(Y)}} \\
&- \frac{\rho(E_X,E_Y)\sqrt{(1-Rel(X))(1-Rel(Y))}}{\sqrt{Rel(X)Rel(Y)}}.
\end{aligned} \tag{2.32}$$

Für lokal unkorrelierte Tests mit $\rho(E_X, E_Y) = 0$ ergibt sich die sogenannte *Verdünnungsformel*:

$$\rho(T_X, T_Y) = \frac{\rho(X, Y)}{\sqrt{Rel(X)Rel(Y)}}. \tag{2.33}$$

Sie erlaubt die Abschätzung der Korrelation zwischen den Personenparametern T_X und T_Y aufgrund der Korrelation der Beobachtungswerte und der Reliabilitätskoeffizienten der beiden Messungen. Gleichung (2.32) zeigt, daß positiv korrelierte Fehler die Personenparameterkorrelation vermindern. Umgekehrt erhöhen bei fester Personenparameterkorrelation positiv korrelierte Fehler die Korrelation der Beobachtungswerte. Dies kann dazu führen, daß bei nicht gerechtfertigter Anwendung von Formel (2.33) zur Berechnung der Personenparameterkorrelation diese deutlich überschätzt wird.

Die Formel (2.33) für lokal unkorrelierte Tests zeigt gleichzeitig wie hoch die Korrelation zwischen zwei lokal unkorrelierten Beobachtungswerten unter optimalen Bedingungen werden kann. Für die Validität ist der optimale Fall dann erreicht, wenn $\rho(T_X, T_Y) = 1$ und $Rel(Y) = 1$. Dann ist

$$\rho(X, Y) = \sqrt{Rel(X)}.$$

Die Validität eines Tests bezüglich eines lokal unkorrelierten Kriteriums ist also höchstens gleich dem *Reliabilitätsindex*, der Quadratwurzel aus der Reliabilität. Unter optimalen Bedingungen könnte also eine Validität eines Tests größer sein als seine Reliabilität. Eine analoge Betrachtung von Gleichung (2.32) zeigt, daß im allgemeinen Fall, also bei korrelierten Fehlern, die Validität eines Tests sogar größer als der Reliabilitätsindex werden kann (Zimmerman & Williams, 1977).

2.3.2 Bestimmung der Validität bei selegierten Stichproben

Testergebnisse werden häufig zu Selektionsentscheidungen benötigt, bei denen Vorhersagen eines bestimmten Kriteriums zu machen sind. Zur Bestimmung der Validität wäre es in diesem Fall notwendig, sowohl den Testwert als auch die Kriteriumsvariable in nicht selegierten Stichproben zu bestimmen. Dies ist aber häufig nicht möglich, weil die Kriteriumsvariable oft nur von einer selegierten Stichprobe erhoben werden kann. Bei einem Testverfahren zur Vorhersage eines Ausbildungserfolges können beispielsweise die Kosten der Ausbildung so hoch sein, daß eine Aufnahme aller Personen aus ökonomischen Gründen nicht zu vertreten wäre. Wird die Stichprobe für die Erhebung des Kriteriums aber mit Hilfe des Testwertes ausgewählt, so findet eine Minderung der Varianz statt, falls der Testwert mit dem Kriterium korreliert ist. Abbildung 2.1 zeigt den Zusammenhang zwischen der Selektionsrate und der Standardabweichung in einer Stichprobe, die nach dem Testwert selegiert ist.

Liegt eine Varianzeinschränkung durch Selektion vor, so kann die Validität nur dann geschätzt werden, wenn spezielle Annahmen über den Zusammenhang zwischen Testwert und Kriterium und über deren Varianzen gemacht werden. Lord und Novick (1968) gehen davon aus, daß die Regression des Kriteriums Y auf den Testwert X linear ist:

$$\mathscr{E}(Y \mid x) = \alpha + \beta x,$$

und daß die bedingte Varianz von Y, gegeben x für alle x gleich ist: Für alle x_1, x_2 gilt

$$\sigma^2(Y \mid x_1) = \sigma^2(Y \mid x_2).$$

Ist die Regressionsfunktion linear, so hängen deren Parameter nicht von der Art der Selektion ab, müssen also in der selegierten und der nicht selegierten Population gleich sein. Wir schreiben X' und Y' für die Variablen X und Y in der selegierten Population. Es gilt also

$$\beta(Y \mid x) \quad = \quad \rho(X, Y)\frac{\sigma(Y)}{\sigma(X)}$$

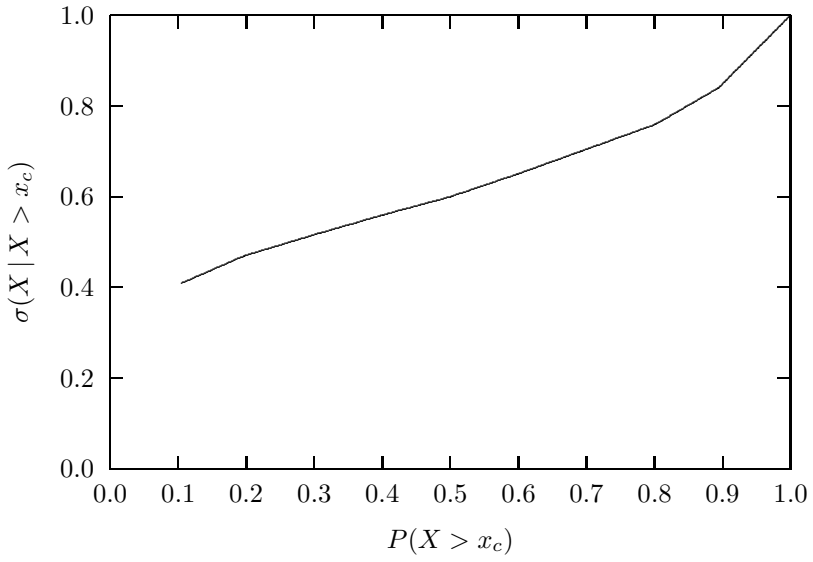

Abb. 2.1. Verminderung der Standardabweichung eines Merkmals durch Selektion. Wird ein Merkmal X beobachtet und die Standardabweichung von X in der Stichprobe geschätzt, so vermindert sich diese, wenn die Schätzung auf einen aufgrund von X selegierten Teil der Stichprobe beschränkt wird (nach Lord & Novick, 1968, Tab. 6.8.1, S. 141).

$$\begin{aligned} &= \rho(X', Y')\frac{\sigma(Y')}{\sigma(X')} \\ &= \beta(Y' \mid x'). \end{aligned} \tag{2.34}$$

Wegen der Unabhängigkeit der Residualvarianz von Y gegeben x von X ist die bedingte Residualvarianz in der selegierten und der nicht selegierten Population identisch:

$$\sigma^2(Y \mid x) = \sigma^2(Y' \mid x').$$

Löst man Gleichung (2.34) nach $\sigma(Y)$ auf, dann erhält man

$$\sigma(Y) = \frac{\rho(X', Y')\,\sigma(Y')\,\sigma(X)}{\rho(X, Y)\,\sigma(X')}.$$

Dies kann man in den Ausdruck

$$\sigma^2(Y')\,(1 - \rho^2(X', Y')) = \sigma^2(Y)\,(1 - \rho^2(X, Y))$$

für die Fehlervarianzen einsetzen und erhält

$$\sigma^2(Y')\,(1 - \rho^2(X', Y')) = \frac{\rho(X', Y')\,\sigma(Y')\,\sigma(X)}{\rho(X, Y)\,\sigma(X')}\,(1 - \rho^2(X, Y)).$$

Daraus erhält man das Ergebnis

$$\rho^2(X, Y) = \frac{1}{1 + \dfrac{\sigma^2(X')}{\sigma^2(X)}\left(\dfrac{1}{\rho^2(X', Y')} - 1\right)} \tag{2.35}$$

$$\sigma^2(Y) = \sigma^2(Y')\left(1 - \rho^2(X', Y') + \rho^2(X', Y')\frac{\sigma^2(X)}{\sigma^2(X')}\right). \tag{2.36}$$

Beispiel. Angenommen man beobachtet in einer selegierten Population mit $\sigma^2(X') = 80$ eine Korrelation von $\rho(X', Y') = 0.40$. Die Varianz in der nicht selegierten Population sei $\sigma^2(X) = 100$. Nach Formel (2.35) ergibt sich dann

$$\begin{aligned} \rho^2(X, Y) &= \frac{1}{1 + \dfrac{80}{100}\left(\dfrac{1}{0.16} - 1\right)} \\ &= 0.1923. \end{aligned}$$

Die Validität in der nicht selegierten Population ist also $\rho(X, Y) = 0.44$. Wie man leicht nachrechnen kann, sinkt die beobachtete Validität $\rho(X', Y')$ auf 0.24 ab, wenn so stark selegiert wird, daß die Varianz $\sigma^2(X')$ nur noch den Wert 32 hat.

2.4 Verbesserung der Reliabilität durch Testverlängerung

Auf Seite 29 wurde darauf hingewiesen, daß parallele Messungen ein Ersatz für Meßwiederholungen sind, mit dessen Hilfe eine Abschätzung des Meßfehlers möglich wird. Meßwiederholungen dienen im allgemeinen nicht nur zur Bestimmung des Meßfehlers, sondern auch zu dessen Verminderung. Ist eine physikalische Messung stark fehlerbehaftet, so wird man die Messung mehrmals durchführen und den Mittelwert aller Einzelmessungen als „Personenparameter“ benutzen. In analoger Weise können parallele Messungen in der Psychodiagnostik zur Verbesserung der Meßgenauigkeit benutzt werden. Sind etwa mehrere parallele Tests vorhanden, so kann man diese zu einem einzigen Test zusammenfügen und erhält dadurch eine Messung mit höherer Genauigkeit. Die Reliabilität des neuen Tests läßt sich sogar aus der Reliabilität der einzelnen Paralleltests berechnen.

Wir betrachten hier nur den einfachen Fall von zwei parallelen Messungen X und Y, die zu einem neuen Test zusammengefügt werden. Für die Varianz des komponierten Tests gilt

$$\begin{aligned} \sigma^2(X+Y) &= \sigma^2(X) + \sigma^2(Y) + 2\sigma(X,Y) \\ &= 2\sigma^2(X) + 2\sigma^2(T_X) \\ &= 2\sigma^2(X) + 2\sigma^2(X)Rel(X) \\ &= 2\sigma^2(X)[1 + Rel(X)]. \end{aligned} \qquad \text{Gl. (1.9)}$$

Für die Reliabilität des komponierten Tests gilt dann

$$\begin{aligned} Rel(X+Y) &= \frac{\sigma^2(T_{X+Y})}{\sigma^2(X+Y)} \\ &= \frac{\sigma^2(T_X + T_Y)}{\sigma^2(X+Y)} \\ &= \frac{4\sigma^2(T_X)}{2\sigma^2(X)[1 + Rel(X)]} \\ &= \frac{2Rel(X)}{1 + Rel(X)}. \end{aligned} \qquad (2.37)$$

Die entsprechende Formel für n parallele Tests wird als *Spearman-Brown-Formel* bezeichnet:

$$Rel(X_1 + \ldots + X_n) = \frac{nRel(X_1)}{1 + (n-1)Rel(X_1)}. \qquad (2.38)$$

Gleichung (2.37) oder die Spearman-Brown-Formel (2.38) können auch zur Abschätzung der Reliabilität benutzt werden, wenn sich ein Test in 2 oder n parallele Teile zerlegen läßt. Man kann dann die Reliabilität der Teiltests wegen (2.25) über ihre Interkorrelationen bestimmen und mit Hilfe der geeigneten Formel auf den Gesamttest hochrechnen.

Sind die Messungen X und Y nicht parallel, so gibt die folgende Formel (Koeffizient α) eine untere Schranke der Reliabilität der Summe $X + Y$ an:

$$\alpha = 2\Big(1 - \frac{\sigma^2(X) + \sigma^2(Y)}{\sigma^2(X+Y)}\Big). \qquad (2.39)$$

Dieser Ausdruck ist gleich der Reliabilität, wenn in allen Teilpopulationen $\sigma^2(X) = \sigma^2(Y)$ und $T_Y = T_X + s$ gilt. Für mehr als zwei Testteile verallgemeinert sich die Formel zu

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n}\sigma^2(X_i)}{\sigma^2(\sum_{i=1}^{n} X_i)}\right). \qquad (2.40)$$

Zur Abschätzung der Reliabilität eines Tests ist der Koeffizient α nur dann zu empfehlen, wenn die Aufteilung in parallele Testteile nicht gelingt.

2.5 Die Reliabilität von Beobachtungswertdifferenzen

Bei Testanwendungen kommt es häufig vor, daß Differenzen von Testwerten bestimmt und beurteilt werden. So kann etwa zur Bestimmung des Behandlungserfolges ein Testergebnis vor der Behandlung mit einem Testergebnis nach der Behandlung verglichen werden. Leider zeigt eine genauere Analyse, daß Testwertdifferenzen häufig nur eine sehr geringe Reliabilität haben. Wir betrachten zwei Beobachtungswerte X und Y und fragen, wie groß bei gegebener Reliabilität von X und Y die Reliabilität der Differenzen $D = X - Y$ ist. Zur Vereinfachung gehen wir davon aus, daß die Varianzen der beiden Tests gleich sind: $\sigma^2(X) = \sigma^2(Y)$, und daß die beiden Tests lokal unkorreliert sind: $\rho(E_X, E_Y) = 0$. Für die Varianz des Beobachtungswertes $D = X - Y$ und des Personenparameters $T_D = T_{X-Y}$ gilt dann

$$\begin{aligned}
\sigma^2(D) &= \sigma^2(X - Y) \\
&= \sigma^2(X) + \sigma^2(Y) - 2\sigma(X, Y) \qquad \text{Gl. (1.10)} \\
&= 2\sigma^2(X)(1 - \rho(X, Y) \\
\sigma^2(T_D) &= \sigma^2(T_{X-Y}) \\
&= \sigma^2(T_X - T_Y) \\
&= \sigma^2(T_X) + \sigma^2(T_Y) - 2\sigma(T_X, T_Y) \\
&= \sigma^2(X)[Rel(X) + Rel(Y)] - 2\sigma^2(X)\rho(X, Y) \\
&= \sigma^2(X)[Rel(X) + Rel(Y) - 2\rho(X, Y)].
\end{aligned}$$

Für die Reliabilität von D ergibt sich

$$\begin{aligned}
Rel(D) &= \frac{\sigma^2(X)[Rel(X) + Rel(Y) - 2\rho(X, Y)]}{2\sigma^2(X)(1 - \rho(X, Y))} \\
&= \frac{[Rel(X) + Rel(Y)]/2 - \rho(X, Y)}{1 - \rho(X, Y)}. \qquad (2.41)
\end{aligned}$$

Die Reliabilität der Differenzen nimmt also mit den Reliabilitätswerten der Einzelmessungen monoton zu, nimmt aber mit der Korrelation der beiden Teilmessungen monoton ab. Da man häufig für die beiden Teilmessungen den gleichen oder zumindest ähnliche Tests benutzen möchte, werden die beiden Teilmessungen meist hoch korreliert sein. Eine hohe Korrelation der Teilmessungen führt aber zu einer geringen Reliabilität der Beobachtungswertdifferenzen. Es ergibt sich die scheinbar paradoxe Situation, daß etwa ein für alle Personen konstanter Behandlungseffekt zur Reliabilität 0 führt. Das ist aber nur auf den ersten Blick überraschend, denn ein konstanter Effekt in $T_X - T_Y$ bedeutet eben, daß keine interindividuellen Unterschiede bzw. keine Populationsvarianz der Personenparameterdifferenzen vorliegt, sondern die Varianz nur auf den Fehler zurückgeht. Für die Bestimmung des Behandlungseffektes ist das optimal, eine Vorhersage interindividueller Unterschiede im Behandlungserfolg ist damit aber nicht möglich.

Für den allgemeinen Fall mit ungleichen Varianzen von X und Y und nicht vorliegender lokaler Unkorreliertheit geben Williams und Zimmerman (1977) für die Reliabilität der Differenzwerte an

$$\begin{aligned}
Rel(D) &= \frac{(\lambda Rel(X) + \lambda^{-1} Rel(Y))/2 - \rho(X, Y)}{(\lambda + \lambda^{-1})/2 - \rho(X, Y)} \\
&\quad + \frac{\rho(E_X, E_Y)\sqrt{(1 - Rel(X))(1 - Rel(Y))}}{(\lambda + \lambda^{-1})/2 - \rho(X, Y)},
\end{aligned}$$

wobei $\lambda = \sigma(X)/\sigma(Y)$. Die Formel zeigt, daß die Reliabilität der Differenzen vom Quotienten der Standardabweichungen der beiden Beobachtungswerte abhängt, nicht von den Varianzen selbst. Außerdem sieht man, daß bei korrelierten Fehlern die Reliabilität der Differenzen größer werden kann, als dies nach Formel (2.41) zu erwarten wäre. Der Grund dafür ist, daß durch die Bildung der Differenzen bei korrelierten Fehlern die Fehlerwerte sich gegenseitig aufheben.

2.6 Parallelisierbare Beobachtungswerte: Lineare Strukturgleichungsmodelle

Wir haben gesehen, daß die klassische Testtheorie nur auf Erwartungswerten, Varianzen und Kovarianzen von Beobachtungswerten aufbaut. Ihre methodische Leistung besteht darin, den Personenparameter und den Meßfehler zu identifizieren. Dies wird mit Hilfe paralleler Messungen erreicht, die bezüglich Erwartungswerten, Varianzen und Kovarianzen die echte Replikation von Beobachtungen ersetzen können. Betrachten wir nun die Kovarianz zweier Beobachtungswerte, dann ist nach (2.18)

$$\sigma(X,Y) = \sigma(T_X,T_Y) + \sigma(E_X,E_Y).$$

Sind X und Y lokal unkorreliert erhält man $\sigma(X,Y) = \sigma(T_X,T_Y)$. Betrachten wir nun statt der Kovarianz zwischen X und Y die Kovarianz zwischen einem Beobachtungswert X' und Y, wobei X' mit X und Y lokal unkorreliert sein soll und für $T_{X'}$ gilt $T_{X'} = aT_X + b$ mit festen reellen Zahlen a und b. Dann gilt

$$\begin{aligned}
\sigma(X',Y) &= \sigma(aT_X+b,T_Y)\\
&= \mathscr{E}[((aT_X+b)-\mathscr{E}(aT_X+b))(T_Y-\mathscr{E}(T_Y))]\\
&= \mathscr{E}[(aT_X+b-a\mathscr{E}(T_X)-b)(T_Y-\mathscr{E}(T_Y))]\\
&= a\,\mathscr{E}[(T_X-\mathscr{E}(T_X))(T_Y-\mathscr{E}(T_Y))]\\
&= a\,\sigma(T_X,T_Y)\\
&= a\,\sigma(X,Y).
\end{aligned}$$

Für beliebige Beobachtungswerte Y gilt also

$$a = \frac{\sigma(X',Y)}{\sigma(X,Y)}, \tag{2.42}$$

und außerdem ist

$$\sigma(X',X) = \sigma(aT_X+b,T_X) = a\sigma^2(T_X). \tag{2.43}$$

Die Transformation $X \mapsto aX+b$ wird als *affine* Transformation bezeichnet. Die obigen Überlegungen zeigen, daß affine Transformationen der Personenparameter bei lokal unkorrelierten Beobachtungswerten unter geeigneten Bedingungen identifizierbar sind. Aus (2.42) kann der multiplikative Faktor der affinen Transformation eindeutig berechnet werden. Darüber hinaus kann über (2.43) die Varianz der Personenparameter berechnet und damit sowohl die Reliabilität von X als auch die von X' bestimmt werden. Da Gleichung (2.42) für beliebige lokal unkorrelierte Beobachtungswerte Y gilt, kann sie auch als Modelltest benutzt werden. Sie verlangt, daß der Quotient der Kovarianzen von X und X' mit allen Variablen Y konstant ist.

Wir betrachten ein weiteres Beispiel von Lord und Novick (1968, S. 216 ff) mit drei affin verwandten Beobachtungswerten. Angenommen, wir haben drei Beobachtungswerte X_1, X_2, X_3, deren Personenparameter affin verwandt sind, so daß eine Zufallsvariable T_0 und Konstanten a_i und b_i existieren, mit denen gilt

$$T_{X_i} = a_iT_0 + b_i \tag{2.44}$$

für $i = 1, 2, 3$. Dann ist $\sigma^2(T_{X_i}) = a_i^2\,\sigma^2(T_0)$, und für die Beobachtungswerte X_i gilt

$$X_i = a_iT_0 + b_i + E_{X_i}.$$

Nimmt man nun an, daß die Fehlerwerte unkorreliert sind, dann erhält man mit $i \neq j$ für die Kovarianzen der Beobachtungswerte

$$\begin{aligned}
\sigma(X_i,X_j) &= \sigma(T_{X_i},T_{X_j})\\
&= a_ia_j\sigma^2(T_0)\\
&= \sigma(T_{X_i})\sigma(T_{X_j}).
\end{aligned}$$

Daraus kann man die Varianzen der Personenparameter T_{X_i} und die Gewichte a_i berechnen. Beispielsweise erhält man für $i = 1$

$$\frac{\sigma(X_1, X_2)\,\sigma(X_1, X_3)}{\sigma(X_2, X_3)} = a_1^2\sigma^2(T_0). \tag{2.45}$$

Die Parameter b_i erhält man aus

$$\begin{aligned} \mathscr{E}(X_i) &= \mathscr{E}(a_i T_0 + b_i + E_{X_i}) \\ &= a_i\mathscr{E}(T_0) + b_i, \end{aligned}$$

wobei $\mathscr{E}(T_0)$ ohne Beschränkung der Allgemeinheit auf 0 gesetzt werden kann. Nicht nur die Parameter der affinen Transformation lassen sich auf diese Weise rekonstruieren, man erhält auch Schätzungen für die Reliabilitäten der Beobachtungswerte mit

$$Rel(X_i) = \frac{\sigma^2(T_{X_i})}{\sigma^2(X_i)}.$$

Damit ist gezeigt, daß nicht nur Tests, die parallel im Sinne von Definition 2.5 sind, zur Schätzung der Reliabilität geeignet sind, sondern auch solche, deren Personenparameter wie in unserem Beispiel affin verwandt sind. Aus der Struktur der Kovarianzmatrix und aus den Mittelwerten solcher affin verwandter Tests lassen sich darüber hinaus die Parameter der affinen Transformationen schätzen, die die Tests echt parallelisieren. Beobachtungswerte, die wie in unserem Beispiel affin verwandte Personenparameter haben, werden von Jöreskog (1971) *kongenerisch* genannt. Systeme psychometrischer Daten mit kongenerischen Beobachtungswerten sind die Grundlage sogenannter „linearer Strukturgleichungsmodelle“. Solche Modelle können neben affin verwandten Personenparametern auch Beobachtungswerte enthalten, deren Personenparameter nicht perfekt korrelieren (man beachte, daß $\rho(T_X, aT_X + b) = 1$).

Die obigen Beispiele zeigen, daß durch affine Beziehungen der Personenparameter strukturelle Restriktionen auf den Varianzen und Kovarianzen von Beobachtungswerten entstehen können. Bei der empirischen Prüfung solcher Modelle wird in der Regel angenommen, daß die Meßfehler der beteiligten Variablen normalverteilt sind. In diesem Fall lassen sich Maximum-Likelihood-Schätzfunktionen für die Modellparameter bzw. für die Kovarianzen angeben. Betrachten wir das zweite oben gegebene Beispiel. Bei n Beobachtungswerten erhält man $n(n-1)/2$ Kovarianzen und $2n$ Parameter für die affinen Transformationen. Für $n = 3$ kann man, wie oben gezeigt, die Parameter berechnen, aber die Struktur der Kovarianzmatrix ist dadurch noch nicht eingeschränkt, da zur Berechnung eines Parameters a_i alle 3 Kovarianzen benötigt werden. Ist dagegen $n = 4$, so ergeben sich 6 Kovarianzen für die Berechnung der 4 Parameter $a_1, \ldots, a_4$. Damit entsteht eine strukturelle Restriktion für die Kovarianzen (2.45), die als Restriktion für die Maximum-Likelihood-Schätzung der Kovarianzen und als Test des Modells benutzt werden kann. Als Modelltest eignet sich ein Likelihoodquotiententest (Jöreskog, 1971). Dieser Test beruht auf einem Vergleich der Likelihood der Daten bei modellkonformen Parameterwerten mit der Likelihood bei freien Parameterwerten. Eine Einführung in die Methoden zur Analyse der sogenannten „linearen Strukturgleichungsmodelle“ („LISREL“) gibt Knoche (1990).

Wir werden später (Def. 2.7) zur Explikation des Konzepts „Test“ ein System psychometrischer Daten nur dann als „Test“ bezeichnen, wenn es parallele Messungen enthält. Die obigen Überlegungen und Beispiele deuten aber bereits darauf hin, daß die Parallelitätsforderung durchaus abgeschwächt werden kann. Auch affin verwandte Messungen sind zur statistischen Analyse eines Tests geeignet. Wir werden hier nicht weiter auf diese Möglichkeiten eingehen. Für eine Vertiefung dieser Methoden, die in gewisser Hinsicht auch als Modelltests benutzt werden können, sei auf Knoche (1990) verwiesen.

2.7 Statistische Parameter einzelner Testaufgaben

Theorem 2.1 zeigt, daß in den Grundkonzepten der klassischen Testtheorie einzelne Testaufgaben nicht vorkommen. Die Primitiva der Theorie sind nicht einzelne Aufgaben, sondern Tests, also Gruppen von Testaufgaben, die zum Beobachtungswert X führen. Für die Konstruktion eines Tests ist es aber in der Regel notwendig, über die inhaltliche Auswahl der Testaufgaben hinaus auch deren statistische Eigenschaften zu untersuchen. Dies geschieht vor allem im Hinblick darauf, ob sich durch eine geeignete Auswahl von Testaufgaben die Reliabilität oder die Validität eines Tests verbessern lassen. Es sei allerdings bereits an dieser Stelle darauf hingewiesen, daß sich eine befriedigende Lösung des Problems der Auswahl geeigneter Aufgaben erst im Rahmen der logistischen Testmodelle angeben läßt, da diese Modelle von den einzelnen Aufgaben als Primitiva ausgehen.

Zur Betrachtung einzelner Testaufgaben führen wir eine neue Zufallsvariable U_j ein, mit der wir die Antworten einer zufällig gezogenen Person bei einer Aufgabe j kodieren: $U_j = 1$, falls die Person die Aufgabe j richtig löst und $U_j = 0$ in allen anderen Fällen.

2.7.1 Die Schwierigkeitsstatistik

Der Erwartungswert der Zufallsvariablen U_j einer Aufgabe j wird in der klassischen Testtheorie als *Schwierigkeitsstatistik* bezeichnet:[4]

$$\pi_j = \mathscr{E}(U_j). \tag{2.46}$$

Da der Beobachtungswert X die Summe aller U_j ist, erhält man für den Erwartungswert von X

$$\begin{aligned} \mathscr{E}(X) &= \mathscr{E}\Big(\sum_{j=1}^{n} U_j\Big) \\ &= \sum_{j=1}^{n} \pi_j. \end{aligned}$$

Die Aufgabenschwierigkeit π_j wird durch die relative Lösungshäufigkeit in der Stichprobe von Probanden geschätzt und ist daher stark von der Verteilung der Testwerte in der Stichprobe abhängig.

Ist bei einzelnen Aufgaben die Wahrscheinlichkeit, eine richtige Lösung zu erraten, sehr hoch, so ist die Schwierigkeitsstatistik nicht gut zur Beschreibung der Aufgabenschwierigkeit geeignet. Man kann alternativ dazu die (fiktive) Wahrscheinlichkeit p_j dafür betrachten, daß ein Proband die richtige Lösung der Aufgabe j kennt, und diese als „Schwierigkeitsindex“ benutzen. Eine richtige Lösung der Aufgabe j kann dadurch zustande kommen, daß der Proband die Lösung kennt, oder dadurch, daß er sie nicht kennt und dafür richtig rät. Sei $1/a$ die Wahrscheinlichkeit, die richtige Lösung zu erraten, dann gilt

$$\begin{aligned} P(U_j = 1) &= p_j + (1 - p_j)\, 1/a \\ &= 1/a + p_j\,(1 - 1/a) \\ &= \pi_j. \end{aligned}$$

Nach p_j aufgelöst erhält man

$$\begin{aligned} p_j &= (\pi_j - 1/a)\,\frac{1}{1 - 1/a} \\ &= (\pi_j - 1/a)\,\frac{a}{a-1}, \end{aligned}$$

wobei π_j wiederum durch die relative Häufigkeit richtiger Lösungen von Aufgabe j geschätzt wird und $1/a$ die Ratewahrscheinlichkeit ist. Für $a \to \infty$ nähert sich p_j der Schwierigkeitsstatistik π_j an.

[4] Die eigentlich üblichen Bezeichnungen für die Schwierigkeitsstatistik und die weiter unten definierte Trennschärfestatistik sind „Schwierigkeit“ und „Trennschärfe“. Diese Bezeichnungen wollen wir aber der logistischen Testtheorie vorbehalten, da sie dort sinnvoller verwendet werden.

2.7.2 Die Trennschärfestatistik

Die Varianz der Zufallsvariablen U_j ist

$$\sigma^2(U_j) = \pi_j(1-\pi_j),$$

da U_j binomial verteilt ist. Die Varianz des Beobachtungswertes X ergibt sich aus der Varianz der einzelnen Aufgaben in folgender Weise:

$$\begin{aligned}
\sigma^2(X) &= \sigma^2\Big(\sum_{j=1}^{n} U_j\Big) \\
&= \sum_{j=1}^{n}\sum_{j'=1}^{n} \sigma(U_j, U_{j'}) \qquad \text{wegen Gl. (1.11)} \\
&= \sum_{j=1}^{n}\sum_{j'=1}^{n} \sigma(U_j)\,\sigma(U_{j'})\,\rho(U_j, U_{j'}). \qquad (2.47)
\end{aligned}$$

Dabei ist $\rho(U_j, U_{j'})$ die Produkt-Moment-Korrelation der beiden Aufgaben j und j'. Die Gesamtvarianz hängt also von der Varianz jeder einzelnen Testaufgaben ab. Diese ist maximal bei $\pi_j = 0.5$, Testaufgaben mit mittlerer Schwierigkeit tragen zur Gesamtvarianz des Testwerts am meisten bei, während Aufgaben mit extremen Schwierigkeitswerten nur wenig beitragen und daher in der Regel nicht erwünscht sind. Gleichung (2.47) zeigt, daß eine hohe Korrelation der Aufgaben untereinander die Varianz des Tests erhöht und daher in der Regel auch erwünscht ist. Eine Ausnahme liegt dann vor, wenn die Korrelation auf ein Merkmal zurückzuführen ist, das nicht das zu messende Merkmal ist.

Die Gesamtvarianz läßt sich auch in Abhängigkeit von der Korrelation einer Aufgabe mit dem Beobachtungswert betrachten:

$$\begin{aligned}
\sigma^2(X) &= \sigma\Big(\sum_{j=1}^{n} U_j, X\Big) \\
&= \sum_{j=1}^{n} \sigma(U_j, X) \\
&= \sum_{j=1}^{n} \sigma(U_j)\,\sigma(X)\,\rho(U_j, X). \qquad (2.48)
\end{aligned}$$

Die Korrelation $\rho(U_j, X)$ wird als *Trennschärfestatistik* der Aufgabe j bezeichnet. Aus (2.48) folgt

$$\sigma(X) = \sum_{j=1}^{n} \sigma(U_j)\,\rho(U_j, X). \qquad (2.49)$$

Ein hoher Wert der Trennschärfestatistik erzeugt also auch eine hohe Varianz des Beobachtungswertes. Die Trennschärfestatistik läßt sich auch als gewichtete Interkorrelation der Testaufgaben schreiben:

$$\begin{aligned}
\rho(U_j, X) &= \rho\Big(\sum_{j=1}^{n} U_j, X\Big) \\
&= \frac{1}{\sigma(X)} \sum_{j'=1}^{n} \sigma(U_{j'})\,\rho(U_j, U_{j'}). \qquad (2.50)
\end{aligned}$$

Schwierigkeits- und Trennschärfestatistik können über inhaltliche Kriterien hinaus zur Auswahl von geeigneten Testaufgaben benutzt werden. Es ist dabei allerdings zu bedenken, daß diese Parameter sehr stark von der Verteilung der Beobachtungswerte in der Stichprobe abhängen. Eine befriedigende Lösung des Selektionsproblems von Testaufgaben aufgrund ihrer statistischen Eigenschaften ist erst im Rahmen logistischer Testmodelle möglich.

2.7.3 Aufgabenvalidität

Für die Kovarianz zwischen einem Test X und einem Kriterium Y gilt

$$\begin{aligned}\sigma(X,Y) &= \sigma\Big(\sum_{j=1}^{n} U_j\Big) \\ &= \sum_{j=1}^{n} \sigma(Y)\,\sigma(U_j)\,\rho(U_j,Y). \qquad (2.51)\end{aligned}$$

Hier ist $\rho(U_j,Y)$ die *Aufgabenvalidität* der Aufgabe j. Die Validität eines Tests bezüglich des Kriteriums Y hängt von den einzelnen Aufgabenvaliditäten ab:

$$\begin{aligned}\rho(X,Y) &= \frac{\sigma(X,Y)}{\sigma(X)\,\sigma(Y)} \qquad \text{wegen (2.49) und (2.51)} \\ &= \frac{\sum_{j=1}^{n} \sigma(U_j)\,\rho(U_j,Y)}{\sum_{j=1}^{n} \sigma(U_j)\,\rho(U_j,X)}.\end{aligned}$$

Die Aufgabenvalidität bestimmt also ganz wesentlich die Gesamtvalidität eines Tests, und es ist sinnvoll, zum Erreichen einer bestimmten Validität Aufgaben auszuwählen, die eine hohe Aufgabenvalidität besitzen.

Zum Schluß dieses Abschnittes sei noch darauf hingewiesen, daß für praktische Berechnungen der Produkt-Moment-Korrelationen zwischen den Zufallsvariablen U_j und den Variablen X oder Y die Formel der *punktbiserialen Korrelation* benutzt werden kann, da die Zufallsvariablen U_j binär sind. Weitere Hinweise zur Aufgabenauswahl nach statistischen Gesichtspunkten findet man bei Lord und Novick (1968).

2.8 Wann ist ein System psychometrischer Daten ein Test?

Ob die klassische Testtheorie einen empirischen Gehalt hat, kann man an der Antwort auf die Frage prüfen, wie ein Datensatz beschaffen sein müßte, damit er die Theorie *nicht* erfüllt und ob es einen solchen Datensatz geben kann. Wenn die Theorie keinen empirischen Gehalt hätte, dann wäre sie nur eine Methode zur statistisch-mathematischen Umformung von Daten, denen über sich selbst hinaus kein Erklärungswert zukäme. Worin könnte dagegen ein möglicherweise vorhandener empirischer Gehalt bestehen? Ein solcher würde beispielsweise dann vorliegen, wenn die Personenparameter tatsächlich stabile Eigenschaften einer Person wären, mit deren Hilfe die Ergebnisse unterschiedlicher Testverfahren oder das Verhalten in Kriteriumssituationen vorhergesagt werden könnten.

Es ist zu prüfen, welche empirisch falsifizierbaren Aussagen die klassische Testtheorie enthält, beziehungsweise welche empirischen Belege bei einer Testkonstruktion vorgelegt werden müssen, um die Erfüllung der Theorie zu begründen. Wie wir gesehen haben, ist zur empirischen Bestimmung der Reliabilität mindestens die Existenz von 2 parallelen Testformen notwendig. Sind nicht mindestens 2 parallele Testformen vorhanden, dann sind alle Aussagen der klassischen Testtheorie nur Definitionen oder Datenumformungen, denn keine der Aussagen von Theorem 2.1 oder deren Folgerungen für einen einzelnen Beobachtungswert ist empirisch falsifizierbar.

Sind dagegen mindestens 2 parallele Testformen vorhanden, dann gibt es empirisch prüfbare Aussagen. Diese folgen aus den Parallelitätsbedingungen (2.20) und (2.21). Diese Bedingungen verlangen, daß parallele Tests in allen Teilpopulationen den gleichen Erwartungswert und die gleiche Varianz haben. Werden also für beide Tests Daten erhoben, kann statistisch geprüft werden, ob die Varianzen und die Erwartungswerte der Tests gleich sind. Ist dies in einer Teilpopulation nicht der Fall, dann muß die Theorie

für die erhobenen Daten verworfen werden. Wird die Nullhypothese gleicher Erwartungswerte und gleicher Varianzen beibehalten, dann ist damit keineswegs nachgewiesen, daß die beiden Tests parallel sind. Selbst wenn man von möglichen statistischen Fehlentscheidungen absieht, ist festzuhalten, daß die Gleichheit der Varianzen zweier Tests ein sehr schwaches Argument für Parallelität ist, besonders dann, wenn sie nur in wenigen, möglicherweise nur in einer Teilpopulation festgestellt wurde.

Eine weitere Prüfmöglichkeit ergibt sich aus Bedingung (2.26), sie verlangt aber die Korrelation der beiden Tests mit einem oder mehreren Außenkriterien. Hier besteht die empirische Prüfung darin, die Gleichheit der Korrelationen zu testen. Noch bessere Argumente für die Erfüllung der klassischen Testtheorie ergeben sich, wenn mehr als zwei parallele Testformen vorliegen. Dann kann neben den schon beschriebenen Tests auch die Gleichheit der Interkorrelationen (2.25) der Paralleltests empirisch geprüft werden.

Methoden zur statistischen Prüfung der Parallelität von Messungen bzw. der Parallelisierbarkeit wurden von Jöreskog (1971) vorgeschlagen. Sie beruhen auf dem Prinzip des Likelihoodquotiententests. Die Wahrscheinlichkeit der beobachteten Daten wird dabei unter zwei unterschiedlichen Hypothesen geschätzt und diese Werte dann verglichen. Die erste Schätzung erfolgt beispielsweise unter der Hypothese, daß Erwartungswerte und Varianzen der beiden Tests, deren Parallelität zu prüfen ist, unabhängig sind. In diesem Fall sind für die beiden Tests 5 Parameter aus den Daten zu bestimmen: 2 Erwartungswerte, 2 Varianzen und die Korrelation. Die zweite Schätzung erfolgt unter der Hypothese der Parallelität. In diesem Fall bleiben nur 3 freie Parameter: Erwartungswert, Varianz und Korrelation (Reliabilität), da dann beide Tests den gleichen Erwartungswert und die gleiche Varianz haben. Mit einer geeigneten Verteilungsannahme kann man nun die Wahrscheinlichkeit der beobachteten Daten unter den beiden Hypothesen berechnen. Sind diese sehr ähnlich, so heißt das, die Reduktion der Parameterzahl von 5 auf 3 führt nicht zu einer Änderung der Likelihoodfunktion und die beiden Messungen können als parallel betrachtet werden. Eine gründliche Einführung in diese Art der Überprüfung des klassischen Testsmodells findet man bei Knoche (1990).

Wir können festhalten: Die klassische Testtheorie macht empirisch wiederlegbare Aussagen über die Struktur von Testdaten, allerdings nur dann, wenn mindestens 2 parallele Testformen vorliegen. Wir fassen dies in der folgenden Definition zusammen.

Definition 2.7. Ein System $\langle \Omega, \mathfrak{A}, P, \{X_k \mid k \in I\}, H\rangle$ psychometrischer Daten mit $|I| \geq 2$ ist ein *psychometrischer Test mit linearer Struktur*, wenn mindestens zwei der Beobachtungswerte X_k parallel sind.

Diese Definition könnte aufgrund der Überlegungen in Abschnitt 2.6 abgeschwächt werden, indem man sich, statt Parallelität zu fordern, mit Parallelisierbarkeit zufrieden gibt. Lineare Strukturgleichungsmodelle sind aber weniger der angewandten Psychodiagnostik als der psychologischen Theorienbildung und der angewandten Forschung zuzuordnen. Für die praktische Anwendung im Rahmen psychodiagnostischer Entscheidungen sind nach wie vor nur solche Verfahren als wissenschaftlich begründet anzusehen, deren psychometrische Eigenschaften sich auf die Existenz paralleler Beobachtungswerte gründet.

2.9 Kritik der klassischen Testtheorie

Die Grundannahmen der klassischen Testtheorie sind auf der Zufallsvariablen X aufgebaut, die den Beobachtungswert bei einem Test beschreibt. Implizit wird angenommen, daß es sich dabei um eine reellwertige, mindestens intervallskalierte Größe handelt, andernfalls wäre beispielsweise die Definition des Fehlerwertes $E_X = X - T_X$ sinnlos. Der Beobachtungswert ist in der Regel einfach die Anzahl richtig gelöster Aufgaben in einem Test. Formal ist X damit tatsächlich reellwertig und intervallskaliert. Auch der Einwand, daß die klassische Testtheorie nicht versucht, die Messung, mit

der sie sich befaßt, auf qualitative Beobachtungen zu gründen, ist nicht gerechtfertigt, denn das Zählen richtig gelöster Aufgaben läßt sich sehr leicht qualitativ begründen.

Als Kritik an der klassischen Testtheorie werden häufig Einwände vorgebracht, die dem Beobachtungswert implizit absprechen, eine Maßzahl für die zu messende Eigenschaft zu sein. So schreibt etwa Fisseni (1990) zum Skalentyp klassischer Tests: „Die klassische Testtheorie setzt Daten auf dem Niveau einer Intervallskala voraus. Denn es werden Mittelwerte und Varianzen berechnet, es werden Differenzen von Meßwerten gebildet. Aber es ist fraglich, ob Testdaten dieses Meßniveau erreichen“ (Fisseni, 1990, S. 101).

Noch präziser formuliert Stelzl (1980) diesen Einwand: „man weiß nicht, ob der Unterschied zwischen $T_X = 5$ und $T_X = 7$ genauso groß ist, wie der Unterschied zwischen $T_X = 13$ und $T_X = 15$.“ Wie ist nun ein solcher Einwand zu interpretieren, denn wörtlich genommen ist er offensichtlich unsinnig, da in beiden Fällen die Differenz eben 2 beträgt. Gemeint ist damit wohl eher die Frage, ob der Unterschied in der der Testbearbeitung zugrundeliegenden Fähigkeit zwischen Personen, mit $T_X = 5$ und $T_X = 7$ genauso groß ist, wie der zwischen Personen mit $T_X = 13$ und $T_X = 15$. Auch so ist aber die Frage nur dann sinnvoll, wenn T_X nicht selbst eine Maßzahl für diese Fähigkeit ist. Genau das wird aber von der klassischen Testtheorie behauptet.

Das mit dieser Behauptung verbundene Problem zeigt sich bei Aussagen über psychologische Größen, die mit Hilfe psychologischer Tests konstruiert wurden. Eine Aussage wie

> die Korrelation zwischen *Intelligenz* und dem Durchschnitt der Schulnoten beträgt 0.4,

enthält ja keinen Bezug zu einem bestimmten Meßverfahren für „Intelligenz“, unterstellt also implizit, daß die Korrelation von Intelligenz mit anderen Variablen unabhängig von dem speziellen Meßverfahren für Intelligenz berechnet und interpretiert werden kann. Durch die indirekte Beobachtungsweise über die Anzahl der richtigen Lösungen in einem Test entsteht daher das Problem, den Einfluß, den das Meßverfahren selbst auf die Beobachtungsdaten hat, zu isolieren. „Intelligenz“ kann ja nicht direkt beobachtet werden, auch ein ordinaler Paarvergleich zwischen Personen, wie er etwa bei sportlichen Leistungen durchaus üblich ist, kann zur Messung der Intelligenz in der Regel nicht durchgeführt werden. Stattdessen werden Testverfahren benutzt, die aus unterschiedlich schwierigen Aufgaben bestehen. Die Anzahl der richtig gelösten Aufgaben ist dann zwar ein problemlos zu erhebendes Datum, es ist aber klar, daß die Anzahl gelöster Aufgaben in der Regel nicht die zu messende Größe sein wird. Man kann dies sehr schön an einem Beispiel von Fischer (1974) sehen:

Angenommen es soll untersucht werden, wie eine bestimmte Trainingsmethode die Leistungsfähigkeit von Kindern mit unterschiedlichem Anfangsniveau verändert. Es werden zwei Gruppen von Kindern gebildet, solche mit niedrigem und solche mit höherem Anfangsniveau, von beiden wird vor dem Training ein Testwert erhoben. Nach der Durchführung des Trainings wird ein zweiter Testwert erhoben, und es wird festgestellt, daß die Kinder mit niedrigem Anfangsniveau einen durchschnittlich höheren Zuwachs in der Anzahl der gelösten Testaufgaben erzielen als die Kinder mit höherem Anfangsniveau. Das Training hat also zumindest bei den Punktwerten die Kinder mit niedrigem Anfangsniveau im Durchschnitt stärker gefördert als die mit höherem. Das bedeutet aber nicht, daß diese Aussage auch für die latente Leistungsfähigkeit gilt. Es könnte nämlich sein, daß die Gruppe mit hohem Anfangsniveau bereits im ersten Test vor dem Training fast alle Aufgaben richtig löst, so daß für einen Leistungszuwachs im oberen Schwierigkeitsbereich zu wenig Aufgaben zur Verfügung stehen. Würde man also eine Aufgabensammlung mit höherem durchschnittlichen Schwierigkeitsgrad wählen, könnte das Ergebnis im Zuwachs der Anzahl richtig gelöster Aufgaben ganz anders aussehen.

Mit den Methoden der klassischen Testtheorie ist dieses Problem nicht wirklich zu lösen. Der Grund ist, daß in der klassischen Testtheorie inner-

halb eines Tests nicht versucht wird, den Einfluß der Aufgaben vom Einfluß der Person auf die direkt beobachtbaren Lösungsanzahlen zu separieren. Die „Lösung“ der klassischen Testtheorie für das Problem, daß die Messung nur indirekt ist und deshalb auch die Testaufgaben die beobachtbaren Größen beeinflussen, besteht darin, nicht einzelne Aufgaben, sondern den Test nur als Ganzes zu betrachten. Sie verlangt, daß für die Messung einer psychologischen Größe immer der „gleiche“ Test benutzt wird, wobei Tests dann als „gleich“ betrachtet werden, wenn sie parallel sind, also prinzipiell für jede Person das gleiche Ergebnis liefern.

Gibt man sich mit einer möglichst hohen Korrelation zwischen Testwert und Außenkriterium als Erfolgskriterium zufrieden, so kann man mit der klassischen Testtheorie zufriedenstellende Ergebnisse erreichen, obwohl auch da paradoxe Effekte auftreten können (Fischer, 1974). Als psychologische Theorie zur Erklärung des Zusammenhangs zwischen dem Verhalten in der Testsituation und dem bei der Bewährung am Außenkriterium ist die klassische Testtheorie nicht zufriedenstellend. Dies liegt, wie wir gesehen haben, im wesentlichen an zwei Punkten: erstens ist das Primitivum der klassischen Testtheorie nicht die einzelne Testaufgabe, sondern das Gesamtergebnis im Test, und zweitens ist die Annahme eines linearen Zusammenhangs zwischen Testleistung und gemessener psychologischer Eigenschaft, wie er in der Beziehung $\mathscr{E}(X \,|\, T_X = \tau) = \tau$ zum Ausdruck kommt, aus psychologischer Sicht zu restriktiv, aber gleichzeitig im Rahmen der klassischen Testtheorie nur schwer zu widerlegen.

3 Logistische Testmodelle

Logistische Testmodelle unterscheiden sich in 3 wesentlichen Punkten von der klassischen Testtheorie: 1. wird zwischen den Personenparametern und den beobachtbaren Variablen kein linearer, sondern ein logistischer Zusammenhang angenommen, 2. wird die theoretische Analyse nicht auf dem Testrohwert als Primitivum, sondern auf dem Ergebnis bei einzelnen Testaufgaben aufgebaut, und 3. wird der Einfluß der Testaufgaben auf das Lösungsverhalten vom Einfluß der Personenfähigkeit abgetrennt. Dieser letzte Punkt ist für die strukturellen Eigenschaften der Testmodelle von größter Bedeutung. Er ist im wesentlichen dafür verantwortlich, daß diese Modelle „spezifisch objektive“ Aussagen über die Personen zulassen. Man versteht darunter Aussagen, die Relationen zwischen Personen herstellen, die unabhängig von den speziellen Testaufgaben gelten. Aussagen mit dieser Qualität sind im Rahmen der klassischen Testtheorie für einen einzelnen Test grundsätzlich unmöglich, da in der Definition eines Systems psychometrischer Daten (Def. 2.1) Testaufgaben nicht vorkommen. Das theoretische Primitivum der klassischen Testtheorie ist der Test als Ganzes. Mit den logistischen Testmodellen vergleichbare, objektive Aussagen setzen im Rahmen der klassischen Testtheorie eine größere Familie paralleler oder parallelisierbarer Tests voraus (Steyer & Eid, 1993), während solche Aussagen im Rahmen der logistischen Testmodelle bereits auf der Basis eines einzelnen Tests möglich sind.

Da Testaufgaben und ihre individuellen Eigenschaften im Konzept der klassischen Testtheorie nicht vorkommen, ist auch das Konzept der Trennschärfestatistik nur ein nachträglich eingeführter, deskriptiver Parameter, der für den Aufbau der Theorie keinerlei Bedeutung hat. Im Gegensatz dazu ist der Einfluß der Testaufgaben auf die Lösungswahrscheinlichkeit expliziter Bestandteil der logistischen Testmodelle. Dies erlaubt die Trennung von Aufgaben- und Personeneinfluß und als Folge davon Aussagen, die sich nur auf Personen bzw. nur auf Aufgaben beziehen. Diese Trennung von Aufgaben- und Personeneinfluß auf die Lösungswahrscheinlichkeit ist kein statistischer Kunstgriff, sondern setzt bestimmte strukturelle Eigenschaften der Lösungswahrscheinlichkeiten voraus. Sind diese in einem vorliegenden Datensatz nicht gegeben, so ist das entsprechende Testmodell auf diesen Datensatz nicht anwendbar. Im Gegensatz zur klassischen Testtheorie sind daher die logistischen Testmodelle mit Hilfe eines einzigen Datensatzes empirisch testbar. Der Test besteht in der empirischen Überprüfung der für die Trennbarkeit von Personen- und Aufgabeneinfluß notwendigen strukturellen Eigenschaften des Systems der Lösungswahrscheinlichkeiten.

Die Entwicklung der logistischen Testmodelle basiert auf den Arbeiten von Rasch (1960) und Birnbaum (1968). Gute Einführungen enthalten die Monographien von Fischer (1974), Hamerle (1982) und Rost (1988). Aus mathematischer Sicht besonders gründlich ist die Darstellung von Lind (1994).

3.1 Die Datenmatrix logistischer Testmodelle

Der Wahrscheinlichkeitsraum, von dem logistische Testmodelle ausgehen, hat als Ergebnisraum nicht eine Kombination aus Personen und Testergebnissen, sondern aus Personen und Antwortvektoren, in denen die Antworten auf jedes einzelne Item kodiert sind. Wir bezeichnen im folgenden die Personen durch die Indizes $i = 1, \ldots, m$ und die Aufgaben durch die Indizes

$j = 1, \ldots, n$. Die Antworten der Personen werden durch Zufallsvariable U_{ij} kodiert:

$$U_{ij} = \begin{cases} 1 & \text{falls Person } i \text{ die Aufgabe } j \text{ richtig löst,} \\ 0 & \text{sonst.} \end{cases}$$

Das Ergebnis der Datenerhebung ist dann eine $m \times n$-Matrix, deren Eintragungen aus den Werten 0 oder 1 besteht. Diese Matrix wird als *Datenmatrix* bezeichnet. Die Datenmatrix enthält alle Daten, sie ist die empirische Grundlage aller statistischen Überlegungen.

Das „lineare Modell" der klassischen Testtheorie kommt in der Gleichung (2.10) zum Ausdruck. Sie gibt den angenommenen Zusammenhang zwischen Personenparameter und Beobachtungswert in einem Test an. Unterschiedliche, aber parallele bzw. parallelisierbare Tests werden als affin verwandt angenommen, so daß man als etwas allgemeineres „lineares Modell" der klassischen Testtheorie die Form

$$\mathscr{E}(X \mid T_0 = \tau; a_X, b_X) = a_X T_0 + b_X \tag{3.1}$$

betrachten kann. Hier ist T_0 ein Basisparameter für die zu untersuchende Fähigkeit mit dem die Personenparameter affin verwandter Tests über die angegebene Transformation zusammenhängen.

Wie bereits oben erwähnt, gehen logistische Modelle von der binären Zufallsvariablen U_{ij} aus. Zwischen dem Erwartungswert der Zufallsvariablen U_{ij} und dem Parameter der Person wird eine logistische Funktion als Zusammenhang angenommen:

$$\mathscr{E}(U_{ij} \mid \theta_i, j) = \frac{\exp[f_j(\theta_i)]}{1 + \exp[f_j(\theta_i)]}. \tag{3.2}$$

Mit θ_i wird hier üblicherweise der Personenparameter der Person i bezeichnet. Die Funktion $f_j()$ ist immer eine affine Transformation der Form $\theta_i \mapsto (a_j\theta_i + b_j)$. Sie gibt an, wie die Aufgabe j den Personenparameter in die logistische Funktion einbringt. Ein Vergleich der Ausdrücke (3.1) und (3.2) zeigt deutlich den oben erwähnten Unterschied in den Annahmen über den Zusammenhang zwischen Beobachtungswerten und Personenparametern. Welche dieser beiden Annahmen besser ist, kann nicht ohne weiteres entschieden werden, da dies offensichtlich sehr stark von der Art der Testaufgaben abhängen wird, vor allem davon, wie die Verteilung der Personenfähigkeiten in den Extrembereichen der Population aussieht. In mittleren Bereichen ist die logistische Funktion einer Geraden sehr ähnlich, so daß sich dort nur wenig Unterschiede zwischen den beiden Gleichungen zeigen.

Die formale Ähnlichkeit der Gleichungen (3.1) und (3.2) wird von Steyer & Eid (1993) zum Anlaß genommen, die logistischen Testmodelle als bedingte Erwartungen der Zufallsvariablen U einzuführen, da formal betrachtet $\mathscr{E}(U_{ij}) = P(U_{ij} = 1)$ ist. Hinter dieser Beziehung steckt aber kein empirisch bedeutsamer Sachverhalt, da die Zufallsvariable U nur ein nominaler Indikator ist. Die Zuweisung der Zahlen 0 und 1 zu den möglichen Werten von U ist vollkommen willkürlich. Werden andere Zahlen gewählt, so gilt die Übereinstimmung zwischen Antwortwahrscheinlichkeit und Erwartungswert nicht mehr.

Ein entscheidender Fortschritt der logistischen Modelle gegenüber dem klassischen Testmodell ergibt sich daraus, daß diese Modelle die Bearbeitung einzelner Aufgaben in den Mittelpunkt der Überlegungen stellen. Dadurch wird eine wesentlich genauere Analyse der formalen Struktur eines Tests möglich. Insbesondere ergeben sich daraus auch Möglichkeiten, dessen formale Struktur empirisch zu prüfen, ohne daß dazu parallele Beobachtungswerte erforderlich sind.

Verschiedene logistische Testmodelle unterscheiden sich durch unterschiedliche Restriktionen für die Funktion $f_j(\theta)$ in (3.2). Wir werden hier nur zwei Modelle ausführlicher betrachten: das *1-parametrige logistische Modell*: $f_j(\theta) = \theta - b_j$ und das *2-parametrige logistische Modell*: $f_j(\theta) = a_j\theta - b_j$. Das 1-parametrige logistische Testmodell wurde von Rasch (1960) vorgeschlagen und ausführlich untersucht. Wir werden es deshalb im folgenden

als *Rasch-Modell* bezeichnen. Das 2-parametrige logistische Testmodell geht auf Birnbaum (1968) zurück und wird als *Birnbaum-Modell* bezeichnet.

Die Darstellung der logistischen Modelle als Aussagen über die bedingten Erwartungswerte der Beobachtungsvariablen U_{ij} zeigt die formale Verwandtschaft dieser Modelle zu den Vorstellungen der klassischen Testtheorie. Für die folgenden Überlegungen werden wir die Modelle aber nicht primär als Aussagen über bedingte Erwartungen auffassen, sondern als Aussagen über die Lösungswahrscheinlichkeiten einzelner Aufgaben durch eine Person. Üblicherweise betrachtet man bei den logistischen Testmodellen die Lösungswahrscheinlichkeit einer zufällig ausgewählten Person, ohne auf die Verteilung der Personenparameter in der Population einzugehen. Dies ist möglich, weil die logistischen Testmodelle im Gegensatz zur klassischen Testtheorie die Bestimmung der Parameter erlauben, ohne deren Verteilung in der Population zu kennen. Man nennt diese Eigenschaft *Populationsunabhängigkeit*. Für die Formulierung der Modelle bedeutet dies, daß man von der bedingten Wahrscheinlichkeit ausgeht, mit der eine Person mit dem Parameter θ_i eine vorgegebene Testaufgabe löst. Da man sich grundsätzlich auch die Testaufgaben aus einer „Population“ von Testaufgaben ausgewählt denken kann, wird das beobachtete Ereignis „Lösung der Aufgabe j durch die Person i“ sowohl auf den Personen- als auch auf die Aufgabenparameter bedingt. Wir betrachten also

$$P(U_{ij} = u_{ij} \,|\, \theta_i, \Lambda_j),$$

die Wahrscheinlicheit dafür, daß eine Person mit dem Parameter θ_i bei einer Aufgabe mit dem Parameter Λ_j das Ergebnis u_{ij} erzielt. Der Aufgabenparameter Λ_j kann mehrdimensional sein. Wir betrachten hier allerdings nur die Fälle, in denen Λ_j eindimensional ($\Lambda_j = \varepsilon_j$ beim Rasch-Modell) oder zweidimensional ($\Lambda_j = (\alpha_j, \varepsilon_j)$ beim Birnbaum-Modell) ist. Für das allgemeine logistische Modell erhalten wir

$$P(U_{ij} = u_{ij} \,|\, \theta_i, \Lambda_j) = \frac{\exp[f_j(\theta_i)]^{u_{ij}}}{1 + \exp[f_j(\theta_i)]}. \tag{3.3}$$

Zur Vereinfachung der Schreibweise werden wir im folgenden die Indizes i und j bei den Parametern immer dann weglassen, wenn auf die Vektoren der Parameter aller Personen bzw. aller Aufgaben Bezug genommen werden soll. In diesem Fall verwenden wir Großbuchstaben, also $\Theta = (\theta_1, \ldots, \theta_i, \ldots, \theta_m)$, $\Lambda = (\lambda_1, \ldots, \lambda_j, \ldots, \lambda_n)$ und $E = (\varepsilon_1, \ldots, \varepsilon_j, \ldots, \varepsilon_n)$. Für den Datenvektor $(U_{i1}, \ldots, U_{in})$ einer Person i schreiben wir $\mathbf{U}_i$, und für die gesamte Datenmatrix $(U_{11}, \ldots, U_{1n}, \ldots, U_{mn})$ schreiben wir $\mathbf{U}$. Entsprechend verwenden wir die Bezeichnungen $\mathbf{u}_i$ und $\mathbf{u}$ für die Realisationen der Zufallsvektoren $\mathbf{U}_i$ und $\mathbf{U}$.

3.2 Lokale stochastische Unabhängigkeit

In der logistischen Testtheorie wird angenommen, daß die Beantwortung der einzelnen Aufgaben durch eine Person stochastisch unabhängig ist von den Antworten der gleichen Person bei den anderen Aufgaben. Diese Annahme wird *lokale stochastische Unabhängigkeit* (vgl. Def. 1.23) genannt. Sie gestattet eine Zerlegung der Verbundwahrscheinlichkeit für einen bestimmten Antwortvektor einer Person in ein Produkt der Antwortwahrscheinlichkeiten für die einzelnen Aufgaben. Damit gilt dann für jede Person i

$$P(\mathbf{U}_i = \mathbf{u}_i \,|\, \theta_i, \Lambda) = \prod_{j=1}^{n} P(U_{ij} = u_{ij} \,|\, \theta_i, \Lambda_j). \tag{3.4}$$

Die lokale stochastische Unabhängigkeit ist analog zur Forderung lokal unkorrelierter Fehler in der klassischen Testtheorie. Man kommt dort mit der etwas schwächeren Forderung der Unkorreliertheit aus, weil ja immer nur die ersten oder zweiten Momente von Verteilungen betrachtet werden. Lokale stochastische Unabhängigkeit impliziert lokal unkorrelierte Fehler, aber nicht umgekehrt.

Es wird natürlich auch angenommen, daß die Antwortvektoren unterschiedlicher Personen stochastisch unabhängig sind. Die empirische Gültigkeit dieser Bedingung kann durch geeignete Testvorgaben gesichert werden, da sie nur dann verletzt ist, wenn die Antworten einer Person von den Antworten einer anderen Person beeinflußt werden. Sind die Antworten einer Person unabhängig von denen anderer Personen, dann läßt sich die Verbundwahrscheinlichkeit für das Auftreten einer bestimmten Datenmatrix als Produkt der Wahrscheinlichkeiten für das Auftreten der einzelnen Zeilen schreiben. Es gilt also

$$\begin{aligned} P(\mathbf{U} = \mathbf{u} \,|\, \Theta, \Lambda) &= \prod_{i=1}^{m} P(\mathbf{U}_i = \mathbf{u}_i \,|\, \theta_i, \Lambda) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{n} P(U_{ij} = u_{ij} \,|\, \theta_i, \Lambda_j) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{n} \frac{\exp[u_{ij} f_j(\theta_i)]}{1 + \exp[f_j(\theta_i)]}. \end{aligned} \tag{3.5}$$

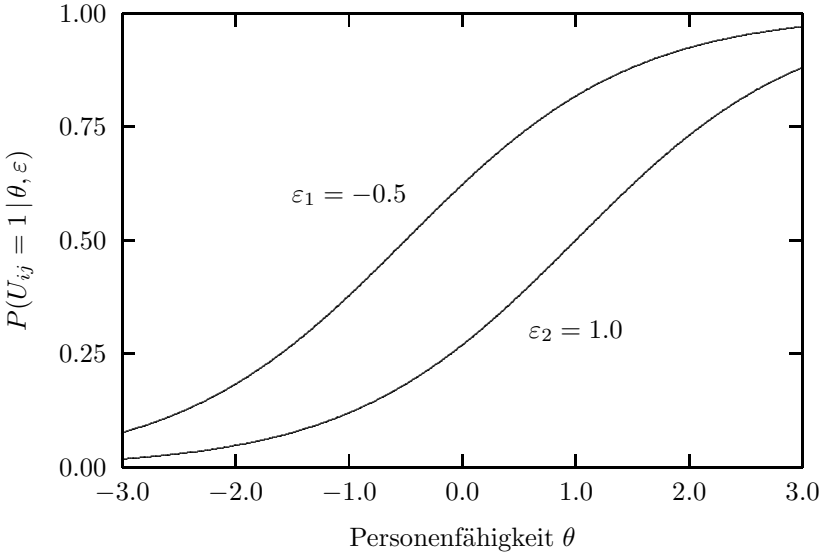

Abb. 3.1. Aufgabenkennlinien zweier Testaufgaben nach dem Rasch-Modell. Für jeden konstanten Funktionswert p mit der Eigenschaft $p = P(U_{ij} = 1 \,|\, \theta, \varepsilon_j) = P(U_{ij} = 1 \,|\, \theta', \varepsilon_k)$ gilt, daß die Differenz $\theta - \theta'$ der Personenparameter, die diesen Funktionswert erzeugen, konstant ist. Alle Aufgabenkennlinien können daher aus einer einzigen Kurve durch eine Verschiebung entlang der Abszisse erzeugt werden. Der horizontale Abstand der Kennlinien ist identisch mit der Differenz der beiden Aufgabenparameter.

3.3 Das Rasch-Modell

Alle bisherigen Annahmen bezogen sich auf die Unabhängigkeit beobachtbarer Ereignisse. Die nächste (und wichtigste) Annahme des Testmodells von Rasch (1960) kann ebenfalls als Unabhängigkeitsannahme aufgefaßt werden und hat auch formal sehr viel Ähnlichkeit mit den Annahmen über die stochastische Unabhängigkeit. Diese Annahme sagt, daß sich die Wahrscheinlichkeit, daß Person i Aufgabe j mit u_{ij} beantwortet, in eine logistische Funktion von zwei subtraktiv verknüpften Parametern θ_i und ε_j zerlegen läßt

$$P(U_{ij} = u_{ij} \,|\, \theta_i, \varepsilon_j) = \frac{\exp[u_{ij}(\theta_i - \varepsilon_j)]}{1 + \exp(\theta_i - \varepsilon_j)}. \tag{3.6}$$

Setzt man Gleichung (3.6) in Gleichung (3.4) ein und definiert

$$r_i = \sum_{j=1}^{n} u_{ij}$$

und

$$s_j = \sum_{i=1}^{m} u_{ij},$$

dann gilt für den Datenvektor der Person i

$$\begin{aligned} P(\mathbf{U}_i = \mathbf{u}_i \,|\, \theta_i, E) &= \prod_{j=1}^{n} \frac{\exp[u_{ij}(\theta_i - \varepsilon_j)]}{1 + \exp(\theta_i - \varepsilon_j)} \\ &= \frac{\exp\Big(\sum_{j=1}^{n} u_{ij}\theta_i - \sum_{j=1}^{n} u_{ij}\varepsilon_j\Big)}{\prod_{j=1}^{n}[1 + \exp(\theta_i - \varepsilon_j)]} \\ &= \frac{\exp\Big(\theta_i \sum_{j=1}^{n} u_{ij} - \sum_{j=1}^{n} u_{ij}\varepsilon_j\Big)}{\prod_{j=1}^{n}[1 + \exp(\theta_i - \varepsilon_j)]} \\ &= \frac{\exp\Big(\theta_i r_i - \sum_{j=1}^{n} u_{ij}\varepsilon_j\Big)}{\prod_{j=1}^{n}[1 + \exp(\theta_i - \varepsilon_j)]} \\ &= \frac{\exp(r_i\theta_i)\,\exp\Big(-\sum_{j=1}^{n} u_{ij}\varepsilon_j\Big)}{\prod_{j=1}^{n}[1 + \exp(\theta_i - \varepsilon_j)]}. \end{aligned} \tag{3.7}$$

Abbildung 3.1 zeigt Beispiele für die Funktionen (3.6) mit konstantem Aufgabenparameter ε in Abhängigkeit von der Personenfähigkeit θ. Wesentliches Merkmal dieser *Aufgabenkennlinie* ist, daß sie sich nur durch eine horizontale Verschiebung unterscheiden.

3.4 Das Birnbaum-Modell

Eine etwas allgemeinere Form der Dekomponierbarkeit der Wahrscheinlichkeit dafür, daß Person i Aufgabe j mit u_{ij} beantwortet, wurde von Birnbaum (1968) vorgeschlagen. Zusätzlich zu den Parametern θ_i und ε_j wird dabei den Testaufgaben ein weiterer Parameter α_j zugeordnet, so daß sich ergibt

$$P(U_{ij} = u_{ij} \,|\, \theta_i, \alpha_j, \varepsilon_j) = \frac{\exp[u_{ij}\alpha_j(\theta_i - \varepsilon_j)]}{1 + \exp[\alpha_j(\theta_i - \varepsilon_j)]}. \tag{3.8}$$

Analog zu Gleichung (3.7) erhält man mit dem Birnbaum-Modell

$$\begin{aligned} P(\mathbf{U}_i = \mathbf{u}_i \,|\, \theta_i, \Lambda) &= \prod_{j=1}^{n} \frac{\exp[u_{ij}\alpha_j(\theta_i - \varepsilon_j)]}{1 + \exp[\alpha_j(\theta_i - \varepsilon_j)]} \\ &= \frac{\exp\Big(\theta_i \sum_{j=1}^{n} \alpha_j u_{ij}\Big)\,\exp\Big(-\sum_{j=1}^{n} u_{ij}\alpha_j\varepsilon_j\Big)}{\prod_{j=1}^{n}(1 + \exp[\alpha_j(\theta_i - \varepsilon_j)])}. \end{aligned} \tag{3.9}$$

Dies zeigt, daß die Parameter α_j als Gewichtungsparameter der Daten fungieren. Im Gegensatz zum Rasch-Modell, wo zur Berechnung von r_i und s_j einfach die Antwortindikatoren u_{ij} aufsummiert wurden, müssen diese beim Birnbaum-Modell mit dem Parameter α_j gewichtet werden. Auf die Probleme, die dadurch entstehen, werden wir später eingehen.

Abbildung 3.2 zeigt Beispiele für die Funktionen (3.8) mit konstanten Aufgabenparametern α und ε in Abhängigkeit von der Personenfähigkeit θ. Die unterschiedlichen Trennschärfeparameter führen dazu, daß sich die einzelnen Aufgabenkennlinien in der Steigung unterscheiden können.

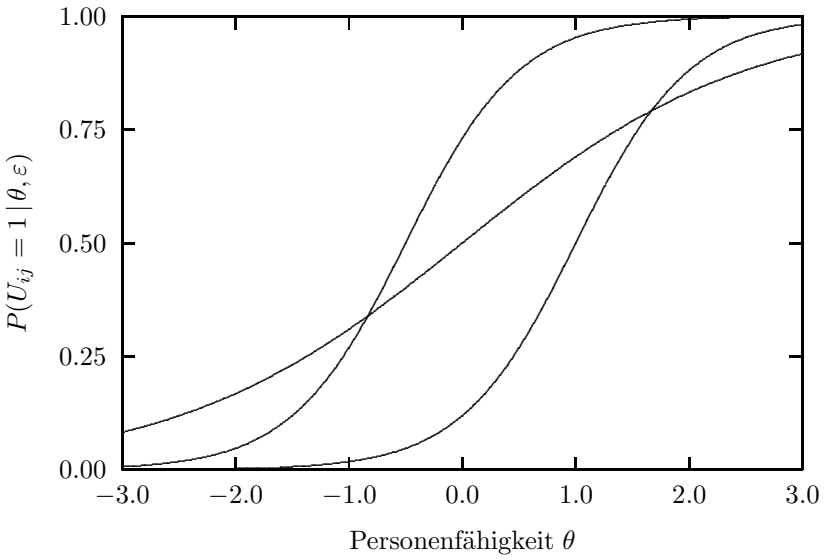

Abb. 3.2. Aufgabenkennlinien von 3 Testaufgaben nach dem Birnbaum-Modell. Im Gegensatz zu den Kennlinien beim Rasch-Modell können sich die Items hier in der Steigung unterscheiden. Je größer der Trennschärfeparameter, desto steiler verlaufen die Kurven am Wendepunkt. Die Parameterwerte der drei Items sind: $(\alpha, \varepsilon) = (0.8, 0.0)$, $(2.0, 1.0)$, $(2.0, -0.5)$.

3.5 Statistische Eigenschaften

Die Annahme, daß die Wahrscheinlichkeiten $P(U_{ij} = u_{ij})$ in der Form von Gleichung (3.6) dekomponiert werden können, erscheint im ersten Moment willkürlich. In Abschnitt 3.3 wurde dafür auch keine besondere Begründung angegeben. Es läßt sich aber zeigen, daß die Form von Gleichung (3.6) auch aus Annahmen über die speziellen statistischen Eigenschaften des Wahrscheinlichkeitsraums der Testkonstruktion abgeleitet werden können. Die Vorteile dieser statistischen Annahmen sind leicht einsehbar und können allgemein als notwendige Eigenschaften psychologischer Skalierungsverfahren betrachtet werden.

Wir werden hier auf zwei unterschiedliche, theoretische Begründungen des Rasch-Modells eingehen. Im folgenden Abschnitt betrachten wir die statistische Fundierung, wie sie im Detail von Andersen (1973a) ausgearbeitet wurde. In Abschnitt 3.8 werden wir dann ein Begründung des Rasch-Modells betrachten, bei der meßtheoretische Argumente im Vordergrund stehen.

3.5.1 Suffiziente Statistiken

Wie bereits Gleichung (3.6) andeutet, wird von Rasch (1960) angenommen, daß es für jede Aufgabe j einen skalaren Parameter ε_j und für jede Person i einen ebenfalls skalaren Parameter θ_i gibt, so daß sich die Wahrscheinlichkeitsfunktion für U_{ij} als in den beiden Argumenten θ_i und ε_j streng monotone Funktionen schreiben läßt:

$$P(U_{ij} = u_{ij} \mid \theta_i, \varepsilon_j) = f(u_{ij} \mid \theta_i, \varepsilon_j). \tag{3.10}$$

Selbst wenn über die Funktion f keine weitergehenden Annahmen gemacht werden, ist dies eine sehr starke Forderung, da sie erzwingt, daß ein einziger Personen- und Aufgabenparameter ausreicht, um den Einfluß der Person und der Aufgabe auf die Lösungswahrscheinlichkeit zu beschreiben. Mit dieser Annahme ist ausgeschlossen, daß sich die Aufgaben in der Trennschärfe unterscheiden, wie dies etwa im Birnbaum-Modell der Fall ist. Eine empirische Prüfung dieser Annahme kann mit Hilfe der von Irtel und Schmalhofer (1982) vorgeschlagenen Methoden durchgeführt werden. Von diesen Autoren wird gezeigt, daß sich aus den Annahmen der lokalen stochastischen Unabhängigkeit (3.4) und der monotonen Dekomponierbarkeit (3.10) empirisch prüfbare Bedingungen, nämlich die sogenannte *Ordnungsunabhängigkeit*, ableiten lassen. Außerdem ist aus der Meßtheorie bekannt, daß für stochastisch unabhängige Zufallsvariable Ordnungsunabhängigkeit der Wahrscheinlichkeitsfunktionen und die Existenz der Parameter θ_i und ε_j äquivalent sind (Fishburn, 1973). Die Parameter sind allerdings in diesem Fall nur eindeutig bis auf streng monotone Transformationen.

Um Parameter mit besseren Eindeutigkeitseigenschaften zu erhalten, ist noch eine dritte Annahme über die Wahrscheinlichkeitsfunktionen der Zufallsvariablen U_{ij} notwendig: Es wird angenommen, daß für jedes θ_i eine minimal suffiziente Statistik $T_i = t(\mathbf{U}_i)$ existiert, die unabhängig von den Parametern $\varepsilon_1, \ldots \varepsilon_n$ ist. Zur Prüfung, ob eine Statistik T suffizient für einen Parameter θ ist, wird häufig die *Faktorisierungsbedingung* benutzt (Hogg & Craig, 1978). Aus ihr folgt, daß eine Statistik $T = t(X_1, \ldots, X_n)$ genau dann suffizient für den Parameter θ ist, wenn zwei nichtnegative Funktionen f_1 und f_2 existieren, so daß

$$f(x_1, \ldots, x_n \,|\, \theta) = f_1[t(x_1, \ldots, x_n) \,|\, \theta]\, f_2(x_1, \ldots, x_n)$$

und dabei für jedes feste $t(x_1, \ldots, x_n)$ die Funktion $f_2(x_1, \ldots, x_n)$ nicht von θ abhängt. Diese Gleichung zeigt sehr schön, daß in der Funktion f_1 alle Information über θ steckt, die man braucht, um von der unbedingten Likelihoodfunktion f_2 der Daten auf die auf θ bedingte Likelihoodfunktion überzugehen.

Um diese Bedingung auf das Rasch-Modell anzuwenden, betrachtet man Gleichung (3.7). Diese läßt sich faktorisieren, so daß

$$f(\mathbf{u}_i \,|\, \theta_i, E) = f_1[t(\mathbf{u}_i) \,|\, \theta_i, E]\, f_2(\mathbf{u}_i \,|\, E),$$

wobei

$$f_1[t(\mathbf{u}_i) \,|\, \theta_i, E] = \frac{\exp(r_i \theta_i)}{\prod_{j=1}^{n} [1 + \exp(\theta_i - \varepsilon_j)]}$$

und

$$f_2(\mathbf{u}_i \,|\, E) = \exp\Big(-\sum_{j=1}^{n} u_{ij} \varepsilon_j\Big).$$

Damit ist $t(\mathbf{U}_i) = r_i$ eine suffiziente Statistik für θ_i. Bei der Anwendung der Faktorisierungsbedingung ist zu beachten, daß die Funktionen f_1 und f_2 *keine* Wahrscheinlichkeitsfunktionen sein müssen. Man sieht auch sofort, daß dies für f_2 zutrifft.

Analog zum Rasch-Modell läßt sich auch bei Gültigkeit des Birnbaum-Modells die Wahrscheinlichkeitsfunktion (3.9) des Datenvektors einer Person faktorisieren. Man erhält in diesem Fall

$$f_1[t(\mathbf{u}_i, A) \,|\, \theta_i, E, A] = \frac{\exp\Big(\theta_i \sum_{j=1}^{n} \alpha_j u_{ij}\Big)}{\prod_{j=1}^{n} [1 + \exp(\alpha_j(\theta_i - \varepsilon_j))]}$$

und

$$f_2(\mathbf{u}_i \,|\, E, A) = \exp\Big(-\sum_{j=1}^{n} u_{ij} \alpha_j \varepsilon_j\Big)$$

Damit ist

$$t(\mathbf{U}_i, A) = \sum_{j=1}^{n} \alpha_j U_{ij} \tag{3.11}$$

eine suffiziente Statistik für θ_i. Die Formel zeigt bereits das damit verbundene Schätzproblem auf: Die suffiziente Statistik hängt von den Gewichten α_j ab. Sie erfüllt damit nicht die Bedingung, daß die suffiziente Statistik für θ_i nicht von den Aufgaben abhängen darf. Die Schätzung der Personenparameter θ_i setzt im Birnbaum-Modell die Kenntnis der Aufgabenparameter α_j voraus. Für die Testanwendung ist dies unproblematisch, da bei der Anwendung die Aufgabenparameter in jedem Fall bekannt sind. Probleme ergeben sich bei der Schätzung der Aufgabenparameter, also bei der Testkonstruktion, da bei Anwendung des Birnbaum-Modells, im Gegensatz zu der des Rasch-Modells, immer eine gleichzeitige Schätzung von Personen- und Aufgabenparameter durchgeführt werden muß.

3.5.2 Das Theorem von Andersen (1973a) als Begründung des Rasch-Modells

Die Annahme von Andersen (1973a), daß für jedes θ_i in Gleichung (3.10) eine minimal suffiziente Statistik existiert, läßt sich also aufgrund der Faktorisierungsbedingung dadurch ausdrücken, daß die Existenz zweier Funktionen f_1 und f_2 verlangt wird, für die gilt

$$f(\mathbf{u}_i \,|\, \theta_i, \varepsilon_1, \ldots, \varepsilon_n) = f_1(T_i \,|\, \theta_i, \varepsilon_1, \ldots, \varepsilon_n)\, f_2(\mathbf{u}_i \,|\, \varepsilon_1, \ldots, \varepsilon_n). \tag{3.12}$$

Für die Statistik T_i muß darüber hinaus gelten, daß

$$t(\mathbf{U}_i) = t[\pi(\mathbf{U}_i)], \tag{3.13}$$

wobei $\pi(\mathbf{U}_i)$ eine Permutation von $\mathbf{U}_i$ ist.

Bedingung (3.12) entspricht der Faktorisierungsbedingung für suffiziente Statistiken, und Gleichung (3.13) bedeutet, daß die Statistik T_i symmetrisch in ihren Argumenten ist. Diese Symmetrie in den Argumenten ist eine Eigenschaft, die minimal suffiziente Statistiken von identisch verteilten Zufallsvariablen haben. Die $\mathbf{U}_i$ sind zwar nicht immer identisch verteilt, da selbst wenn (3.10) gilt, die Parameter $\varepsilon_1, \ldots, \varepsilon_n$ verschieden sein können. Es wird aber angenommen, daß die Statistiken T_i unabhängig von den $\varepsilon_1, \ldots, \varepsilon_n$ sind. Sind in diesem Fall alle ε_j gleich: $\varepsilon_1 = \ldots = \varepsilon_n$, so sind wegen der strengen Monotonie von f auch die Zufallsvariablen U_{ij} identisch verteilt. Damit folgt aus der Annahme der Minimalsuffizienz der Statistiken T_i für θ_i bei gleichen ε_j, daß T_i symmetrisch ist. Da T_i aber unabhängig von $\varepsilon_1, \ldots, \varepsilon_n$ ist, gilt dies für alle Werte von $\varepsilon_1, \ldots, \varepsilon_n$, so daß (3.13) allgemein gelten muß.

Von Andersen (1973a) wird gezeigt, daß die Annahmen (3.4) und (3.10) bis (3.13) hinreichend sind dafür, daß sich die Wahrscheinlichkeit $P(U_{ij} = u_{ij})$ in der Form (3.6) dekomponieren läßt. Damit sind diese Annahmen hinreichend dafür, daß ein psychologischer Test durch das Rasch-Modell beschrieben werden kann.

THEOREM 3.1. *(Andersen, 1973a). Seien* $\mathbf{U}_i$ *mit* $i = 1, \ldots, m$ *n-dimensionale Zufallsvariable mit dichotomen Komponenten, die lokal stochastisch unabhängig sind, so daß für festes* i

$$P(\mathbf{U}_i = \mathbf{u}_i) = \prod_{j}^{n} P(U_{ij} = u_{ij}).$$

Die Wahrscheinlichkeitsfunktionen $P(U_{ij} = u_{ij})$ *lassen sich in der Form*

$$P(U_{ij} = u_{ij}) = \frac{exp[u_{ij}(\theta_i - \varepsilon_j)]}{1 + exp(\theta_i - \varepsilon_j)}$$

mit skalaren Parametern θ_i, $i = 1, \ldots, m$ *und* ε_j, $j = 1, \ldots, n$ *darstellen, falls folgende Bedingungen gelten:*

1. Es gibt skalare Parameter $\theta_1, \ldots, \theta_m$ und $\varepsilon_1, \ldots, \varepsilon_n$ und eine in den Argumenten θ_i und ε_j für festes u_{ij} streng monotone Funktion f, so daß

$$P(U_{ij} = u_{ij}) = f(u_{ij} \,|\, \theta_i, \varepsilon_j).$$

2. Für jedes θ_i, $i = 1, \ldots, m$ gibt es eine minimal suffiziente Statistik $T_i = t(\mathbf{U}_i)$, die nicht von $\varepsilon_1, \ldots, \varepsilon_n$ abhängt.

Beweis (nach Andersen, 1973a). Der Beweis dieses Theorems stützt sich im wesentlichen auf die Faktorisierungsbedingung (3.12) für suffiziente Statistiken und auf die Symmetrie (3.13) minimal suffizienter Statistiken. Wir nehmen an, daß für alle $i = 1, \ldots, m$ und alle $j = 1, \ldots, n$ die skalaren Parameter θ_i und ε_j existieren, so daß

$$P(U_{ij} = u_{ij}) = f(u_{ij} \,|\, \theta_i, \varepsilon_j),$$

wobei f eine Funktion ist, die für festes u_{ij} in den Argumenten θ_i und ε_j streng monoton ist (vgl. (3.43)). Ferner sei $T_i = t(\mathbf{U}_i)$ eine minimal suffiziente Statistik für θ_i, die nicht von $\varepsilon_1, \ldots, \varepsilon_n$ abhängt. Mit $E = (\varepsilon_1, \ldots, \varepsilon_n)$ gilt dann wegen der lokalen stochastischen Unabhängigkeit

$$\begin{aligned} P(U_i = \mathbf{u}_i) &= f(\mathbf{u}_i \,|\, \theta_i, E) \\ &= \prod_{j=1}^{n} f(u_{ij} \,|\, \theta_i, \varepsilon_j). \end{aligned} \tag{3.14}$$

Wegen der Suffizienz der Statistik T_i für θ_i gibt es aufgrund des Faktorisierungstheorems Funktionen f_1 und f_2, so daß

$$f(\mathbf{u}_i \,|\, \theta_i, E) = f_1(T_i \,|\, \theta_i, E)\, f_2(\mathbf{u}_i \,|\, E).$$

Wir betrachten nun die Wahrscheinlichkeitsfunktionen zweier Personen mit den Parametern θ_i und θ_0 für einen bestimmten Datenvektor $\mathbf{u}_i$. Es gilt dann für den Quotienten dieser beiden Funktionen:

$$\begin{aligned} \frac{f(\mathbf{u}_i \,|\, \theta_i, E)}{f(\mathbf{u}_i \,|\, \theta_0, E)} &= \frac{f_1(T_i \,|\, \theta_i, E)\, f_2(\mathbf{u}_i \,|\, E)}{f_1(T_i \,|\, \theta_0, E)\, f_2(\mathbf{u}_i \,|\, E)} \\ &= \frac{f_1(T_i \,|\, \theta_i, E)}{f_1(T_i \,|\, \theta_0, E)}. \end{aligned} \tag{3.15}$$

Damit hängt dieser Quotient nicht mehr direkt vom Datenvektor $\mathbf{u}_i$ ab, sondern hängt von den Daten nur indirekt über die Statistik T_i ab. Der Wert des Quotienten ist damit für alle Datenvektoren, die den gleichen Wert T_i ergeben, konstant. Wegen der Minimalsuffizienz von T_i für θ_i und der daraus folgenden Symmetrie von T_i in allen Argumenten ist aber T_i für alle Permutationen von $\mathbf{u}_i$ konstant (vgl. (3.13)). Damit erhalten wir aus Gleichung (3.15)

$$\frac{f(\mathbf{u}_i \,|\, \theta_i, E)}{f(\mathbf{u}_i \,|\, \theta_0, E)} = \frac{f[\pi(\mathbf{u}_i) \,|\, \theta_i, E]}{f[\pi(\mathbf{u}_i) \,|\, \theta_0, E]}$$

und wegen Gleichung (3.14)

$$\prod_{j=1}^{n} \frac{f(u_{ij} \,|\, \theta_i, \varepsilon_j)}{f(u_{ij} \,|\, \theta_0, \varepsilon_j)} = \prod_{j=1}^{n} \frac{f(\pi(u_{ij}) \,|\, \theta_i, \varepsilon_j)}{f(\pi(u_{ij}) \,|\, \theta_0, \varepsilon_j)}, \tag{3.16}$$

wobei $\pi(\mathbf{u}_i)$ eine beliebige Permutation von $\mathbf{u}_i$ ist.

Da diese Beziehung auch für alle Teilmengen von $\varepsilon_1 \ldots, \varepsilon_n$ gelten muß, betrachten wir sie für ε_1 und jeweils ein ε_j mit $j = 2, 3, \ldots, n$. Mit $\pi(u_{i1}) = u_{ij}$ und $\pi(u_{ij}) = u_{i1}$ erhalten wir dann

$$\frac{f(u_{i1} \,|\, \theta_i, \varepsilon_1) f(u_{ij} \,|\, \theta_i, \varepsilon_j)}{f(u_{i1} \,|\, \theta_0, \varepsilon_1) f(u_{ij} \,|\, \theta_0, \varepsilon_j)} = \frac{f(u_{ij} \,|\, \theta_i, \varepsilon_1) f(u_{i1} \,|\, \theta_i, \varepsilon_j)}{f(u_{ij} \,|\, \theta_0, \varepsilon_1) f(u_{i1} \,|\, \theta_0, \varepsilon_j)} \tag{3.17}$$

und hieraus

$$\frac{f(u_{ij} \,|\, \theta_i, \varepsilon_j)}{f(u_{i1} \,|\, \theta_i, \varepsilon_j)} = \frac{f(u_{ij} \,|\, \theta_i, \varepsilon_1) f(u_{ij} \,|\, \theta_0, \varepsilon_j) f(u_{i1} \,|\, \theta_0, \varepsilon_1)}{f(u_{i1} \,|\, \theta_i, \varepsilon_1) f(u_{i1} \,|\, \theta_0, \varepsilon_j) f(u_{ij} \,|\, \theta_0, \varepsilon_1)}. \tag{3.18}$$

Wir definieren nun

$$\xi_i := \frac{f(1 \mid \theta_i, \varepsilon_1)}{f(0 \mid \theta_i, \varepsilon_1)}$$

$$\sigma_j := \frac{f(1 \mid \theta_0, \varepsilon_j)\, f(0 \mid \theta_0, \varepsilon_1)}{f(0 \mid \theta_0, \varepsilon_j)\, f(1 \mid \theta_0, \varepsilon_1)},$$

so daß gilt

$$\xi_i \sigma_j = \frac{f(1 \mid \theta_i, \varepsilon_j)}{f(0 \mid \theta_i, \varepsilon_j)}.$$

Da aber $f(0 \mid \theta_i, \varepsilon_j) = 1 - f(1 \mid \theta_i, \varepsilon_j)$, ergibt sich

$$f(1 \mid \theta_i, \varepsilon_j) = \frac{\xi_i \sigma_j}{1 + \xi_i \sigma_j} \tag{3.19}$$

und

$$f(0 \mid \theta_i, \varepsilon_j) = \frac{1}{1 + \xi_i \sigma_j}.$$

Die Definition von ξ_i und σ_j hängt nicht davon ab, welche speziellen Elemente θ_0 und ε_1 ausgewählt wurden. Dies kann man daran erkennen, daß Gleichung (3.18) ebenso wie Gleichung (3.15) als Konsequenzen von (3.16) für alle beliebigen θ_i, θ_0, ε_j und ε_1 gelten muß. Es läßt sich auch leicht nachprüfen, daß mit der Definition von ξ_i und σ_j die kritische Bedingung (3.16) erfüllt wird.

Gleichung (3.19) muß für alle i und alle j gelten. Wir können daher $\varepsilon_1 = 0$ setzen und $\varepsilon_2, \ldots, \varepsilon_n$ durch $\varepsilon_j = -\ln \sigma_j$ für $j = 2, \ldots, n$ und für alle $i = 1, \ldots, n$ $\theta_i = \ln \xi_i$ definieren. Damit hat $f(u_{ij} \mid \theta_i, \varepsilon_j)$ die Form

$$f(u_{ij} \mid \theta_i, \varepsilon_j) = \frac{\exp[u_{ij}(\theta_i - \varepsilon_j)]}{1 + \exp(\theta_i - \varepsilon_j)}.$$

□

3.6 Parameterschätzung

Durch die Existenz minimal suffizienter Statistiken für beide Parameter ist es im Rasch-Modell möglich, einen Parameter unabhängig von den anderen Parameterwerten zu schätzen. Dieses Verfahren beruht auf der Maximum-Likelihood-Methode, man geht dabei jedoch nicht von der Likelihoodfunktion (3.7) aus, sondern betrachtet die Wahrscheinlichkeitsfunktion der Daten unter der Bedingung, daß die suffizienten Statistiken für eine der beiden Parameterklassen bestimmte Ausprägungen angenommen haben. Diese Methode wird *bedingte Maximum-Likelihood-Methode* genannt, da sie von einer bedingten Likelihoodfunktion ausgeht.

3.6.1 Bedingte Maximum-Likelihood-Schätzung für das Rasch-Modell

Für die Likelihoodfunktion des Datenvektors einer Person gilt im Rasch-Modell nach Gleichung (3.7)

$$f(\mathbf{u}_i \mid \theta_i, E) = \frac{\exp(r_i \theta_i)\, \exp\Big(-\sum_{j=1}^{n} u_{ij} \varepsilon_j\Big)}{\prod_{j=1}^{n} [1 + \exp(\theta_i - \varepsilon_j)]}. \tag{3.20}$$

In Abschnitt 3.5 wurde gezeigt, daß die Summen r_i suffiziente Statistiken für die θ_i und die Summen s_j suffiziente Statistiken für die Parameter ε_j sind. Die Methode der bedingten Maximum-Likelihood-Schätzung der Parameter ε_j besteht nun darin, die bedingte Wahrscheinlichkeitsfunktion der Zufallsgrößen U_{ij} zu betrachten, gegeben, daß für die suffizienten Statistiken r_i für die Parameter θ_i bestimmte Werte beobachtet wurden. Gesucht ist

also die bedingte Likelihoodfunktion $h(\mathbf{u}_i \,|\, r_i, \theta_i, E)$. Es wird sich zeigen, daß in dieser Likelihoodfunktion die Parameter θ_i nicht mehr vorkommen. Aus diesem Grund können beim Rasch-Modell die Aufgabenparameter geschätzt werden, ohne daß die Personenparameter bekannt sein müssen bzw. ebenfalls geschätzt werden müssen. Dies ist neben besonderen statistischen Eigenschaften ein wesentlicher Vorteil der bedingten Maximum-Likelihood-Schätzung. Bei der sonst üblichen unbedingten Maximum-Likelihood-Methode wird Gleichung (3.20) direkt als zu maximierende Likelihoodfunktion benutzt, wobei jedoch gleichzeitig Aufgaben- und Personenparameter geschätzt werden müssen. Die Möglichkeit, eine bedingte Maximum-Likelihood-Methode zu verwenden, ist beim Rasch-Modell in der Existenz suffizienter Statistiken für die Modellparameter begründet. Aus statistischer Sicht ist diese Eigenschaft der bedeutendste Vorzug, den das Rasch-Modell gegenüber anderen Testmodellen hat.

Sei nun $g(r_i \,|\, \theta_i, E)$ die Wahrscheinlichkeitsfunktion, die die Wahrscheinlichkeit dafür angibt, daß die Randsumme den Wert r_i annimmt. Aus der Definition der bedingten Wahrscheinlichkeit folgt dann, daß

$$f(\mathbf{u}_i \,|\, \theta_i, E) = g(r_i \,|\, \theta_i, E)\, h(\mathbf{u}_i \,|\, r_i, \theta_i, E)$$

gelten muß, denn das Auftreten eines bestimmten Datenvektors $\mathbf{u}_i$ bestimmt die Randsumme r_i vollständig. Für die gesuchte Funktion h gilt demnach

$$h(\mathbf{u}_i \,|\, r_i, \theta_i, E) = \frac{f(\mathbf{u}_i \,|\, \theta_i, E)}{g(r_i \,|\, \theta_i, E)}. \tag{3.21}$$

Die Funktion f ist aus Gleichung (3.20) bekannt, wir müssen nun die Funktion g bestimmen. g ist die Wahrscheinlichkeit dafür, daß bei gegebenen Aufgabenparametern und gegebenem Personenparameter eine bestimmte Randsumme auftritt. Diese Wahrscheinlichkeit kann grundsätzlich einfach dadurch erhalten werden, daß alle Funktionswerte von Gleichung (3.20) aufsummiert werden, bei denen ein Datenvektor $\mathbf{u}_i$ beobachtet wird, dessen Zeilensumme gleich r_i ist:

$$\begin{aligned}
g(r_i \,|\, \Theta, E) &= \sum_{\sum_{j=1}^{n} u_{ij} = r_i} f(\mathbf{u}_i \,|\, \theta_i, E) \\
&= \sum_{\sum_{j=1}^{n} u_{ij} = r_i} \prod_{j=1}^{n} \frac{\exp[u_{ij}(\theta_i - \varepsilon_j)]}{1 + \exp(\theta_i - \varepsilon_j)} \\
&= \sum_{\sum_{j=1}^{n} u_{ij} = r_i} \frac{\exp(r_i\theta_i)\, \exp\Big(-\sum_{j=1}^{n} u_{ij}\varepsilon_j\Big)}{\prod_{j=1}^{n}[1 + \exp(\theta_i - \varepsilon_j)]} \\
&= \frac{\exp(r_i\theta_i)}{\prod_{j=1}^{n}[1 + \exp(\theta_i - \varepsilon_j)]} \sum_{\sum_{j=1}^{n} u_{ij} = r_i} \exp\Big(-\sum_{j=1}^{n} u_{ij}\varepsilon_j\Big).
\end{aligned}$$

Für die Summe im letzten Teil des obigen Ausdrucks benutzt man üblicherweise die Bezeichnung $\gamma(r_i, E)$ mit der Definition

$$\gamma(r_i, E) = \sum_{\sum_{j=1}^{n} u_{ij} = r_i} \exp\Big(-\sum_{j=1}^{n} u_{ij}\varepsilon_j\Big). \tag{3.22}$$

Die Gleichung $\sum u_{ij} = r_i$ unter dem Summenzeichen soll andeuten, daß die Summation im Ausdruck $\exp(-\sum u_{ij}\varepsilon_j)$ über alle möglichen Antwortvektoren $\mathbf{u}_i$ zu erfolgen hat, deren Zeilensumme gleich r_i ist. Man erhält dann

$$g(r_i \,|\, \Theta, E) = \frac{\exp(r_i\theta_i)\, \gamma(r_i, E)}{\prod_{j=1}^{n}[1 + \exp(\theta_i - \varepsilon_j)]}.$$

Die Funktionen f und g können nun in Gleichung (3.21) zur Berechnung der bedingten Wahrscheinlichkeitsfunktion h eingesetzt werden, und man erhält nach Kürzen:

$$h(\mathbf{u}_i \,|\, r_i, E) = \frac{\exp\Big(-\sum_{j=1}^{n} u_{ij}\varepsilon_j\Big)}{\gamma(r_i, E)} \tag{3.23}$$

Die Personenparameter θ_i kommen im rechten Teil von Gleichung (3.23) nicht mehr vor. Das Argument θ_i von h kann deshalb weggelassen werden. Dies ist Ausdruck der Tatsache, daß die suffizienten Statistiken r_i bereits die gesamte Information über die Parameter θ_i enthalten, die zur Berechnung der Likelihood h notwendig ist.

3.6.2 Schätzung der Aufgabenparameter

Als bedingte Maximum-Likelihood-Schätzfunktionen für ε_j werden diejenigen Werte $\hat{\varepsilon}_j$ bezeichnet, die bei gegebener Datenmatrix $\mathbf{u}$ die bedingte Likelihood (3.23) maximieren. Um diese Schätzfunktionen zu berechnen, muß Gleichung (3.23) für die vollständige Datenmatrix erweitert werden:

$$\begin{aligned} h(\mathbf{u}\,|(r_1,\ldots,r_m), E) &= \frac{\exp\Big(-\sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}\varepsilon_j\Big)}{\prod_{i=1}^{m}\gamma(r_i, E)} \\ &= \frac{\exp\Big(-\sum_{j=1}^{n} s_j\varepsilon_j\Big)}{\prod_{i=1}^{m}\gamma(r_i, E)} \end{aligned} \tag{3.24}$$

Dies wird logarithmiert:

$$\log h(\mathbf{u}\,|(r_1,\ldots,r_m), E) = -\sum_{j=1}^{n} s_j\varepsilon_j - \sum_{i=1}^{m}\log\gamma(r_i, E).$$

Als Ableitung nach ε_j erhält man

$$\frac{\partial \log h}{\partial \varepsilon_j} = -\,s_j - \sum_{i=1}^{m}\frac{\dfrac{\partial\gamma(r_i, E)}{\partial\varepsilon_j}}{\gamma(r_i, E)}.$$

Nach Nullsetzen ergibt sich daraus für die Schätzfunktionen $\hat{\varepsilon}_j$ bzw. $\hat{E}$:

$$-\,s_j = \sum_{i=1}^{m}\frac{\dfrac{\partial\gamma(r_i, \hat{E})}{\partial\varepsilon_j}}{\gamma(r_i, \hat{E})}.$$

Eine genaue Betrachtung der Definition von $\gamma(r, E)$ zeigt, daß die partielle Ableitung von

$$\gamma(r, (\varepsilon_1, \ldots, \varepsilon_j, \ldots, \varepsilon_n))$$

nach ε_j gleich

$$\gamma(r-1, (\varepsilon_1, \ldots, \varepsilon_{j-1}, \varepsilon_{j+1}, \ldots, \varepsilon_n))$$

ist. Schreibt man hierfür $\gamma(r-1, E^{(j)})$, dann erhalten wir also

$$-\;s_j = \sum_{i=1}^{m}\frac{\gamma(r_i - 1, \hat{E}^{(j)})}{\gamma(r_i, \hat{E})} \tag{3.25}$$

für alle $j = 1, \ldots, n$. Diese Gleichung kann nicht explizit nach $\hat{\varepsilon}_j$ aufgelöst werden, sie muß daher mit Hilfe iterativer Verfahren gelöst werden. Eine

brauchbare Methode zur Berechnung der Funktionen γ und ihrer Ableitungen ist bei Fischer (1974) zu finden. Dort ist auch ein Computerprogramm zur Lösung der Gleichungen (3.25) abgedruckt.

Die Schätzfunktionen $\hat{\varepsilon}_j$ sind approximativ normalverteilt. Der Erwartungswert von $\hat{\varepsilon}_j$ ist ε_j, und die Varianz-Kovarianz-Matrix ist die Inverse der *Informationsmatrix* I. Die Zellen I_{jk} dieser Informationsmatrix sind

$$I_{jk} = \mathscr{E}\left(\frac{\partial \ln h}{\partial \varepsilon_j}\frac{\partial \ln h}{\partial \varepsilon_k}\right),$$

wobei h für Gleichung (3.24) steht. Eine ausführliche Ableitung der Werte I_{jk} ist bei Fischer (1974, S. 235 ff) zu finden. Die Informationsmatrix eines Tests ist von besonderem Interesse, weil sie es erlaubt, Konfidenzintervalle für die Schätzfunktionen $\hat{\varepsilon}_j$ und $\hat{\theta}_i$ anzugeben.

3.6.3 Schätzung der Personenparameter

Für die Schätzung der Personenparameter kann nun angenommen werden, daß die Aufgabenparameter bereits bekannt sind. Als Likelihoodfunktion für den Antwortvektor $\mathbf{u}_i$ einer Person ergibt sich

$$\begin{aligned} f(\mathbf{u}_i \,|\, \theta_i, E) &= \prod_{j=1}^{n} \frac{\exp[u_{ij}(\theta_i - \varepsilon_j)]}{1 + \exp(\theta_i - \varepsilon_j)} \\ &= \exp(r_i\theta_i)\, \frac{\exp\Big(-\sum_{j=1}^{n} u_{ij}\varepsilon_j\Big)}{\prod_{j=1}^{n}[1 + \exp(\theta_i - \varepsilon_j)]}. \end{aligned}$$

Nach Logarithmieren erhält man hieraus

$$\log f = r_i\theta_i - \sum_{j=1}^{n} u_{ij}\varepsilon_j - \sum_{j=1}^{n} \log(1 + \exp(\theta_i - \varepsilon_j)).$$

Dies wird nach θ_i differenziert

$$\frac{\partial \log f}{\partial \theta_i} = r_i - \sum_{j=1}^{n} \frac{\exp(\theta_i - \varepsilon_j)}{1 + \exp(\theta_i - \varepsilon_j)},$$

so daß sich nach Nullsetzen für jedes θ_i die Schätzgleichung

$$r_i = \sum_{j=1}^{n} \frac{\exp(\hat{\theta}_i - \varepsilon_j)}{1 + \hat{\theta}_i\varepsilon_j}$$

ergibt. Da die r_i und die ε_j bekannt sind, können damit durch eine iterative Methode die Personenparameter bestimmt werden.

3.6.4 Die Maximum-Likelihood-Methode

Die bedingte Maximum-Likelihood-Methode ist zwar eine mathematisch elegante Lösung des Schätzproblems, wird aber in der Praxis nicht angewandt (Baker, 1992). Zwei Gründe sind dafür ausschlaggebend: Sie ist nur beim Rasch-Modell, nicht bei anderen logistischen Testmodellen anwendbar, und der damit verbundene Rechenaufwand ist zu groß. Insbesondere erfordert eine hinreichend genaue Berechnung der Funktionen (3.22) einen zu großen rechentechnischen Aufwand.

Eine Alternative zur bedingten Maximum-Likelihood-Methode ist die unbedingte Maximum-Likelihood-Methode. Bei ihr wird einfach die Likelihood der Datenmatrix für das entsprechende Modell ((3.7) für das Rasch- oder (3.9) für das Birnbaum-Modell) maximiert. Allerdings ist die unbedingte Maximum-Likelihood-Methode in der üblichen Form nicht ohne weiteres anwendbar, da im vorliegenden Fall nicht sicher ist, daß die damit

berechneten Schätzwerte konsistent sind. Der Grund dafür ist, daß mit jeder Vergrößerung der Stichprobe um eine Person auch die Anzahl der zu schätzenden Parameter zunimmt. Die bedingte Maximum-Likelihood-Methode für das Rasch-Modell löst dieses Problem durch die suffizienten Statistiken für die Personenparameter, wodurch dann in der zu maximierenden Likelihoodfunktion keine Personenparameter mehr vorkommen. Da für alle anderen logistischen Testmodelle keine suffizienten Statistiken für die Personenparameter vorliegen, die ohne Kenntnis der Aufgabenparameter berechnet werden können, ist die bedingte Maximum-Likelihood-Methode nur beim Rasch-Modell anwendbar. Mit der unbedingten Maximum-Likelihood-Methode ist das Schätzproblem durch das Einbeziehen der Verteilung der Personenparameter in der Population zu lösen. Dies geschieht in der von Bock und Lieberman (1970) vorgeschlagenen und und Bock und Aitkin (1981) weiterentwickelten *marginalen Maximum-Likelihood-Methode*.

Die marginale Maximum-Likelihood-Methode geht davon aus, daß die Personenparameter eine bestimmte Verteilungsdichte $g(\theta \mid \eta)$ in der Population haben, wobei η für die Parameter der Dichtefunktion der Personenparameter in der Population steht. Damit wird es möglich, über die Verteilung der Personenparameter zu integrieren und die Aufgabenparameter in der Randverteilung zu schätzen, in der die einzelnen Personenparameter nicht mehr enthalten sind. Zur Anwendung kommt dabei die Formel von Bayes (1.1) für Verteilungsdichten:

$$P(\theta_i \mid U_i, \eta, \Lambda) = \frac{P(U_i \mid \theta_i, \Lambda)\, g(\theta \mid \eta)}{\int P(U_i \mid \theta_i, \Lambda)\, g(\theta \mid \eta)\, d\theta}$$

Für $P(U_i \mid \theta_i, \Lambda)\, g(\theta \mid \eta)$ wird die Likelihoodfunktion des Datenvektors der Person i entsprechend dem Modell eingesetzt. Die Form der Dichtefunktion $g(\theta \mid \eta)$ ist zwar unbekannt, es genügt aber, sie innerhalb des Verfahrens approximativ festzulegen. Eine detailierte Beschreibung des Schätzverfahrens findet man bei Baker (1992). Wie bei der bedingten Maximum-Likelihood-Methode ergeben sich bei der marginalen Maximum-Likelihood-Methode im ersten Schritt nur die Aufgabenparameter. Die Personenparameter können dann mit Hilfe der oben beschriebenen unbedingten Maximum-Likelihood-Methode bestimmt werden[1].

3.6.5 Die statistische Information einer Testaufgabe

Zur Berechnung eines Konfidenzintervalls für die Parameter θ_i kann die oben bereits kurz erwähnte, statistische Information I benutzt werden. Die Information, die eine Aufgabe j für die Schätzfunktion θ_i liefert, ist

$$I_j(\theta_i) = \mathscr{E}\left[\left(\frac{\partial \ln f(u_{ij})}{\partial \theta_i}\right)^2\right],$$

wobei hier zur Vereinfachung kurz $f(u_{ij})$ für die Likelihoodfunktion des Datums u_{ij} geschrieben wurde, ohne genauer anzugeben, von welchen Parametern diese abhängt. Dies ist sinnvoll, da die Überlegungen dieses Abschnitts sowohl für das Rasch- als auch für das Birnbaum-Modell gelten. In den folgenden Ausdrücken schreiben wir f' für die erste Ableitung von f. Alle Ableitungen sind nach θ_i. Wir erhalten

$$I_j(\theta_i) \quad = \quad \mathscr{E}\left[\left(\frac{f'(u_{ij})}{f(u_{ij})}\right)^2\right].$$

Wir bilden nun den Erwartungswert. Die Zufallsvariable u_{ij} kann nur die Werte 0 und 1 annehmen:

$$I_j(\theta_i) \quad = \quad \frac{f'^2(1)}{f^2(1)} f(1) + \frac{f'^2(0)}{f^2(0)} f(0)$$

[1] Ein modernes Computerprogramm zur Parameterschätzung für das Rasch-Modell wird vom Institut für die Pädagogik der Naturwissenschaften (Kiel) vertrieben (von Davier, 1994). Die Modelle, die mit diesem Programm untersucht werden können, werden von Rost & von Davier (1995) und von Davier & Rost (1995) beschrieben.

$$= \frac{f(0)f'^2(1) + f(1)f'^2(0)}{f(0)f(1)}.$$

Da $f(0) = 1 - f(1)$ ist $f'(0) = -f'(1)$, und wir erhalten

$$\begin{aligned} I_j(\theta_i) &= \frac{f(0)f'^2(1) + f(1)[-f'(1)]^2}{f(0)\,f(1)} \\ &= \frac{f'^2(1)[f(0) + f(1)]}{f(0)\,f(1)} \\ &= \frac{f'^2(1)}{f(0)\,f(1)}. \end{aligned}$$

Für das Rasch-Modell ergibt sich damit

$$I_j(\theta_i) = \frac{\exp(\theta_j - \varepsilon_j)}{[1 + \exp(\theta_i - \varepsilon_j)]^2}, \tag{3.26}$$

denn in diesem Fall ist

$$\begin{aligned} f'(1\,|\,\theta_i, \varepsilon_j) &= \frac{\partial}{\partial\theta_i} \frac{\exp(\theta_i - \varepsilon_j)}{1 + \exp(\theta_i - \varepsilon_j)} \\ &= \frac{[1 + \exp(\theta_i - \varepsilon_j)]\exp(\theta_i - \varepsilon_j) - [\exp(\theta_i - \varepsilon_j)]^2}{[1 + \exp(\theta_i - \varepsilon_j)]^2} \\ &= \frac{\exp(\theta_i - \varepsilon_j)}{[1 + \exp(\theta_i - \varepsilon_j)]^2}. \end{aligned}$$

Abbildung 3.3 zeigt den Verlauf der Informationsfunktion einer Aufgabe zusammen mit ihrer Aufgabenkennlinie $f(1\,|\,\theta_i, \varepsilon_j)$.

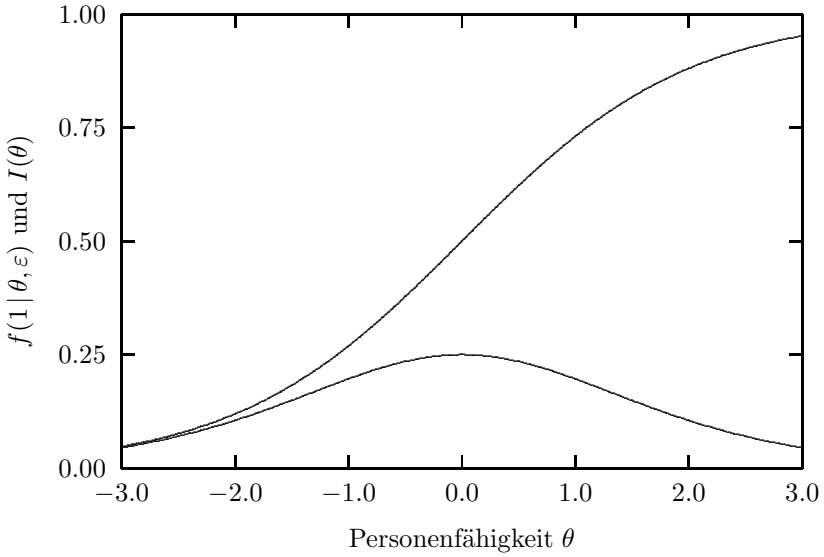

Abb. 3.3. Kennlinie einer Aufgabe mit dem Parameter $\varepsilon = 0.0$ und Informationsfunktion der Testaufgabe nach Gleichung (3.6). Die Informationsfunktion ist dort maximal, wo die Kennlinie die größte Steigung hat. Dieser Punkt ist genau dort, wo $\theta = \varepsilon$.

Für das Birnbaum-Modell erhält man den Informationsbeitrag

$$I_j(\theta_i) = \frac{\alpha_j^2 \exp[\alpha_j(\theta_i - \varepsilon_j)]}{[1 + \exp(\alpha_j(\theta_i - \varepsilon_j))]^2}. \tag{3.27}$$

Der Parameter α wird deshalb *Trennschärfe* genannt. Je größer sein Wert ist, desto mehr Information liefert die Aufgabe in dem Leistungsbereich, in dem $\theta \sim \varepsilon$. Aufgaben mit hohen α-Werten sind also gut geeignet, um Personen mit $\theta < \varepsilon$ von solchen mit $\theta > \varepsilon$ zu trennen. Abbildung 3.4 zeigt Informationsfunktionen für Aufgaben aus einem Birnbaum-Test.

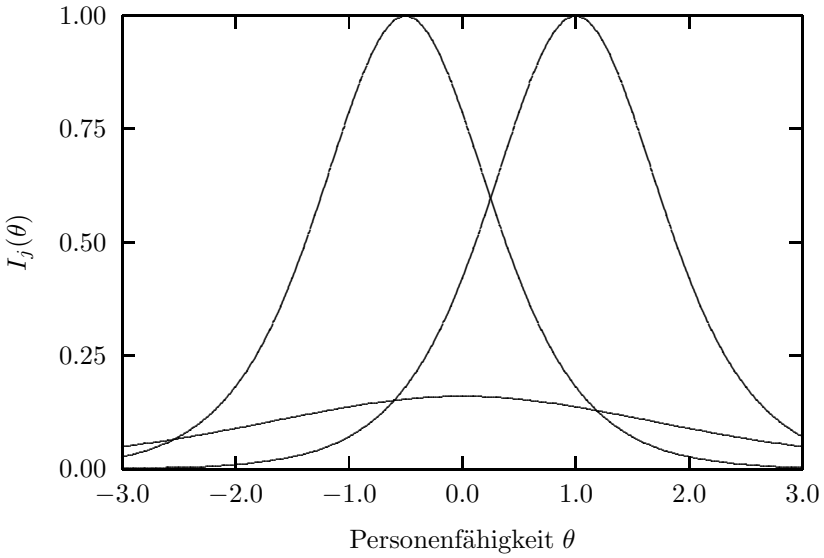

Abb. 3.4. Informationsfunktionen für die 3 Aufgaben, deren Kennlinien in Abbildung 3.2 dargestellt sind.

Es läßt sich zeigen, daß das Informationsmaß additiv ist. Der Informationsbeitrag mehrerer Testaufgaben $j = 1, \ldots, n$ für die Schätzung eines Parameters θ_i ist gleich der Summe der Informationsbeiträge jeder einzelnen Aufgabe:

$$I(\theta_i) = \sum_{j=1}^{n} I_j(\theta_i).$$

Eine genauere Betrachtung der Gleichungen (3.26) und (3.27) zeigt, daß in beiden Fällen das Maximum der Informationsfunktion dort ist, wo $\theta_i = \varepsilon_j$. Eine Testaufgabe liefert also genau für den Personenparameter θ_i am meisten Information, der den gleichen Wert hat wie der Schwierigkeitsgrad der Aufgabe.

3.6.6 Konfidenzintervalle für die Personenparameter

Wie bereits oben erwähnt, sind Maximum-Likelihood-Schätzungen asymptotisch normalverteilt, wobei die Fehlervarianz asymptotisch gleich dem Kehrwert der Informationsfunktion ist. Wir können deshalb für die Parameter θ_i mit Hilfe der Informationsfunktion ein Konfidenzintervall zum Niveau α angeben:

$$\hat{\theta}_i - \frac{z_{1-\alpha/2}}{I(\hat{\theta}_i)^{1/2}} \;\leq\; \theta_i \;\leq\; \hat{\theta}_i + \frac{z_{1-\alpha/2}}{I(\hat{\theta}_i)^{1/2}}.$$

In dieser Formel ist $z_{1-\alpha/2}$ der Abweichungswert der Standardnormalverteilung für den $P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$ gilt. Die Formel zeigt, daß die Breite des Konfidenzintervalls proportional zum Schätzfehler bzw. umgekehrt proportional der Wurzel aus der Information ist[2].

3.6.7 Adaptives Schätzen der Personenparameter

Im Gegensatz zur klassischen Testtheorie ist es bei den logistischen Testmodellen nicht notwendig, daß für die Schätzung des Personenparameters alle Testaufgaben vorgelegt werden. Mit der in Abschnitt 3.6.3 beschriebenen Maximum-Likelihood-Methode zur Schätzung der Personenparameter für den Fall, daß die Itemparameter schon bekannt sind, kann der Personenparameter auch bei unvollständigen Daten geschätzt werden. Es wird in diesem Fall eben einfach die Likelihood des vorliegenden Datenvektors der Person maximiert. Damit ergibt sich die Möglichkeit einer adaptiven

[2]Bei Fischer (1974, S. 297, Formeln 15.3.9 bis 15.3.11) ist an dieser Stelle ein Druckfehler.

Schätzmethode. Die Überlegungen in Abschnitt 3.6.5 zeigen, daß nicht alle Testaufgaben gleich viel zur Schätzung des Parameters θ einer bestimmten Person beitragen. Man könnte daher die Testvorgabe erheblich abkürzen, wenn nur die Aufgaben vorgelegt würden, die einen nennenswerten Informationsbeitrag zur Schätzung des Personenparameters leisten. Am meisten Information liefern Aufgaben, deren Schwierigkeitsparameter ε gleich dem Personenparameter θ sind. Man muß also versuchen, Aufgaben vorzugeben, deren Parameter die Information maximieren. Welche das sind, kann man aber erst feststellen, wenn der Personenparameter bekannt ist.

Eine Lösung für dieses Dilemma ist die adaptive Schätzung des Personenparameters. Man versucht *während* des Testvorgangs anhand der Ergebnisse bei den bisher vorgelegten Aufgaben eine Schätzung des Personenparameters zu berechnen und gibt danach die optimale Aufgabe vor. Diesen Vorgang wiederholt man nach jeder Antwort der Testperson. Der Test kann abgebrochen werden, wenn genügend Information gesammelt ist, also etwa dann, wenn das Konfidenzintervall für den Personenparameter hinreichend klein ist.

Beim adaptiven Testen entstehen einige spezielle Probleme, die in dieser Form bei einer vollständigen Testvorgabe nicht auftreten:

1. Am Anfang des Verfahrens kann keine Schätzung durchgeführt werden, man benötigt daher für den Anfang ein anderes Auswahlverfahren für die Testaufgaben. Beispielsweise beginnt man mit einer Aufgabe mittlerer Schwierigkeit und wählt dann, je nachdem, ob die Versuchsperson eine richtige oder falsche Antwort gibt, eine schwierigere oder eine leichtere Aufgabe. Man geht so lange auf diese Weise vor, bis eine Maximum Likelihood-Schätzung des Personenparameters möglich ist.

2. Die Abstimmung der Aufgaben auf die Personenfähigkeit führt nur dann zu einer höheren Effizienz des Testvorgangs, wenn für jede Personenfähigkeit geeignete Aufgaben vorhanden sind. Man benötigt daher für ein adaptives Verfahren insgesamt erheblich mehr Testaufgaben als für einen Test fester Länge.

3. Das adaptive Testverfahren setzt die Kenntnis der Aufgabenparameter voraus. Die Bestimmung der Aufgabenparameter kann selbst aber nicht ohne weiteres adaptiv erfolgen, da die Maximum-Likelihood-Methode bei zwei Aufgaben nur dann Schätzwerte für die Parameter liefern kann, wenn hinreichend viele Personen beide Aufgaben bearbeitet und nicht alle Personen bei einer der beiden Aufgaben das gleiche Ergebnis erzeugt haben. Die Aufgabenparameter müssen daher bereits vor der adaptiven Testvorgabe durch eine volle Testvorgabe bestimmt werden. Dafür genügt natürlich eine Stichprobe der Population, in der auch die adaptive Vorgabe durchgeführt werden soll. Häufig werden die Aufgabenparameter neuer Items auch dadurch bestimmt, daß die neuen Items während einer adaptiven Vorgabe bekannter Items eingefügt werden.

4. Aus der Gültigkeit eines Testmodells für eine Sammlung von Testaufgaben bei normaler Vorgabe kann nicht auf die Modellgültigkeit bei adaptiver Vorgabe geschlossen werden. Auch die Aufgabenparameter, die bei einer festen Vorgabe bestimmt werden, müssen nicht mit denen identisch sein, die bei einer adaptiven Vorgabe wirksam sind. Ein Grund dafür ist offensichtlich: Bei einer Vorgabe der Aufgaben mit fester Reihenfolge wirken Lerneffekte als Veränderung der Aufgabenschwierigkeiten, die in die Aufgabenparameter eingeht. Wegen der Konfundierung der Aufgabenreihenfolge und der Bearbeitungsfolge fällt dies weder beim Modelltest als Verletzung der lokalen stochastischen Unabhängigkeit noch bei den Aufgabenparametern auf. Gibt man aber die Aufgaben in einer anderen Reihenfolge vor, sind solche Kontexteffekte als Verletzungen der Modellannahmen wirksam.

Eine ausführliche Diskussion der Probleme adaptiver Testvorgaben findet man bei Wainer (1990).

3.7 Ein Modelltest

Von Andersen (1973b) wurde ein statistischer Test des Rasch-Modells vorgeschlagen, der auf der bedingten Maximum-Likelihood-Schätzung der Aufgabenparameter aufbaut. Der Grundgedanke des Tests besteht darin, mehrere Schätzungen der Aufgabenparameter aus verschiedenen Teilmengen der Personen mit einer Schätzung der Aufgabenparameter zu vergleichen, die auf der gesamten Personenmenge basiert. Man teilt also die Personenmenge A in zwei disjunkte Teilmengen A_1 und A_2 und berechnet die Aufgabenparameter einmal mit den Daten der Personen A_1 und einmal mit den Daten der Personen A_2. Seien E_1 und E_2 die jeweiligen Schätzwerte für die Aufgabenparameter. E seien die Schätzwerte für die Aufgabenparameter, die mit den Daten aller Personen A berechnet werden. $\mathbf{u}_{(1)}$ und $\mathbf{u}_{(2)}$ seien die jeweiligen Antwortmatrizen der Teilgruppen A_1 und A_2. Da die Parameterschätzungen E_1 und E_2 jeweils optimal für die Personengruppen A_1 und A_2 sind, muß für die bedingten Likelihoodfunktionen (3.24) bezogen auf die Teilmengen A_1 und A_2 folgende Beziehung gelten:

$$h(\mathbf{u}\,|(r_1,\ldots,r_m),E) \;\leq\; h(\mathbf{u}_{(1)}\,|(r_1,\ldots,r_m),E_1)h(\mathbf{u}_{(2)}\,|(r_1,\ldots,r_m),E_2) \tag{3.28}$$

Falls nun die Daten das Rasch-Modell erfüllen, dann sind die Parameterschätzungen E, E_1 und E_2 identisch, da in allen Fällen die gleichen Aufgaben beantwortet wurden. Die Ungleichung (3.28) wird dann zur Gleichung. Je stärker in Gleichung (3.28) jedoch der rechte Teil den linken überwiegt, desto stärker unterscheiden sich die Parameterschätzungen in den Teilgruppen von der in der Gesamtmenge, so daß das Rasch-Modell verworfen werden muß.

Von Andersen (1973b) wird dieses Argument statistisch ausgearbeitet. Er betrachtet nicht nur eine Aufteilung in zwei Teilmengen, sondern schlägt vor, die Daten in so viele Gruppen aufzuteilen wie unterschiedliche Werte von r_i beobachtet wurden. Ausgenommen werden die Werte $r = 0$ und $r = n$, so daß sich dann maximal $n-1$ verschiedene Personengruppen ergeben. Für jede dieser Personengruppen kann mit Hilfe von Gleichung (3.24) eine Schätzung der Aufgabenparameter durchgeführt werden. Sei nun k analog der obigen Notation der Index der Personengruppe A_k, mit r_k richtigen Antworten, und m_k sei die Anzahl dieser Personen. Wir betrachten dann die Ungleichung

$$h(\mathbf{u}\,|(r_1,\ldots,r_m),E) \;\leq\; \prod_{k=1}^{n-1} h(\mathbf{u}_{(k)}\,|\,r_k,E_k),$$

wobei auch bei $\mathbf{u}$ die Daten der Personen mit $r_i = 0$ und $r_i = n$ nicht berücksichtigt werden. Hieraus folgt, daß

$$h(\mathbf{u}_{(k)}\,|\,r_k,E_k) = \frac{\exp\Big(-\sum_{j=1}^{n} s_{kj}\varepsilon_{kj}\Big)}{\gamma(r_k,E_k)^{m_k}},$$

wobei ε_{kj} der Parameter der Aufgabe j in der Personengruppe mit r_k richtigen Antworten und s_{kj} die Anzahl von richtigen Lösungen von Aufgabe j in dieser Personengruppe ist. Der Likelihoodquotient

$$\lambda = \frac{h(\mathbf{u}\,|(r_1,\ldots,r_m),E)}{\prod_{k=1}^{n-1} h(\mathbf{u}_{(k)}\,|\,r_k,E_k)} \tag{3.29}$$

hat bei Gültigkeit des Rasch-Modells für die wahren Parameter E und E_k den Wert 1, da dann $E = E_k$ sein muß. Von Andersen (1973b) wird gezeigt, daß bei Maximum-Likelihood-Schätzung von E und E_k die Verteilung des Ausdrucks $Z = -2\log\lambda$ unter der Nullhypothese der Gültigkeit des Rasch-Modells gegen eine χ^2-Verteilung mit $(n-1)(n-2)$ Freiheitsgraden strebt, falls m_k für alle $k = 1,\ldots,n-1$ hinreichend groß wird. Die Anzahl der

Freiheitsgrade bei diesem Test ergibt sich daraus, daß im Nenner von Gleichung (3.29) genau $(n-1)$ Vektoren E_k mit jeweils n Parametern enthalten sind. Von den n Parametern in E_k sind aber wegen des Skalenniveaus der Parameter nur $(n-1)$ unabhängig, so daß sich im Nenner $(n-1)(n-1)$ unabhängige Parameter ergeben. Der Zähler enthält analog dazu $(n-1)$ unabhängige Parameter. Insgesamt hat dann die Prüfgröße Z dieses Tests $(n-1)(n-1)-(n-1)=(n-1)(n-2)$ Freiheitsgrade.

3.8 Meßtheoretische Aspekte logistischer Modelle

Aus meßtheoretischer Sicht wird bei einem Testverfahren eine beobachtbare Größe in zwei nicht direkt beobachtbare Einflußgrößen dekomponiert. Im Fall der linearen Testtheorie ist es die Dekomposition der beobachtbaren Größe „Beobachtungswert“ in die beiden Einflußgrößen „Personeneigenschaft“ und „Testeigenschaft“. Im Fall der logistischen Modelle ist es die „beobachtbare“ Größe „Lösungswahrscheinlichkeit“, die in eine „Personeneigenschaft“ und eine „Aufgabeneigenschaft“ dekomponiert wird. Strukturen dieser Art werden in der Meßtheorie als *verbundene Meßstrukturen* bezeichnet. Ihr formales Merkmal ist, daß sie auf Daten aufbauen, die auf faktorisierbaren Mengen definiert sind.

3.8.1 Tests als verbundene Meßstrukturen

In der Regel ist die Faktorisierung durch die Datenerhebung vorgegeben. So hat man bei den logistischen Modellen als *Objektbereich* das Mengenprodukt $A\times V$ aus allen möglichen Paaren (a,v) von Personen $a\in A$ und Aufgaben $v\in V$. Für die klassischen Testmodelle wird V als Menge von Tests interpretiert. Die Daten bestehen in Aussagen über Objektpaare, also etwa in der Aussage „die Lösungswahrscheinlichkeit von Person a bei der Aufgabe v ist mindestens so groß wie die von Person b bei Aufgabe w“. Die Daten werden in der Meßtheorie als Relationen auf dem Objektbereich formalisiert. Die obige Aussage über den Vergleich des „Objekts“ (a,v) mit dem Objekt (b,w) schreibt man als

$$(a,v)\succsim(b,w).$$

Die Relation $\succsim$ (sprich: „... ist mindestens so ... wie ...“) ist also eine Teilmenge von $(A\times V)\times(A\times V)$.

Die Theorie einer verbundenen Meßstruktur dieser Art gibt an, unter welchen Bedingungen eine Dekomposition der Daten in der Form möglich ist, daß für Funktionen θ auf A und λ auf V die Bedingung

$$(a,v)\succsim(b,w)\quad\text{gdw.}\quad F[\theta(a),\lambda(v)]\geq F[\theta(b),\lambda(w)]$$

für eine noch zu bestimmende Funktion F gilt. Bei der Anwendung meßtheoretischer Methoden auf diese Fragestellung ergeben sich mehrere Probleme:

1. Übersetzt man quantitative Aussagen über Antwortwahrscheinlichkeiten in die qualitative Sprache der Relation $\succsim$, dann geht Information verloren. So wird etwa der Unterschied der Lösungswahrscheinlichkeiten nicht kodiert, kleine und große Unterschiede können zur gleichen Ordnungsaussage führen. Man wertet also nur die Ordnungsinformation in den Wahrscheinlichkeiten aus. Das gleiche gilt für die Anwendung auf Beobachtungswerte. Auch dort wird dann nur die Ordnungsinformation ausgewertet.

2. Die meßtheoretischen Strukturen bieten keine adäquate Möglichkeit, das Problem der statistischen Unsicherheit zu behandeln. Bei den logistischen Modellen wird dieses Problem dadurch „gelöst“, daß man nicht von echten Beobachtungsdaten ausgeht, sondern von Lösungswahrscheinlichkeiten. Da die Wahrscheinlichkeiten keine statistische Unsicherheit mehr enthalten, ist das Problem damit umgangen. Für die Anwendung meßtheoretischer Methoden auf klassische Testmodelle umgeht Steyer (1989) das Problem auf ähnliche Weise: Er formuliert die Theorie auf den bedingten Erwartungen T_X und nicht auf den Beobachtungswerten X. Auch damit bleiben der Meßtheorie Aussagen zum stochastischen Charakter der Daten erspart.

3. In der Meßtheorie sind bisher für qualitative Daten nur sehr einfache Modelle bekannt. Das sind im wesentlichen nur ordinal dekomponierbare und additiv verbundene Strukturen. Gemeinsam ist diesen beiden Modellen, daß sie die Abbildungen θ und λ als eindimensional voraussetzen. Das Modell affin verwandter Personenparameter und das logistische Birnbaum-Modell können damit nicht als fundamentale Meßstrukturen behandelt werden[3].

Eine Folge dieser Probleme, insbesondere des zweiten oben erwähnten Punktes, ist, daß meßtheoretische Methoden bei der praktischen Testkonstruktion keine Rolle spielen. Allerdings sind meßtheoretische Überlegungen von großer Bedeutung für die theoretischen Grundlagen der Modelle und insbesondere für die Frage, welchen theoretischen Status Testskalen besitzen. Dies betrifft die Frage des Skalenniveaus und der Bedeutsamkeit von Aussagen, die mit Hilfe von Testskalen möglich sind. Vor allem im Zusammenhang mit dem Rasch-Modell gab es in der Literatur einige Unklarheiten über das Skalenniveau (Fischer, 1988). Fragen der Bedeutsamkeit von Aussagen über Testskalen hängen eng mit dem Konzept der *spezifischen Objektivität* zusammen, das von Rasch (1960) allerdings ohne Kenntnis der Meßtheorie eingeführt wurde. Von Irtel (1987, 1993, 1995) wird gezeigt, daß sich dieses Konzept als Bedeutsamkeitsproblem einer Meßstruktur behandeln läßt. Wir werden darauf später genauer eingehen.

3.8.2 Das Rasch-Modell als Spezialfall einer additiv verbundenen Struktur

Das wesentliche Merkmal des Rasch-Modells in Bezug auf die oben angedeutete verbundene Meßstruktur ist, daß die Parameter θ und λ eindimensional sind, und daß diese subtraktiv in die Funktion $F[\theta(a), \lambda(v)]$ eingehen. Wie früher benutzen wir für den eindimensionalen Aufgabenparameter des Rasch-Modells die Bezeichnung ε. Dem Rasch-Modell liegt also eine Bedingung der Form

$$(a, v) \succsim (b, w) \quad \text{gdw.} \quad \theta(a) - \varepsilon(v) \geq \theta(b) - \varepsilon(w)$$

zugrunde. Berücksichtigt man nun, daß die qualitative Relation $\succsim$ durch Lösungswahrscheinlichkeiten $P(a, v)$ definiert ist, dann erhält man

$$P(a, v) \geq P(b, w) \quad \text{gdw.} \quad \theta(a) - \varepsilon(v) \geq \theta(b) - \varepsilon(w). \tag{3.30}$$

Dies bedeutet, daß die Funktionen P und $(\theta - \varepsilon)$ isoton sein müssen, so daß eine streng monotone Funktion F existiert, daß

$$P(a, v) = F[\theta(a) - \varepsilon(v)]. \tag{3.31}$$

Eine Meßstruktur dieser Form wurde bereits von Pfanzagl (1971) untersucht. Er konnte zeigen, daß die empirischen Restriktionen eines solchen Modells denen additiv verbundener Meßstrukturen im wesentlichen äquivalent sind. Aus der Theorie der additiv verbundenen Messung ist bekannt, daß die Skalen θ und ε aus Bedingung (3.30) eindeutig bis auf Transformationen der Form $r\theta + s_1$ und $r\varepsilon + s_2$ sind, es sich also um Intervallskalen handelt. Die additiv verbundene Struktur des Rasch-Modells begründet also Intervallskalen der Personen- und Aufgabenparameter. Aus der Modellgleichung

$$P(a, v) = \frac{\exp[\theta(a) - \varepsilon(v)]}{1 + \exp[\theta(a) - \varepsilon(v)]} \tag{3.32}$$

sieht man aber sofort, daß θ und ε nicht nur Intervall- sondern sogar Differenzenskalen sind, denn wenn θ und ε zulässige Parameter sind, können $\alpha\theta$ und $\alpha\varepsilon$ keine zulässigen Parameter mehr sein. Einen Beweis dafür gibt Theorem 3.2.

[3] Bei den Strukturen, die Steyer & Eid (1993) für affin verwandte Messungen und wir in Def. 3.2 auf Seite 68 für das Birnbaum-Modell angeben, handelt es sich um *abgeleitete* Messungen. Diese gehen nicht von qualitativen Daten, sondern von quantitativen Daten aus.

Aus den unterschiedlichen Skaleneigenschaften folgt, daß die Restriktionen des Rasch-Modells an die Daten stärker sein müssen als die der additiv verbundenen Messung. Fischer (1987) zeigt, daß dies auch der Fall ist. Der Unterschied zwischen der Form (3.31) und Gleichung (3.32) ist, daß in (3.31) für die Funktion F jede beliebige, streng monotone Funktion eingesetzt werden kann, während in (3.32) die entsprechende Funktion F fest definiert ist, nämlich die logistische Funktion $F(x) = \exp(x)/(1 + \exp(x))$.

Für „empirische“ Restriktionen bedeutet dies, daß aus (3.31) die Bedingung

$$F^{-1}[P(a,v)] - F^{-1}[P(b,v)] = F^{-1}[P(a,w)] - F^{-1}[P(b,w)] \tag{3.33}$$

folgt und aus (3.32) die Bedingung

$$\ln \frac{P(a,v)}{1-P(a,v)} \ln \frac{1-P(b,v)}{P(b,v)} = \ln \frac{P(a,w)}{1-P(a,w)} \ln \frac{1-P(b,w)}{P(b,w)}, \tag{3.34}$$

denn die Umkehrfunktion der logistischen Funktion ist

$$F^{-1}[P(a,v)] = \ln \frac{P(a,v)}{1-P(a,v)}.$$

Damit ist der Unterschied der empirischen Restriktionen zwischen einem einfachen additiv verbundenem Modell und dem Rasch-Modell deutlich: Bedingung (3.33) ist empirisch nicht vollständig testbar, da die Funktion F nicht bekannt ist. Von ihr ist nur bekannt, daß sie streng monoton ist. Bedingung (3.34) dagegen ist exakt spezifiziert und empirisch prüfbar, vorausgesetzt die Wahrscheinlichkeiten P sind bekannt. Wir werden später in Theorem 3.2 zeigen, daß Bedingung (3.34) nicht nur notwendig, sondern sogar hinreichend für die Definition der Personen- und Aufgabenparameter ist. Zur empirischen Prüfung des Modells (3.31) bleiben nur die empirischen Bedingungen für allgemeine additive verbundene Messungen, die sämtlich schwächer sind als Bedingung (3.34). Wir fassen diese Überlegungen im folgenden Theorem zusammen. Zur Illustration der meßtheoretischen Skalenkonstruktion ist auch der Beweis mit aufgeführt.

Definition 3.1. Ein Paarvergleichssystem $\langle A \times V, P\rangle$, in dem A und V Mengen und P eine Abbildung von $A \times V$ in das offene, reelle Intervall $(0,1)$ ist, heißt genau dann *Rasch-skalierbar*, wenn es reellwertige Funktionen θ und ε gibt, die auf A bzw. V definiert sind, so daß für alle a in A und alle v in V gilt

$$P(a,v) = \frac{\exp[\theta(a) - \varepsilon(v)]}{1 + \exp[\theta(a) - \varepsilon(v)]}. \tag{3.35}$$

Das Paar $\langle \theta, \varepsilon\rangle$ heißt dann *Rasch-Repräsentation für* $\langle A \times V, P\rangle$.

Theorem 3.2. *Ein Paarvergleichssystem $\langle A \times V, P\rangle$ in dem A und V Mengen und P eine Abbildung von $A \times V$ in das offene, reelle Intervall $(0,1)$ ist, ist genau dann Rasch-skalierbar, wenn für alle a, b in A und alle v, w in V gilt:*

$$\frac{P(a,v)}{1-P(a,v)} \frac{1-P(b,v)}{P(b,v)} = \frac{P(a,w)}{1-P(a,w)} \frac{1-P(b,w)}{P(b,w)}. \tag{3.36}$$

Sind ferner $\langle \theta, \varepsilon\rangle$ und $\langle \theta', \varepsilon'\rangle$ zwei Rasch-Repräsentationen für $\langle A \times V, P\rangle$ dann gibt es eine reellwertige Konstante s, so daß $\theta'(a) = \theta(a) + s$ und $\varepsilon'(v) = \varepsilon(v) + s$ für alle a in A und alle v in V.

Beweis. Zur Vereinfachung der Schreibweise kürzen wir mit $L(x)$ die Umkehrfunktion der logistischen Funktion, die sogenannte „Logit-Transformation“ ab:

$$L(x) = \ln \frac{x}{1-x}, \tag{3.37}$$

ab. Die Multiplikationsbedingung (3.36) ist dann äquivalent zur Bedingung

$$L[P(a,v)] - L[P(b,v)] = L[P(a,w)] - L[P(b,w)]. \tag{3.38}$$

Um zu zeigen, daß die Multiplikationsbedingung (3.38) hinreichend ist, definiert man

$$\theta(a) := L[P(a, w_0)] - L[P(b_0, w_0)]$$

für ein festes b_0 in A und ein festes w_0 in V. Ferner wird $\varepsilon(v)$ für alle v in V durch

$$\varepsilon(v) := -L[P(b_0, v)]$$

definiert. Man erhält dann unter Benutzung der Multiplikationsbedingung (3.38) für $\theta(a) - \varepsilon(v)$

$$\begin{aligned} \theta(a) - \varepsilon(v) &= L[P(a, w_0)] - L[P(b_0, w_0)] + L[P(b_0, v)] \\ &= L[P(a, v)] - L[P(b_0, v)] + L[P(b_0, v)] \\ &= L[P(a, v)]. \end{aligned}$$

Setzt man dies in die Modellgleichung (3.35) ein, dann zeigt sich, daß die so definierten Funktionen θ und ε das Modell erfüllen. Damit ist gezeigt, daß Bedingung (3.38) hinreichend ist. Um zu zeigen, daß sie auch notwendig ist, genügt es, die Modellgleichung (3.35) in die linke Seite von (3.38) einzusetzen. Mit der Gleichung

$$L[P(a, v)] = \theta(a) - \varepsilon(v), \tag{3.39}$$

die aus (3.35) folgt, sieht man sofort, daß dann (3.38) gelten muß, sie braucht dazu nur logarithmiert zu werden:

$$\begin{aligned} L[P(a, v)] - L[P(b, v)] &= [\theta(a) - \varepsilon(v)] - [\theta(b) - \varepsilon(v)] \\ &= \theta(a) - \theta(b) \\ &= [\theta(a) - \varepsilon(y)] - [\theta(b) - \varepsilon(y)] \\ &= L[P(a, w)] - L[P(b, w)]. \end{aligned}$$

Um die Eindeutigkeitseigenschaften der Abbildungen θ und ε zu untersuchen, nehmen wir an, daß $\langle \theta, \varepsilon \rangle$ und $\langle \theta', \varepsilon' \rangle$ Rasch-Repräsentationen für $\langle A \times V, P \rangle$ sind. Aus Gleichung (3.39) folgt dann, daß für alle a in A und alle v in V

$$\theta(a) - \varepsilon(v) = \theta'(a) - \varepsilon'(v)$$

gelten muß. Damit ist $\theta'(a) = \theta(a) + s$ mit $s = \varepsilon'(v) - \varepsilon(v)$. Da die obige Beziehung für beliebige v gelten muß, ist s unabhängig von v und somit auch für alle anderen Elemente von A konstant. Für ε' gilt dann $\varepsilon'(v) = \varepsilon(v) + s$. □

Der Eindeutigkeitsteil des Theorems 3.2 begründet, warum die Aussage, daß im Rasch-Modell Personen und Aufgaben auf einer „gemeinsamen Differenzenskala“ gemessen werden, sinnvoll ist. Auch wenn die Funktionen θ und ε unterschiedliche Definitionsbereiche haben, so sind doch ihre Funktionswerte eng miteinander verknüpft. Die zulässigen Transformationen der beiden Funktionen haben nur einen gemeinsamen Parameter. Es sind also Aussagen wie etwa „$\theta(a) - \varepsilon(v) = k$“ invariant gegenüber zulässigen Transformationen und damit empirisch bedeutsam.

Fischer (1987, 1988) gibt als Skalenniveau für die Parameter des Rasch-Modells Intervallskalen an. Der Grund dafür ist, daß er eine von Definition 3.1 abweichende Definition des Modells benutzt. Ein Paarvergleichssystem $\langle A \times V, P \rangle$ ist nach Fischer *latent subtraktiv* skalierbar, wenn es reellwertige Funktionen θ und ε und reelle Konstanten r und s gibt, so daß gilt

$$P(a, v) = \frac{\exp[r\,(\theta(a) - \varepsilon(v)) + s]}{1 + \exp[r\,(\theta(a) - \varepsilon(v)) + s]}. \tag{3.40}$$

Es ist klar, daß in diesem Fall eine Skalentransformation der Form $\theta \mapsto r'\theta + s'$ durch geeignete Wahl der Konstanten r und s kompensiert werden kann. Daß diese Darstellung des Rasch-Modells nicht optimal ist, folgt

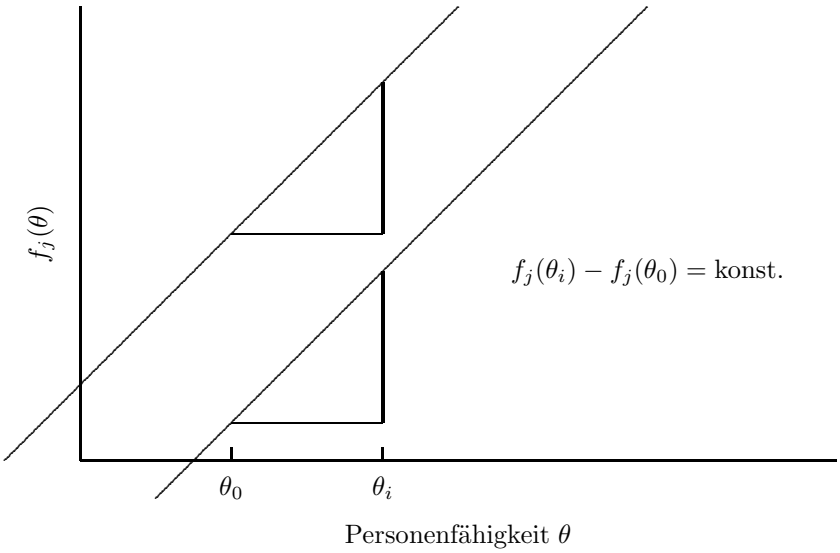

Abb. 3.5. Veranschaulichung der Repräsentationsbedingung (3.36) für das Rasch-Modell. Die Ordinate ist hier durch eine Logit-Transformation von Gleichung (3.3) verzerrt: $f_j(\theta_i) = \ln[P(U_{ij} = 1 \,|\, \theta_i, \varepsilon_j)/[(1 + P(U_{ij} = 0 \,|\, \theta_i, \varepsilon_j))]$. Die Aufgabenkennlinien werden dadurch zu parallelen Geraden. Die Repräsentationsbedingung verlangt, daß für jedes Item die vertikalen Differenzen $[f_j(\theta_i) - f_j(\theta_0)]$ gleich sind.

aus der Tatsache, daß Gleichung (3.40) die gleichen empirischen Restriktionen impliziert wie Gleichung (3.35).[4] Theorem 3.2 zeigt aber, daß diese zur Konstruktion von Differenzenskalen ausreichen. Natürlich ist es grundsätzlich immer möglich, für Daten, die einen bestimmten Skalierbarkeitsgrad erfüllen, Skalen mit einem geringeren Skalenniveau zu konstruieren. Es ist aber klar, daß dabei Information verschenkt wird.

3.8.3 Das Birnbaum-Modell

Definition 3.2. Ein Paarvergleichssystem $\langle A{\times}V, P\rangle$, in dem A und V Mengen und P eine Abbildung von $A{\times}V$ in das offene, reelle Intervall $(0, 1)$ ist, heißt genau dann *Birnbaum-skalierbar*, wenn es reellwertige Funktionen θ, α und ε gibt, die auf A bzw. V definiert sind, so daß für alle a in A und alle v in V gilt

$$P(a, v) = \frac{\exp[\alpha(v)(\theta(a) - \varepsilon(v))]}{1 + \exp[\alpha(v)(\theta(a) - \varepsilon(v))]}. \tag{3.41}$$

Das Tripel $\langle \theta, \alpha, \varepsilon\rangle$ heißt dann *Birnbaum-Repräsentation für* $\langle A{\times}V, P\rangle$.

Theorem 3.3. *Ein Paarvergleichssystem* $\langle A{\times}V, P\rangle$*, in dem* A *und* V *Mengen und* P *eine Abbildung von* $A{\times}V$ *in das offene, reelle Intervall* $(0, 1)$ *ist, ist genau dann Birnbaum-skalierbar, wenn für alle* a*,* b*,* c *in* A *und alle* v*,* w *in* V *gilt:*

$$\frac{L[P(a, v)] - L[P(b, v)]}{L[P(c, v)] - L[P(b, v)]} = \frac{L[P(a, w)] - L[P(b, w)]}{L[P(c, w)] - L[P(b, w)]}. \tag{3.42}$$

Sind ferner $\langle \theta, \alpha, \varepsilon\rangle$ *und* $\langle \theta', \alpha', \varepsilon'\rangle$ *zwei Birnbaum-Repräsentationen für* $\langle A{\times}V, P\rangle$*, dann gibt es reellwertige Konstanten* r *und* s*, so daß* $\theta'(a) = r\theta(a) + s$*,* $\varepsilon'(v) = r\varepsilon(v) + s$ *und* $\alpha' = \alpha/r$ *für alle* a *in* A *und alle* v *in* V*.*

[4] Fischer (1988b) beweist zwei Theoreme. Im Theorem 1 zeigt er die Intervallskalierbarkeit von Modellen der Form (3.31). Im Theorem 2 werden den Voraussetzungen von Theorem 1 weitere Annahmen hinzugefügt und daraus das Rasch-Modell der Form (3.35) abgeleitet. Er übernimmt dann die Ergebnisse von Theorem 1 für Theorem 2 ohne zu berücksichtigen, daß Theorem 2 von stärkeren Restriktionen ausgeht als Theorem 1.

Beweis. Zur Definition der Parameter θ, ε und α wählt man zwei Personen b_0 und b_1 und eine Aufgabe w_0 derart, daß $P(b_1, w_0) \neq 1/2$ und $P(b_1, w_0) \neq P(b_0, w_0)$. Dann definiert man die Parameter in folgender Weise:

$$\begin{aligned} \theta(a) &= \frac{L[P(a, w_0)] - L[P(b_0, w_0)]}{L[P(b_1, w_0)] - L[P(b_0, w_0)]} \\ \alpha(v) &= L[P(b_1, v)] - L[P(b_0, v)] \\ \varepsilon(v) &= -L[P(b_0, v)]/\alpha(v). \end{aligned}$$

Der Rest des Beweises ist dann eine einfache Übungsaufgabe. □

Abbildung 3.6 auf Seite 70 zeigt eine graphische Veranschaulichung der Repräsentationsbedingung (3.42).

3.8.4 Spezifische Objektivität

Das Konzept der spezifischen Objektivität wurde in den 60er Jahren von G. Rasch in die Psychologie eingeführt (Rasch, 1960). Er wollte mit diesem Konzept eine Charakterisierung wissenschaftlicher Aussagen ermöglichen, die für Sozial- und Naturwissenschaften gleichermaßen geeignet ist. Der Kern des Konzepts ist eine Invarianzforderung: Vergleiche zwischen einzelnen Objekten aufgrund bestimmter Verfahren sollen unabhängig davon sein, welche anderen Objekte noch zur Anwendung des Verfahrens benötigt werden. Will man die Breite eines Tisches mit der Breite einer Tür vergleichen, durch die der Tisch befördert werden soll, dann darf das Ergebnis des Vergleiches nicht von dem als Maßstab verwendeten Objekt abhängen. Dem entsprechend soll der Vergleich einer bestimmten Fähigkeit zweier Personen unabhängig davon sein, welches spezielle Meßinstrument zur Bestimmung der Fähigkeit benutzt wird.

Die Forderung der spezifischen Objektivität, daß im Rahmen eines bestimmten Systems Aussagen über einzelne Objekte oder über Beziehungen zwischen Objekten nicht von den anderen Elementen des Systems abhängen sollen, ist sicher eine wünschenswerte Eigenschaft wissenschaftlicher Aussagen. Von Rasch (1977) und anderen Autoren wurden jedoch aus dieser Forderung sehr weitreichende Konsequenzen abgeleitet, die nicht ungeprüft übernommen werden können. So schreibt Fischer (1974): „Die besprochenen Varianten des linearen logistischen Modells sind nicht ein Modelltyp neben vielen anderen gleichwertigen oder vielleicht besseren, sondern der einzige, welcher die Eigenschaft der Stichprobenunabhängigkeit, der spezifischen Objektivität von Aussagen über Medienwirkungen, besitzt. Dies verleiht ihm, gerade angesichts des Gegensatzes, in dem er mit den geläufigen Theorien über die Wirkung von Kommunikationen steht, und angesichts der vielen scheinbar widersprüchlichen empirischen Resultate, eine besondere Bedeutung; erweist sich nämlich das Modell empirisch als unanwendbar, dann sind spezifisch objektive Aussagen über diesen Problemkreis nicht möglich“ (Fischer, 1974, S. 421).

Um von einer verhältnismäßig allgemeinen Begriffsbestimmung, wie sie oben für das Konzept der spezifischen Objektivität gegeben wurde, zu einer so weitreichenden, auch formal begründbaren, Schlußfolgerung zu kommen, sind einige zusätzliche, formale Annahmen zu machen. Wir werden sehen, daß in diesen Annahmen sehr starke, empirisch nicht begründbare Forderungen stecken, die in keinem inhaltlich begründeten Zusammenhang zum Konzept der spezifischen Objektivität stehen. Verzichtet man auf diese technischen Forderungen, dann läßt sich zweierlei zeigen:

1. Es gibt neben dem Rasch-Modell andere, wesentlich verschiedene Modelle, die spezifisch objektive Aussagen zulassen.
2. Das Konzept der spezifischen Objektivität ist verwandt mit dem Bedeutsamkeitskonzept der Meßtheorie. Spezifisch objektive Aussagen lassen sich als empirisch bedeutsame Aussagen in Meßstrukturen definieren.

Wir haben bisher „spezifische Objektivität“ nur verbal umschrieben. Um die Frage zu untersuchen, welche Theorien spezifisch objektive Aussagen zulassen, ist eine genauere Definition notwendig. Wir werden uns dabei im

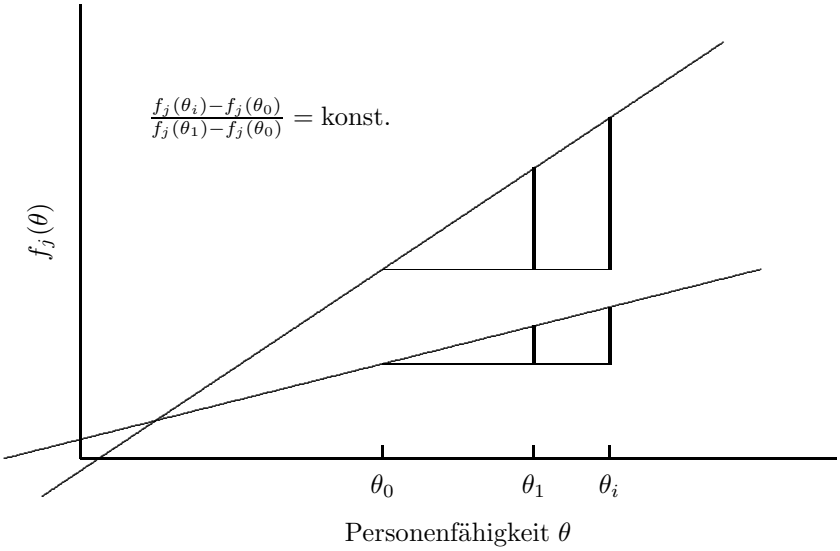

Abb. 3.6. Veranschaulichung der Repräsentationsbedingung (3.42) für das Birnbaum-Modell. Die Ordinate ist hier wie in Abb. 3.5 als logistische Funktion der Antwortwahrscheinlichkeiten aufgetragen. Die Aufgabenkennlinien werden dadurch zu Geraden, die sich in der Steigung unterscheiden können. Die Repräsentationsbedingung verlangt, daß für jedes Item die Quotienten aus den markierten vertikalen Differenzen $[f_j(\theta_i) - f_j(\theta_0)]$ und $[f_j(\theta_1) - f_j(\theta_0)]$ gleich sind. Wie Gleichung (3.45) zeigt, überträgt sich diese Unabhängigkeit des Quotienten von Logitdifferenzen auf die Personenskala θ. Die Graphik zeigt dies ebenfalls.

wesentlichen an die Definition halten, die von Irtel (1995) in Anlehnung an Rasch (1977) gegeben wird.

Wir gehen von einem Paarvergleichssystem $\langle A{\times}V, P\rangle$ aus. Dabei ist etwa A eine Menge von Personen und V eine Menge von Testaufgaben. $A{\times}V$ ist dann die Menge aller Paare von Personen und Testaufgaben. P soll eine Abbildung von $A{\times}V$ in eine Teilmenge R der reellen Zahlen sein. P kann etwa eine Indikatorfunktion sein, deren Funktionswert $P(a, v)$ für a in A und v in V angibt, ob die Person a die Aufgabe v richtig löst oder nicht. $P(a, v)$ kann auch die Wahrscheinlichkeit sein, mit der die Person a die Aufgabe v löst. Für die folgenden Überlegungen ist die empirische Interpretation von P nur insofern von Bedeutung, als sich dadurch der Wertebereich von P ändern kann.

Das System $\langle A{\times}V, P\rangle$ nimmt aus meßtheoretischer Sicht die Rolle des empirischen Relativs ein. Das Paarvergleichssystem enthält die gesamte Empirie, über die von der darauf aufbauenden Theorie Aussagen gemacht werden können. Die Mengen A und V bilden den Gegenstandsbereich der Theorie, und die Funktion P beinhaltet die Daten. Strukturen, die auf dem Kreuzprodukt aus zwei oder mehr verschiedenen Mengen formuliert sind, werden in der Meßtheorie *verbundene Meßstrukturen* genannt (Luce & Tukey, 1964). Sie sind zur Formalisierung mehrfaktorieller Experimente geeignet. Spezifisch objektive Aussagen sollen Aussagen über Personen, also Elemente der Menge A, sein, die unabhängig davon sind, welche Aufgaben, also Elemente aus V, in der Aussage vorkommen. Zur Formalisierung des Konzepts der spezifischen Objektivität ist deshalb eine zweifaktorielle Struktur notwendig.

Eine Formalisierung von „spezifisch objektiv“ verlangt, den Begriff des „Vergleichs“ bezogen auf das System $\langle A{\times}V, P\rangle$ genauer zu spezifizieren. Rasch (1977) benützt hierzu eine Funktion K, die auf Wertepaaren von P definiert ist und Werte aus einer nicht weiter bestimmten Menge B annimmt. Das Ergebnis des Vergleichs zweier Objekte a und b aus A mit Hilfe des Objekts v aus V ist dann der Funktionswert $K[P(a, v), P(b, v)]$. Wir betrachten hier nicht nur Wertepaare von P, sondern n-Tupel endlicher Länge.

DEFINITION 3.3. Sei $\langle A{\times}V, P\rangle$ ein Paarvergleichssystem und K eine Abbildung von n-Tupeln ($n > 1$ und endlich) von Werten von P in eine Menge B, so daß für jedes n-Tupel der Form $[P(a,v), P(b,w), \ldots]$ der Funktionswert $K[P(a,v), P(b,w), \ldots]$ definiert ist. Die Funktion K ist genau dann eine *spezifisch objektive Komparatorfunktion für A in $\langle A{\times}V, P\rangle$*, wenn $K[P(a,v), P(b,w), \ldots]$ für alle a, b, ... in A unabhängig von v, w in V ist. Die Funktion K ist genau dann eine *spezifisch objektive Komparatorfunktion für V in $\langle A{\times}V, P\rangle$*, wenn $K[P(a,v), P(b,w), \ldots]$ für alle v, w, ... in V unabhängig von allen a, b, ... in A ist.

Die Komparatorfunktion K ist also für Objekte A spezifisch objektiv, wenn der Funktionswert von K für n Objekte aus A und ein Objekt aus V für alle beliebigen v identisch ist, also nicht von v abhängt. Welches spezielle Objekt v zur Berechnung der Komparatorfunktion benutzt wird, ist deshalb irrelevant. An dieser Definition der spezifischen Objektivität sind folgende Punkte zu beachten:

1. Spezifische Objektivität ist als Eigenschaft einer Funktion definiert, nicht als Eigenschaft einer Theorie oder eines Modells. Innerhalb einer Theorie, die spezifisch objektive Komparatorfunktionen enthält, gibt es auch Komparatorfunktionen, die nicht spezifisch objektiv sind.

2. Jede spezifisch objektive Komparatorfunktion ist auf das Paarvergleichssystem beschränkt, innerhalb dessen sie definiert wird. Dies war der Grund für Rasch, die Eigenschaft mit dem Attribut „spezifisch“ zu versehen.

3. Die spezifische Objektivität einer Komparatorfunktion ist für beide Faktoren der Paarvergleichsstruktur unabhängig definiert. Wir werden sehen, daß es Modelle gibt, bei denen nur für einen der beiden Faktoren spezifisch objektive Komparatorfunktionen der oben definierten Art existieren.

3.8.5 Spezifisch objektive Meßmodelle

Das Rasch-Modell

Die Multiplikationsbedingung (3.36) für das Rasch-Modell erfüllt genau die Definition einer spezifisch objektiven Komparatorfunktion, sowohl für die Personen als auch für die Aufgaben. Setzt man in die linke Seite von (3.36) die Modellgleichung (3.35) ein, dann erhält man

$$\frac{P(a,v)}{1-P(a,v)}\,\frac{1-P(b,v)}{P(b,v)} = \theta(a) - \theta(b).$$

Der Wert der Komparatorfunktion ist also die Skalenwertdifferenz der beiden Personen. Dies zeigt, daß mit Hilfe des Rasch-Modells itemunabhängige Aussagen über die Skalenwertdifferenz von zwei Personen möglich sind. Entsprechendes gilt für den Vergleich zweier Aufgaben. Dies ist auch Ausdruck der Tatsache, daß es sich bei den Skalen des Rasch-Modells um Differenzenskalen handelt.

Spezifisch objektive Vergleiche mit Ordinalskalen

Aus meßtheoretischer Sicht ist das Rasch-Modell nur ein spezielles Dekompositionsmodell für verbundene Paarvergleichssysteme. Es gibt andere Modelle, die ebenfalls spezifisch objektive Aussagen zulassen. Ein Modell dieser Art ist das ordinale Modell, das von Mokken (1971) und Irtel und Schmalhofer (1982) für psychodiagnostische Messungen vorgeschlagen wurde. Auch dieses Modell geht von einem Paarvergleichssystem $\langle A{\times}X, P\rangle$ aus und gibt die notwendigen und hinreichenden Bedingungen an, für die Dekomposition von P in Funktionen θ und ε, die eindeutig bis auf streng monotone Transformationen sind. Es gilt dann

$$P(a,v) \geq P(b,v) \;\text{ gdw. }\; \theta(a) \geq \theta(b) \tag{3.43}$$

und

$$P(a,v) \geq P(a,w) \;\text{ gdw. }\; \varepsilon(v) \leq \varepsilon(y). \tag{3.44}$$

Auch dieses Modell besitzt spezifisch objektive Komparatorfunktionen für Personen und Aufgaben:

$$K_A[P(a,v),P(b,v)] = \begin{cases} 1 \text{ falls } P(a,v) \geq P(b,v) \\ 0 \text{ sonst} \end{cases}$$

$$K_X[P(a,v),P(a,w)] = \begin{cases} 1 \text{ falls } P(a,v) \leq P(a,w) \\ 0 \text{ sonst.} \end{cases}$$

Wie man sofort sieht, erfüllen diese Komparatorfunktionen die Definition 3.3. Das Unabhängigkeitsmodell ist daher ein Beispiel für ein Modell, das spezifisch objektive Komparatorfunktionen besitzt, aber nicht isomorph zum Rasch-Modell ist.

Das Birnbaum-Modell

Ein Meßmodell, bei dem der Personenmenge eine Intervallskala zugeordnet wird, ist das Birnbaum-Modell (Birnbaum, 1968) nach Gleichung (3.41). Die numerische Aussage

$$\frac{\theta(a)-\theta(b)}{\theta(c)-\theta(b)} = r$$

ist hier im Sinne der Meßtheorie empirisch bedeutsam. Diese Aussage ist unabhängig von den zur Messung benutzten Testaufgaben, da in ihr kein Aufgabenparameter vorkommt. Sie ist auch übersetzbar in eine Aussage über die Paarvergleichsfunktion P. Es gilt nämlich

$$\frac{\theta(a)-\theta(b)}{\theta(c)-\theta(b)} = \frac{L[P(a,v)]-L[P(b,v)]}{L[P(c,v)]-L[P(b,v)]}, \tag{3.45}$$

wobei L die Umkehrfunktion (3.37) der logistischen Funktion ist. Wegen der Repräsentationsbedingung (3.42) ist (3.45) aber unabhängig von v und damit ist

$$K[P(a,v),P(b,v),P(c,v)] = \frac{L[P(a,v)]-L[P(b,v)]}{L[P(c,v)]-L[P(b,v)]}, \tag{3.46}$$

eine spezifisch objektive Komparatorfunktion für die Personen. Das Birnbaum-Modell erlaubt daher im Sinne von Definition 3.3 spezifisch objektive Aussagen über Personen.

Es ist hier sehr wichtig, die Bedeutung des Konzepts der spezifischen Objektivität im Sinne von Definition 3.3 zu berücksichtigen, denn in vielen Publikationen wird das Birnbaum-Modell als „nicht spezifisch objektiv“ behandelt. So schreibt etwa Kubinger (1988) über das Birnbaum-Modell: „Wegen der Existenz der Parameter α wird dieses probabilistische Modell der Eigenschaft der Spezifischen Objektivität verlustig“ (Kubinger, 1988, S. 40). Allerdings betrachtet Kubinger als Vergleichsfunktionen nur Differenzen der Skalenwerte von 2 Personen, und diese sind eben bei keiner Intervallskala bedeutsam, da sie nicht invariant gegenüber den zulässigen Transformationen der Skala sind. Wie Gleichung (3.45) zeigt, besteht für Vergleiche von 3 Personen diese Beschränkung nicht mehr. Sie besteht im übrigen auch nicht für ordinale Vergleiche der Skalenwerte des Birnbaum-Modells, da die Ordnung der Parameterwerte zweier Personen bei zulässigen Transformationen einer Intervallskala erhalten bleibt. Eine Aussage wie „$\theta(a) > \theta(b)$“ ist also auch bei Skalenwerten eines Birnbaum-Modells spezifisch objektiv. Es läßt sich nämlich leicht zeigen, daß Bedingung (3.43) auch im Birnbaum-Modell gilt. Bedingung (3.44) gilt dagegen nicht, denn die Aufgaben haben 2 Parameter, und 2-dimensionale Objekte lassen sich nicht einfach anordnen.

Es gibt im Birnbaum-Modell auch spezifisch objektive Vergleiche von Aufgaben. Da jede Aufgabe zwei Parameter hat, sind auch zwei unabhängige Formen quantitativer Aussagen möglich. Die Trennschärfe zweier Aufgaben läßt sich mit der Funktion

$$K_X[P(a,v),P(b,v),P(a,w),P(b,w)] := \frac{L[P(a,v)]-L[P(b,v)]}{L[P(a,w)]-L[P(b,w)]} \tag{3.47}$$

vergleichen. Es ist leicht zu zeigen, daß dieser Vergleich unabhängig von den Personen a und b ist, die in ihm vorkommen. Quantitative Aussagen über die Aufgabenschwierigkeit sind nur über Aufgabentripel möglich. Man benötigt dazu wiederum 2 Personen, die Vergleichsfunktion

$$K'_X[P(a,v),P(b,v),P(a,w),P(b,w),P(a,x),P(b,x)] := \frac{\dfrac{L[P(a,v)]}{L[P(a,v)]-L[P(b,v)]}-\dfrac{L[P(a,w)]}{L[P(a,w)]-L[P(b,w)]}}{\dfrac{L[P(a,x)]}{L[P(a,x)]-L[P(b,x)]}-\dfrac{L[P(a,w)]}{L[P(a,w)]-L[P(b,w)]}} \tag{3.48}$$

ist aber unabhängig davon, welche speziellen Personen a und b in ihr vorkommen.

3.9 Zulässige und nicht zulässige Transformation der Skalenwerte

Wir betrachten noch einmal ein Rasch-skalierbares Paarvergleichssystem im Sinne von Definition 3.1 und Theorem 3.2. Die Modellgleichung schreiben wir in Form von (3.31), wo die logistische Funktion durch F abgekürzt ist:

$$P(a,v) = F[\theta(a) - \varepsilon(v)]. \tag{3.49}$$

Diese Gleichung zeigt, daß zulässige Transformationen der Skalen θ und ε den Wert der Differenz $\theta(a) - \varepsilon(v)$ nicht ändern dürfen. Es ist aber auch möglich, Transformationen auf θ und ε anzuwenden, die durch eine Änderung der Funktion F kompensiert werden, so daß der gesamte Funktionswert, der ja durch das „Datum“ P vorgegeben ist, konstant bleibt. Das einfachste Beispiel hierfür sind Transformationen der Form $\theta \to r\,\theta + s$ und $\varepsilon \to r\,\varepsilon + s$. Diese können durch die Änderung $F(x) \to F(x/r)$ kompensiert werden:

$$F[(r\,\theta(a) + s - (r\,\varepsilon(v) + s))/r] = F[\theta(a) - \varepsilon(v)].$$

Allgemein führen alle Transformationen h und k der Skalen θ und ε, zu denen eine Kompensationsfunktion K existiert, so daß

$$\theta - \varepsilon = K[h(\theta), k(\varepsilon)] \tag{3.50}$$

gilt, zu neuen Skalen $h(\theta)$ und $k(\varepsilon)$, die ebenfalls als Rasch-Repräsentation des Paarvergleichssystems $\langle A \times X, P\rangle$ geeignet sind. Ein häufig benutztes Beispiel solcher Transformationen ist durch die Funktionen

$$h(x) = \exp(x) \tag{3.51}$$

$$k(x) = \exp(-x) \tag{3.52}$$

$$K(x,y) = \ln(xy)$$

definiert. Die Modellgleichung (3.35) erscheint dann in der Form

$$P(a,v) = \frac{\theta'(a)\varepsilon'(v)}{1 + \theta'(a)\varepsilon'(v)}, \tag{3.53}$$

wobei $\theta' = h(\theta) = \exp(\theta)$ und $\varepsilon' = k(\varepsilon) = \exp(-\varepsilon)$.

Wichtig ist hier vor allem, daß die Transformationen (3.51) und (3.52) *keine* zulässigen Transformationen sind und damit auch nichts über das Skalenniveau aussagen. Die Möglichkeit, diese Transformationen zu definieren, hat keine Folgen für das Skalenniveau, da diese Transformationen zu einer Änderung der repräsentierenden Relation, also des Modells führen. Man sieht auch sofort, daß Gleichung (3.53) andere zulässige Transformationen hat als Gleichung (3.35). Von Colonius (1980) wurde allerdings gezeigt, daß alle Systeme, die als Lösungen der Funktionalgleichung (3.50) in Frage kommen, 1-parametrige zulässige Transformationen haben. Alle diese Skalen sind also isomorph zu einer Verhältnisskala.[5]

[5]Eine Verwechslung der Lösungen von Gleichung (3.50) mit den zulässigen Transformationen eines bestimmten Modells führte bei manchen Autoren zur Annahme, die Skalen des Rasch-Modells seien nur Ordinal- (Wottawa, 1979) oder Intervallskalen (Fischer, 1988).

3.10 Wann messen zwei verschiedene Tests die gleiche Eigenschaft?

Einige interessante Überlegungen zur Frage, welchen theoretischen Status Skalen auf der Grundlage von logistischen Modellen besitzen, werden von Wottawa (1980) angestellt. Angenommen, man hat einen Test, der das Birnbaum-Modell erfüllt, und die einzelnen Testaufgaben zerfallen in zwei Gruppen, wobei alle Aufgaben in einer Gruppe die gleichen Trennschärfeparameter besitzen, die Trennschärfeparameter der beiden Gruppen jedoch verschieden sind. Ein solcher Test erfüllt das Rasch-Modell nicht. Betrachtet man allerdings jede einzelne der beiden Gruppen für sich alleine als Test, dann erfüllt jeder dieser beiden Tests das Rasch-Modell. Die an dieser Überlegung zu diskutierende Frage ist nun, ob die beiden Rasch-Tests die gleiche Fähigkeit messen oder nicht.

Im Rahmen der klassischen Testtheorie kann man definieren: *Zwei Tests im Sinne der klassischen Testtheorie messen genau dann die gleiche Fähigkeit, wenn sie parallelisierbar sind.* Im Rahmen der logistischen Testtheorie könnte man etwa definieren: *Zwei Tests im Sinne der logistischen Testtheorie messen genau dann die gleiche Fähigkeit, wenn ihre Vereinigung wieder ein Test im Sinne der logistischen Testtheorie ist.* Das Problem wäre damit auf die Frage verlagert, wann ein Test ein Test im Sinne der logistischen Testtheorie ist. Läßt man hier nur Tests nach dem Birnbaum- und dem Rasch-Modell zu, so kann man die obige Definition noch genauer fassen: *Zwei Tests im Sinne der logistischen Testtheorie messen genau dann die gleiche Fähigkeit, wenn ihre Vereinigung ein Test im Sinne des Birnbaum-Modells ist.*

Es liegt also immer dann, wenn die Vereinigung von zwei Rasch-Tests einen Test nach dem Birnbaum-Modell ergibt, eine Messung der gleichen Fähigkeit vor. Die Messung geschieht bei den beiden Rasch-Tests eben nur mit unterschiedlicher Trennschärfe. Wesentlich ist, daß die Personenparameter aus den beiden Tests ineinander überführt werden können. Allerdings stellt die Funktion, die diese Überführung leistet, *keine* zulässige Skalentransformation des Rasch-Modells dar. Einige einfache Überlegungen zeigen, daß der Personenparameter $\theta_1(a)$ einer Person a im ersten Test mit dem der gleichen Person im zweiten Test $\theta_2(a)$ über eine Ähnlichkeitsabbildung zusammenhängt, vorausgesetzt, bei beiden Skalen wurde die gleiche Person als Nullpunkt gewählt. Es muß also eine Person b geben für die gilt $\theta_1(b) = \theta_2(b) = 0$. Dann gibt es unter den genannten Voraussetzungen einen Faktor s derart, daß $\theta_1(a) = s\,\theta_2(a)$.

Die oben gegebene Definition dafür, wann zwei Tests die gleiche Fähigkeit messen, läuft darauf hinaus, daß Tests im Sinne der logistischen Testtheorie genau dann die gleiche Fähigkeit messen, wenn die Personenparameter durch eine Abbildung der Form $f(x) = s\,x$ ineinander überführt werden können. Läßt man die Forderung nach der Normierung auf die gleiche Person fallen, so ergibt sich eine affine Abbildung der Form $f(x) = s\,x + r$ als Bedingung. Interessanterweise ist diese Bedingung identisch mit der, die in Abschnitt 2.6 für die Parallelisierbarkeit von Tests im Sinne der klassischen Testtheorie gefordert wurde.

Es stellt sich allerdings die Frage, wieso nur affine Abbildungen zulässig sein sollen, um die Gleichheit gemessener Fähigkeiten zu definieren. Dieses Problem wird bereits von Lord und Novick (1968, S. 219) angesprochen. Sie weisen darauf hin, daß grundsätzlich auch Messungen parallelisierbar sein können, die über allgemeine Polynome ineinander transformierbar sind. Zur Identifikation der Parameter einer solchen Transformation genügen allerdings die ersten beiden Momente einer Verteilung nicht, sondern es sind auch höhere Momente zu berücksichtigen.

Auch im Rahmen logistischer Modelle gibt es keinen besonderen Grund sich auf affine Transformationen zu beschränken. Es ist leicht zu zeigen, daß alle eineindeutigen Transformationen zugelassen werden können. Das würde also heißen, daß *zwei Tests genau dann die gleiche Fähigkeit messen, wenn die Personenparameter durch eine eineindeutige Transformation ineinander überführt werden können.* Daß diese Definition sinnvoll ist, kann man

sich leicht an folgender Überlegung klarmachen: Angenommen zwei Tests würden unterschiedliche Fähigkeiten θ_1 und θ_2 messen. Dann müßte es Personen geben, die in der Fähigkeit θ_1 die gleiche, in der Fähigkeit θ_2 aber unterschiedliche Ausprägungen haben. Damit gibt es aber keine eineindeutige Abbildung von θ_1 auf θ_2. Gibt es umgekehrt eine solche Abbildung, dann kann für jede Person aus dem Parameter θ_1 der Parameter θ_2 eindeutig berechnet werden. Eine unabhängige Variation der beiden Parameter ist daher nicht möglich.

Übungsaufgaben

1. Zeigen Sie daß im Birnbaum-Modell (3.41) Bedingung (3.43) gilt.
2. Zeigen Sie, daß die Funktionen K_X und K'_X in Gleichung (3.47) und (3.48) personenunabhängig sind.

4 Entscheidungstheorie

4.1 Elemente psychodiagnostischer Entscheidungen

Psychologische Diagnostik vermittelt die methodischen Grundlagen für psychologische Entscheidungen und Prognosen, in die neben allgemeinpsychologischen Erkenntnissen vor allem individuelle Daten der Personen eingehen, die von den Entscheidungen betroffen sind oder deren Verhalten vorhergesagt werden soll. Diagnostische Fragestellungen treten in allen Teilgebieten der Angewandten Psychologie auf. Die gemeinsamen methodischen Probleme dieser Fragestellungen werden in der Psychologischen Diagnostik behandelt. Das Ziel, Prognosen aufgrund individueller Daten zu treffen, führt zu mehreren Teilproblemen:

Testtheorie: Die Theorie der Konstruktion psychologischer Tests ist die wesentliche Grundlage für das Einbringen individueller Merkmale in eine Prognose oder Entscheidung.

Multivariate Verfahren: Durch die Verwendung mehrerer Variablen kann die Güte einer Prognose häufig wesentlich verbessert werden.

Klassifikation: In vielen Fällen können ähnliche Merkmalsmuster zu Gruppen zusammengefaßt werden, für deren Mitglieder eine gemeinsame Behandlung möglich ist.

Neben diesen Teilgebieten kann die *Entscheidungstheorie* als ein übergreifendes Konzept der Psychologischen Diagnostik betrachtet werden. Angewandte Diagnostik führt immer zu Entscheidungen. Sei es die Zulassung zu einem bestimmten Ausbildungsgang, die Aufnahme in eine therapeutische Fallgruppe oder die Empfehlung für eine bestimmte Berufslaufbahn, angewandte Diagnostik ist kein Selbstzweck, sondern ist mit Konsequenzen verbunden, die durch diagnostische Entscheidungen herbeigeführt werden. Die Entscheidungstheorie als übergreifendes Konzept analysiert den gesamten diagnostischen Prozess und stellt so einen Rahmen für dessen Optimierung dar.

Die entscheidungstheoretische Analyse der Psychologischen Diagnostik benutzt die Methoden der statistischen Entscheidungstheorie (Bamberg, 1972; Berger, 1980; Chernoff & Moses, 1959; Lindgren, 1971, 1976). Zu einer Entscheidungssituation, wie sie von der statistischen Entscheidungstheorie analysiert wird, gehören folgende Elemente:

- Der *Zustandsraum* Θ ist die Menge von Zuständen des Systems, über das eine Entscheidung getroffen werden soll. Das wesentliche Merkmal der Zustände besteht darin, daß sie unabhängig vom Entscheidungsprozess sind. Ihr Vorliegen kann nicht durch Personen festgestellt werden, die selbst am Entscheidungsprozess teilnehmen, da sonst eine objektive Validierung des Entscheidungsverfahrens unmöglich würde.

- Der *Entscheidungsraum* A ist die Menge der zur Verfügung stehenden Alternativen. Als Ergebnis des Entscheidungsverfahrens wird immer eine der Alternativen aus A bestimmt. Besteht die Aufgabe eines Entscheidungsverfahrens darin, die Zustände des Systems zu identifizieren, dann korrespondieren die Alternativen mit den Zuständen. Auch in diesen Fällen ist aber in der Regel eine der Alternativen „weitere Informationen einholen“.

- Die *Schadensfunktion* s gibt für jedes Paar (θ, a) von Zuständen aus Θ und Alternativen aus A den Schaden $s(\theta, a)$ an. Sie beschreibt damit den Schaden, der auftritt, wenn beim Vorliegen des Zustands θ die Entscheidung a getroffen wird.
- Der *Stichprobenraum* $\mathfrak{X}$ ist die Menge aller zur Verfügung stehenden Daten. Diese liegen als Zufallsvariable X vor, deren Realisierungen Informationen über die Zustände des Systems enthalten sollen.
- Werden Entscheidungen aufgrund von Daten getroffen, dann muß die Menge Δ aller *Entscheidungsregeln* betrachtet werden. Dies sind die Verfahren δ, die jedem Datum X eine bestimmte Entscheidung a zuordnen. Diese Zuordnung kann auch mit Hilfe von Wahrscheinlichkeiten geschehen, man spricht dann von *randomisierten* oder *gemischten* Entscheidungen.
- Die Auswahl einer bestimmten Entscheidungsregel erfolgt nach dem *Optimalitätskriterium* K, mit dessen Hilfe aus der Menge der möglichen Entscheidungsregeln Δ eine als im Sinne des Kriteriums optimal ausgezeichnet werden kann.

Die explizite und formale Angabe aller einzelnen Elemente einer Entscheidungssituation, die schließlich zu einer bestimmten Entscheidung führen, dient nicht zuletzt der Transparenz des Entscheidungsverfahrens. Sowohl dem Auftraggeber als auch dem Betroffenen ist es damit möglich, die Gründe für einzelne Entscheidungen nachzuvollziehen, bzw. die Ziele des Entscheidungsverfahrens zu definieren.

4.1.1 Diagnostizierbare Zustände

Die entscheidungstheoretischen „Zustände“ sind die psychodiagnostischen Kategorien, über deren Vorliegen eine Entscheidung getroffen werden soll. Dabei kann es sich um aktuell gegebene Kategorien handeln, wie etwa bestimmte Krankheitsbilder, oder es kann sich um Merkmale handeln, deren Verifikation erst in der Zukunft erfolgen kann, wie etwa der erfolgreiche Besuch einer höheren Schule oder ein bestimmter Studienabschluß. Solche Zustände dürfen nicht mit Persönlichkeitsmerkmalen verwechselt werden. Auch ein Zustand wie „geeignet für eine bestimmte therapeutische Maßnahme“ ist denkbar. Der Begriff „Zustand“ bezieht sich also nicht in jedem Fall auf etwas „Psychisches“, sondern wird als abstrakte Kategorie für das betrachtet, was durch das Entscheidungsverfahren diagnostiziert werden soll.

Wie bereits oben erwähnt, ist das wesentliche Merkmal eines solchen Zustandes, daß sein Vorliegen unabhängig vom Entscheidungsverfahren selbst sein muß. Ein Zustand kann nur dann Ziel einer Entscheidung sein, wenn sein Vorliegen prinzipiell auch ohne das Entscheidungsverfahren feststellbar ist. Diese Feststellung kann entweder sehr aufwendig sein, wie etwa eine mehrtägige, stationäre Beobachtung in einer Klinik, oder sie kann erst in der Zukunft möglich sein, wie etwa dann, wenn die Bewährung eines Kindes in der Schule vorhergesagt werden soll. Die vom Entscheidungsverfahren unabhängige Feststellbarkeit des „wahren“ Zustandes ist notwendig, damit die Validität des Entscheidungsverfahrens geprüft werden kann. Es wäre sinnlos, ein aufwendiges Entscheidungsverfahren durchzuführen, wenn die Güte der durch das Verfahren getroffenen Entscheidungen nicht evaluiert werden könnte.

4.1.2 Entscheidungsalternativen

Die im Entscheidungsraum zusammengefaßten Alternativen sind die Handlungen, die als Ergebnis des Entscheidungsverfahrens ausgeführt werden. Dies kann die Zuordnung zu einem bestimmten therapeutischen Verfahren, die Zuweisung zu einer Arbeitsstelle, die Anerkennung oder Verneinung der

Glaubwürdigkeit eines Zeugen vor Gericht oder eine ähnliche psychodiagnostische Entscheidung sein. Als Alternativen können auch Rückweisungsentscheidungen auftreten. Reicht die vorliegende Information nicht aus, um geforderte Mindestkriterien zu erfüllen, kann entschieden werden, daß weitere Daten zu erheben sind, ‚oder, wenn dies nicht möglich ist, das Entscheidungsverfahren abgebrochen wird. Da auch mehrstufige Entscheidungsprozesse möglich sind, können die Alternativen auf einzelnen Stufen auch aus einer definitiven Beurteilung oder einer Zuweisung zu einem weitergehenden Entscheidungsprozess bestehen.

4.1.3 Daten als empirische Grundlage von Entscheidungen

Die Psychodiagnostik ist diejenige Teildisziplin der Psychologie, die ihre Erklärungen und Vorhersagen aufgrund individueller Eigenschaften von Personen trifft. Im expliziten Bezug auf die individuellen Eigenschaften — den Persönlichkeitsmerkmalen — unterscheidet sie sich von der Allgemeinen Psychologie oder der Sozialpsychologie, bei deren Vorhersagen individuelle Merkmale eher als Störfaktoren betrachtet werden. Die individuellen Eigenschaften als Grundlage der Vorhersage von Verhalten im weitesten Sinn machen die Methoden der Identifikation solcher Merkmale zu einem zentralen Problem der Psychodiagnostik. Die Merkmale gehen über den Stichprobenraum $\mathfrak{X}$ in den Entscheidungsprozess ein. Von wesentlicher Bedeutung für eine Entscheidung ist, ob die erfaßten Merkmale Information über die zu diagnostizierenden Zustände enthalten, also die Frage nach der *Validität* der Merkmale. In der Entscheidungssituation wird die Validität der Daten mit Hilfe der bedingten Verteilungen der Merkmalsvektoren für die einzelnen Elemente des Zustandsraumes untersucht.

Ein Merkmal X kann nur dann bezüglich der Zustände θ_1 und θ_2 valide sein, wenn die Wahrscheinlichkeit $f(X \mid \theta_1)$ von X beim Vorliegen des Zustandes θ_1 und die Wahrscheinlichkeit $f(X \mid \theta_2)$ von X beim Vorliegen von Zustand θ_2 nicht gleich sind. Ist etwa die Verteilung eines bestimmten Testergebnisses in der Gruppe der überdurchschnittlich erfolgreichen Medizinstudenten gleich der in der Gruppe der unterdurchschnittlich erfolgreichen, dann eignet sich der entsprechende Test nicht zur Prognose des Studienerfolgs. Nur wenn die bedingten Verteilungen eines Merkmals unter den verschiedenen Zuständen auch verschieden sind, kann dieses Merkmal zur Diagnose des Zustandes hilfreich sein.

4.1.4 Kosten von Entscheidungen

Mit jeder Entscheidung sind Kosten verbunden. Einerseits treten bereits mit der Durchführung des Entscheidungsverfahrens zum Teil erhebliche Kosten auf, und andererseits können als Folge des Verfahrens und der getroffenen Entscheidungen sowohl für die betroffene Person, als auch für die entscheidende Institution Kosten entstehen. Diese Kosten sind häufig nicht materieller Art. Wird etwa einem Kind durch eine Fehlentscheidung eine falsche Ausbildungsrichtung empfohlen, tritt ein Schaden ein, der nicht ohne weiteres materiell zu beziffern ist. Ähnliches gilt für die Zuweisung einer Person zur falschen Therapie oder für Glaubwürdigkeitsurteile in Gerichtsverfahren.

Das Einbringen von Kostenüberlegungen in ein formal begründetes Entscheidungsverfahren setzt numerische Angaben über die Kosten voraus. Diese müssen für jedes Paar aus Zustand θ und Entscheidungsalternative a die bei diesem Paar anfallenden Kosten $s(\theta, a)$ angeben. Man nennt dies die *Schadensfunktion*. Sie gibt an, welcher Schaden oder Nutzen auftritt, wenn die Entscheidungsalternative a gewählt wird und der Zustand θ vorliegt.

Das Erstellen einer Schadensfunktion kann auch als psychologisches Skalierungsproblem aufgefasst werden. In diesem Fall ist die Schadensfunktion eine subjektive Bewertung der verschiedenen Kombinationen aus Zustand und Entscheidung. Die Bewertung kann von Betroffenen oder auch von Experten der entscheidenden Institution erhoben werden. Wie immer bei formalen Theorien und Methoden wird die empirische Interpretation des theo-

retischen Konzepts „Schadensfunktion“ durch den Formalismus der Entscheidungstheorie nicht festgelegt. Es wird lediglich verlangt, daß die Interpretation gewissen formalen Randbedingungen genügt. Im Fall der Schadensfunktion heißt dies, daß für jede Kombination aus Zustand und Entscheidungsalternative ein Zahlenwert vorliegen muß, und daß dieser Zahlenwert mindestens die Eigenschaften einer Ordinalskala besitzt, also eine Ordnung der Paare von Zuständen und Alternativen erzeugt. Einige Optimierungskriterien setzen darüber hinaus Intervallskalenniveau voraus, da sie mit Erwartungswerten arbeiten.

4.1.5 Entscheidungsregeln

Das Ziel aller Vorüberlegungen der Entscheidungstheorie ist eine Entscheidungsregel. Sie ordnet jedem Befund X eine Entscheidungsalternative a zu und stellt daher eine Funktion vom Stichprobenraum $\mathfrak{X}$ auf die Menge der Alternativen A dar. Bei allen nichttrivialen Entscheidungssituationen gibt es mehrere mögliche Entscheidungsregeln. Das eigentliche Problem der Entscheidungstheorie besteht daher darin, aus der Menge der möglichen oder in Frage kommenden Entscheidungsregeln eine bestimmte auszuwählen. Die Auswahl einer Entscheidungsregel ist problematisch, weil die Güte oder der tatsächlich auftretende Schaden bzw. Nutzen einer Entscheidung in der Regel vom vorliegenden (aber unbekannten) Zustand abhängt. Wenn dies nicht der Fall wäre, könnte man auf die Entscheidungstheorie verzichten, da dann unabhängig vom Zustand einfach diejenige Entscheidung zu wählen wäre, die den kleinsten Schaden erzeugt bzw. den größten Nutzen bringt. Für die Schadensfunktion hieße dies, daß für alle Alternativen a aus A der Funktionswert $s(\theta, a)$ für alle θ in Θ gleich wäre.

4.1.6 Optimalitätskriterien

Um bei nichttrivialen Problemen eine optimale Entscheidungsregel zu finden, wird ein Optimalitätskriterium benötigt. Wünschenswert ist ein Optimalitätskriterium, mit dessen Hilfe alle Einscheidungsregeln δ aus Δ linear geordnet werden können. In diesem Fall gibt es eine optimale Entscheidungsregel. Allerdings ermöglichen nicht alle Optimalitätskriterien immer eine lineare Ordnung der Entscheidungsregeln. Es kann vorkommen, daß gleich gute Entscheidungsregeln auftreten, zwischen denen dann nach anderen Gesichtspunkten ausgewählt werden muß. Formal ist ein Optimalitätskriterium K eine Abbildung der Menge Δ in die Menge $\mathbb{R}$ der reellen Zahlen.

Das Optimalitätskriterium und die Schadensfunktion können im Sinne von Tack (1976) als „Zielsetzungen“ betrachtet werden, da sie langfristig die Güte der Entscheidungen bestimmen. Optimalitätskriterien versuchen, die Kosten und Schäden einer Entscheidung zu minimieren.

4.1.7 Teilprobleme der Psychodiagnostik

Aus den einzelnen Elementen eines Entscheidungsproblems lassen sich mehrere Teilprobleme der entscheidungstheoretischen Diagnostik herausgreifen.

Die Identifikation von Zuständen

In der Regel handelt es sich bei den zu diagnostizierenden Zuständen um von außen vorgegebene Kategorien. Zumindest kann, wie bereits früher angedeutet wurde, die Identifikation der Zustände nicht innerhalb des Entscheidungsverfahrens geschehen, da sonst die Unabhängigkeit der Zustände von den Entscheidungen nicht mehr gewährleistet wäre. Häufig sind die Zustände Bewährungskriterien, die mit Hilfe nicht-psychologischer Methoden festgestellt werden, etwa die Rückfallhäufigkeit bei Straftätern oder der erfolgreiche Abschluß eines Studiums. Ein Zustand wie „geeignet für Therapie T“ muß ebenfalls unabhängig vom Entscheidungsverfahren definiert sein, etwa durch die Erfolgskriterien der entsprechenden Therapie. Die Identifikation der Zustände muß also innerhalb des Entscheidungsprozesses als unproblematisch betrachtet werden.

Das Erstellen einer Schadensfunktion

Nur in den wenigsten Fällen wird es sich bei den Kosten oder Schäden eines Entscheidungsproblems um bekannte und vorgegebene Beträge handeln. Für die formale Behandlung des Entscheidungsproblems ist aber eine Bewertung der Entscheidungen bei gegebenen Zuständen unerläßlich. Es wäre falsch, aus dem Fehlen einer numerischen Bewertung der Schäden zu folgern, es sei in einem solchen Fall besser, ohne Berücksichtigung der Grundsätze der Entscheidungstheorie vorzugehen. Dies würde nur bedeuten, daß die Bewertung nicht explizit gemacht wird, sondern implizit.

Tatsächlich ist eine Entscheidung nach den Grundsätzen der Entscheidungstheorie nicht möglich, wenn die Paare (θ, a) aus Zuständen und Alternativen nicht bezüglich eines Schadens- oder Nutzenkriteriums geordnet werden können. In einem solchen Fall ist es auch sinnlos, von „optimalem“ oder „rationalem“ Vorgehen zu sprechen, weil Optimalität immer eine Ordnung voraussetzt und rationales Vorgehen immer begründbar sein muß. Fehlt aber die Bewertung der Fehler oder Erfolge, gibt es auch keine Möglichkeit, etwas zu begründen.

Die Konstruktion psychologischer Meßverfahren

Im Mittelpunkt der Datengewinnung steht die Definition und Messung von Persönlichkeitsmerkmalen. Für psychodiagnostische Fragestellungen spielen zwar auch anamnestische Daten eine wichtige Rolle, ihre Erhebung ist aber erheblich weniger problematisch als die Messung von Persönlichkeitsmerkmalen, den psychologischen Daten im engeren Sinn. Die Konstruktion psychologischer Meßverfahren — üblicherweise „Tests“ genannt — stellt daher das wichtigste Teilgebiet der Psychodiagnostik dar. Der Zweck eines psychologischen Tests besteht im Rahmen der entscheidungstheoretisch orientierten Diagnostik darin, in verhältnismäßig kurzer Zeit und mit geringen Kosten diejenigen Informationen zu erheben, die für eine Diagnostizierung des Zustandes hinreichen. Die psychologischen Daten sollen also die unabhängige Bestimmung des Zustandes ersetzen, da diese wie oben erläutert, entweder sehr aufwendig oder erst in der Zukunft möglich ist.

Die Validierung der Messungen

Psychologische Daten sind für die Identifikation von Zuständen oder Vorhersagen von Verhalten nur dann brauchbar, wenn sie in irgendeiner Weise mit diesen Zuständen oder dem Verhalten zusammenhängen. Der Nachweis dafür ist durch Validierungsstudien zu erbringen. Zu diesen gehören Klassifikationsprobleme, bei denen untersucht wird, ob die psychologischen Daten geeignet sind, Personen in Gruppen einzuteilen, die den zu diagnostizierenden Zuständen entsprechen. Solche Klassifikationen erfolgen in der Regel mit multivariaten Methoden, bei denen nicht nur eine, sondern mehrere psychologische Größen benutzt werden, um Gruppen zu differenzieren, oder die Vorhersage externer Kriterien zu optimieren.

Die Bestimmung eines Optimalitätskriteriums

Die Abhängigkeit der Schadensfunktionen von den Zuständen erzwingt eine Verrechnung der Schäden zu einer eindimensionalen Größe, mit deren Hilfe die Entscheidungsregeln geordnet werden können. Dabei können verschiedene Zielsetzungen verfolgt werden. Gängige Optimierungskriterien sind die Minimierung des auf lange Sicht zu erwartenden Schadens oder die Minimierung des maximal zu erwartenden Schadens.

4.1.8 Zusammenfassung

Definition 4.1. Ein Sextupel $\langle \Theta, A, s, \mathfrak{X}, \Delta, K \rangle$, in dem Θ eine Menge von Zuständen, A eine Menge von Alternativen, s eine Abbildung von $\Theta \times A$ in $\mathbb{R}$, eine Schadensfunktion, $\mathfrak{X}$ der Stichprobenraum einer Zufallsvariablen X, Δ eine Menge von Entscheidungsregeln δ, die selbst Abbildungen von $\mathfrak{X}$

auf A oder im Fall gemischter Entscheidungen von $\mathfrak{X}$ in $[0,1]^{(|A|-1)}$ sind, und K eine Abbildung von Δ in $\mathbb{R}$, ein Optimalitätskriterium ist, heißt *Entscheidungsproblem.*

4.2 Optimale Entscheidungen ohne Daten

Die wesentliche Rechtfertigung eines psychologisch begründeten Entscheidungsverfahrens muß darin bestehen, daß die psychologische Grundlage zu besseren Entscheidungen führt als sie ohne psychologische Daten möglich wären. Als Maßstab für die Güte eines psychologischen Entscheidungsverfahrens haben daher optimale Entscheidungsverfahren zu dienen, die keine individuellen, psychologischen Daten benutzen. Solche Verfahren werden als „no data“ - Entscheidungen bezeichnet.

Wir betrachten ein Beispiel von Lindgren (1976, S. 365). Er nimmt an, die vorherzusagenden Zustände seien „es wird regnen“ (θ_1) und „es wird nicht regnen“ (θ_2). Als Handlungsalternativen stehen zur Verfügung „zu Hause bleiben“ (a_1), „ausgehen, ohne einen Schirm mitzunehmen“ (a_2) und „ausgehen und einen Schirm mitnehmen“ (a_3). Tabelle 4.1 sei eine Schadenstabelle für unser Beispiel. Die Auswahl einer optimalen Alternative setzt voraus, daß die Alternativen nach einem Optimalitätskriterium geordnet werden können. Bei den Schäden in Tabelle 4.1 ist dies nicht möglich, da die Ordnung der Alternativen für die beiden Zustände verschieden ist. Ein Optimalitätskriterium muß also jeder Alternative oder Zeile von Tabelle 4.1 genau eine Zahl zuordnen, mit deren Hilfe dann die Alternativen geordnet werden können.

Tabelle 4.1. Schadenstabelle eines Beispiels zur Entscheidung ohne Daten von Lindgren (1976). Falls Ungewissheit darüber besteht, ob „Regen“ oder „kein Regen“ eintreten wird, kann kein eindeutiges Minimum der Schadensfunktion angegeben werden, da die Ordnungen der Alternativen für die beiden zu erwartenden Zustände nicht gleich sind. Es ist daher ein Optimalitätskriterium anzugeben, das eine Alternative unabhängig von den Zuständen eindeutig auszeichnet.

	θ_1 (Regen)	θ_2 (kein Regen)
(zu Hause bleiben) a_1	4	4
(weggehen ohne Schirm) a_2	5	0
(weggehen mit Schirm) a_3	2	5

4.2.1 Der unvermeidbare Schaden

Der Schaden, der bei einer Entscheidung a auftritt, wenn der Zustand θ vorliegt, wird durch den Wert der Schadensfunktion $s(\theta, a)$ angegeben. Tritt bei einem Zustand für alle Entscheidungen ein von null verschiedener Schaden auf, dann gibt es bei diesem Zustand einen *unvermeidbaren* Schaden, der sogar dann einträte, wenn die „wahren“ Zustände bekannt wären. Will man den Vergleich verschiedener Entscheidungsregeln nur auf den vermeidbaren Schaden stützen, so ist statt der Schadensfunktion $s(\theta, a)$ die *Regretfunktion* $r(\theta, a)$ zu betrachten. Sie entsteht aus der Schadensfunktion durch Verminderung um den für einen Zustand unvermeidbaren Schaden:

$$r(\theta, a) = s(\theta, a) - \min_a s(\theta, a). \tag{4.1}$$

Für unsere weiteren Überlegungen ist es unerheblich, ob die Schadens- oder die Regretfunktion benutzt wird, wir werden daher immer von der Schadensfunktion ausgehen. Es ist aber klar, daß bei Benutzung der Regretfunktion, auch wenn die formalen Ausdrücke gleich sind, die getroffenen Entscheidungen verschieden sein können. Dies liegt daran, daß der Ausdruck $(\min s(\theta, a))$ aus Gleichung (4.1) nicht für alle Zustände θ gleich sein muß.

4.2.2 Randomisierte Entscheidungen

Für eine Reihe späterer Überlegungen sind randomisierte Entscheidungen sehr nützlich. Bezogen auf einen Entscheidungsraum $A = \{a_1, \ldots, a_k\}$ wird ein k-Tupel $(p_1, \ldots, p_k)$ mit $p_i \geq 0$ und $p_1 + \cdots + p_k = 1$ als *randomisierte* oder *gemischte Entscheidung* bezeichnet. Die Parameter p_i geben die Wahrscheinlichkeiten für die Auswahl der Alternative a_i an. Bei einer gemischten Entscheidung wird also nicht direkt eine Alternative ausgewählt, sondern es wird ein Zufallsexperiment durchgeführt, in dem Ergebnisse $e_1, \ldots, e_k$ mit den Wahrscheinlichkeiten $p_1, \ldots, p_k$ auftreten können. Eine Alternative a_i wird erst dann gewählt, wenn im Zufallsexperiment das Ergebnis e_i beobachtet wird. Die Alternativen $a_1, \ldots, a_k$ werden als *reine Entscheidungen* bezeichnet.

Ein Effekt gemischter Entscheidungen ist, daß die Schadensfunktion für jeden Zustand zu einer Zufallsvariablen wird. Man betrachtet dann den Schaden, der beim Vorliegen eines Zustandes θ zu erwarten ist:

$$\begin{aligned} s(\theta, p) &= \sum_{i=1}^{k} s(\theta, a_i) p_i \\ &= \mathscr{E}[s(\theta, a)]. \end{aligned} \tag{4.2}$$

Die ursprünglichen Alternativen erhält man durch randomisierte Entscheidungen mit extremen Wahrscheinlichkeitsfunktionen. So führt etwa $p = (1, 0, 0, \ldots)$ immer zur Alternative a_1. Bei nur zwei Zuständen können auch die Schadenserwartungen in die Schadensebene von Abbildung 4.1 mit den Achsen $s(\theta_1, *)$ und $s(\theta_2, *)$ eingezeichnet werden. Jede reine Entscheidung definiert einen Punkt dieser Ebene. Da es sich bei randomisierten Entscheidungen um konvexe Kombinationen handelt, erzeugt eine Mischung von zwei Entscheidungen alle Punkte auf der Verbindungsstrecke. Aus drei reinen Entscheidungen kann ein Dreieck in der Schadensebene bestimmt werden, wie es in Abbildung 4.1 dargestellt ist.

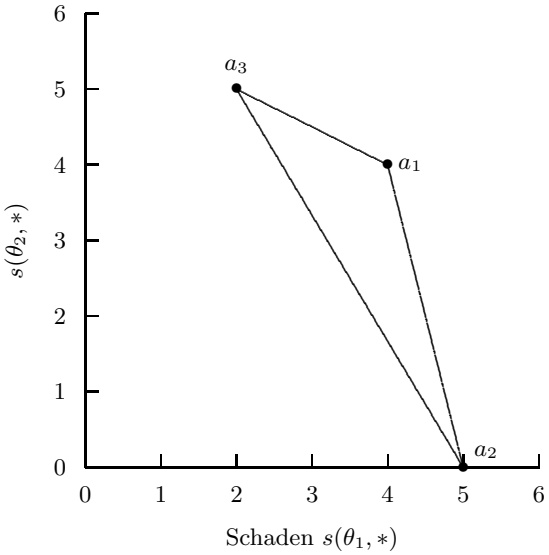

Abb. 4.1. Die Schadensebene aus dem Beispiel in Tabelle 4.1. Die Achsen sind die Schadenswerte bei den beiden Zuständen θ_1 und θ_2. Für jede Entscheidungsalternative ist ein Punkt dargestellt. Alle Punkte auf den Verbindungslinien und innerhalb des dargestellten Dreiecks können durch randomisierte Entscheidungen erreicht werden.

Randomisierte Entscheidungen sind unter anderem dann empfehlenswert, wenn die Entscheidungen im Rahmen eines „Spiels“ zu treffen sind, bei dem die Zustände den Entscheidungen eines intelligenten Spielgegners (der „Natur“) entsprechen. In diesem Fall kann eine reine Entscheidungsregel dem Gegner die Möglichkeit geben, sich auf diese Strategie einzustellen und für sich auszunutzen. Eine randomisierte Entscheidungsregel dagegen

erlaubt es dem Gegner nicht, die eigenen Aktionen mit Sicherheit vorherzusagen.

4.2.3 Das Minimax-Prinzip

Nach dem Minimax-Prinzip wird für jede Alternative der maximale Schaden über die Zustände betrachtet, und es wird diejenige Alternative gewählt, für die dieser Wert am kleinsten ist. Das Optimalitätskriterium M' ist also die Funktion

$$M'(a) = \max_{\theta} s(\theta, a).$$

Tabelle 4.2 zeigt in der letzten Spalte die Werte von M' für die jeweiligen Alternativen. Die Minimax-Lösung ist die Alternative, die M' minimiert. Da bei Alternative a_1 der kleinste Wert von M' auftritt, ist nach dem Minimax-Prinzip Alternative a_1 zu wählen.

Tabelle 4.2. Schadenstabelle des Beispiels aus Tabelle 4.1 mit den Werten des Minimax-Optimalitätskriteriums. Entscheidet man nach diesem Kriterium, so ist die Alternative a_1 („zu Hause bleiben“) zu wählen, da sie den kleinsten aller für die zu erwartenden Zustände maximalen Schadenswerte aufweist.

	θ_1	θ_2	$M'(a)$
a_1	4	4	4
a_2	5	0	5
a_3	2	5	5

Wir betrachten die graphische Darstellung der Schadensebene in Abbildung 4.2 mit den Achsen $s(\theta_1, *)$ und $s(\theta_2, *)$. Jeder Punkt (x, y) repräsentiert einen Schadenswert. Ist $s(\theta_1, a) < s(\theta_2, a)$, dann ist für die Entscheidung $a, s(\theta_2, a)$ der maximale Schaden. Von allen Punkten über der Linie $s(\theta_1, a) = s(\theta_2, a)$ ist der Punkt mit dem kleinsten Ordinatenwert zu suchen, um das minimale Schadensmaximum anzugeben. Ist $s(\theta_1, a) > s(\theta_2, a)$, liegt also die Entscheidung unter der Winkelhalbierenden $s(\theta_1, a) = s(\theta_2, a)$, dann ist $s(\theta_1, a)$ das Maximum, und um das Minimum all dieser Punkte zu finden, ist derjenige mit dem kleinsten Abszissenwert zu suchen.

Berücksichtigt man nun sowohl die Punkte über als auch die Punkte unter der Winkelhalbierenden, dann ist die Minimax-Entscheidung der Punkt, der, wenn er unter der Winkelhalbierenden liegt, den kleinsten Abszissenwert, wenn er über der Winkelhalbierenden liegt, den kleinsten Ordinatenwert hat. Bewegt man also vom Ursprung aus einen rechten Winkel in Richtung der Winkelhalbierenden, dann ist die Minimax-Entscheidung genau die, auf die dieser Winkel zuerst trifft.

Für die Minimax-Lösung aus der Menge der randomisierten Entscheidungen gilt

$$5\gamma + 2(1-\gamma) = 5(1-\gamma)$$

mit der Lösung $\gamma = \frac{3}{8}$. Damit ist $s(\theta_1, a) = s(\theta_2, a) = \frac{25}{8}$. Minimax-Lösungen bei gemischten Entscheidungen führen immer dazu, daß $s(\theta_1, p_M) = s(\theta_2, p_M)$, so daß für beide Zustände die gleiche Schadenserwartung vorliegt.

Die Minimax-Schadenserwartung hat im obigen Beispiel bei der randomisierten Entscheidung mit den Wahrscheinlichkeiten $\left(0, \frac{3}{8}, \frac{5}{8}\right)$ den Wert $\frac{25}{8}$. Der Minimax-Schaden bei den reinen Entscheidungen war 4. Damit ist demonstriert, daß durch die Mischung von Entscheidungen nach dem Minimax-Kriterium eine geringere maximale Schadenserwartung erzielt werden kann als durch reine Entscheidungen.

4.2.4 Das Bayes-Prinzip

Wenn bekannt ist, daß die Zustände θ aus Θ nicht mit der gleichen Häufigkeit auftreten, so wird man dies bei der Entscheidung auch dann berücksichtigen, wenn keine individuellen Daten vorliegen. Wir gehen von einem

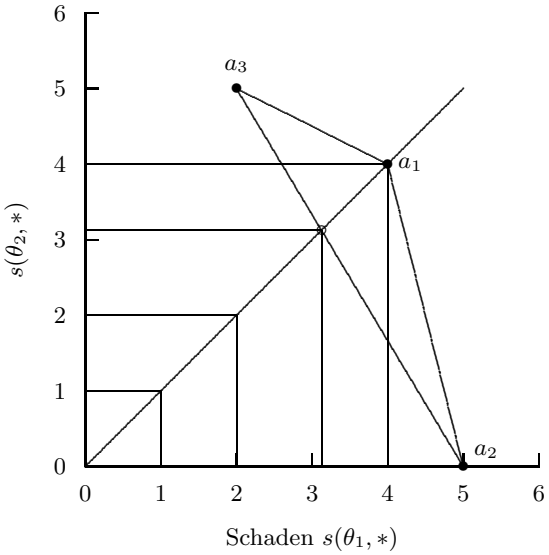

Abb. 4.2. Die Schadensebene aus dem Beispiel in Tabelle 4.1 und die Minimax-Lösungen für reine und gemischte Entscheidungen. Die Minimax-Entscheidung ist diejenige Entscheidung, auf die ein Rechteck, das vom Ursprung entlang der Winkelhalbierenden bewegt wird, zuerst trifft. Die Minimax-Lösung unter den reinen Entscheidungen ist a_1, unter den gemischten Entscheidungen der Vektor $p = (0, \frac{3}{8}, \frac{5}{8})$. Das Verfahren zeigt, daß Minimax-Lösungen bei gemischten Entscheidungen immer zur Folge haben, daß $s(\theta_1, p_M) = s(\theta_2, p_M)$, für beide Zustände also die gleiche Schadenserwartung vorliegt.

diskreten Zustandsraum Θ aus und bezeichnen mit $\pi(\theta)$ die Wahrscheinlichkeit für das Auftreten eines bestimmten Zustandes θ aus der Menge Θ aller möglichen Zustände. Als „Bayesschen“ Schaden einer Alternative a bezeichnet man das Optimalitätskriterium, das den Erwartungswert des Schadens dieser Alternative über die verschiedenen Zustände angibt:

$$\begin{aligned} B'(a) &= \mathscr{E}[s(\theta, a)] \\ &= \sum_{\theta} s(\theta, a)\pi(\theta). \end{aligned} \tag{4.3}$$

Als „Bayessche“ Handlung wird diejenige Alternative a_B bezeichnet, die den minimalen Bayesschen Schaden ergibt:

$$B'(a_B) = \min_a B'(a).$$

Die Wahrscheinlichkeiten $\pi(\theta)$ werden als „a priori“ Wahrscheinlichkeiten bezeichnet. Diese Bezeichnug wird erst klarer werden, wenn Daten in die Betrachtung mit einbezogen werden. Diese führen nämlich dann zu einer „a posteriori“ Verteilung, die bei validen Daten mehr Information enthält als die a priori Verteilung der Zustände. Das Optimalitätskriterium bei einer Bayesschen Vorgehensweise ist also der Erwartungswert des Schadens. Dessen Berechnung setzt voraus, daß ein gewisses Maß an Vorinformation über die Zustände vorhanden ist, nämlich deren Wahrscheinlichkeitsverteilung. Tabelle 4.3 gibt ein Beispiel für eine a priori Verteilung von Zuständen und den Wert des Bayes-Kriteriums.

Für eine bestimmte a priori Verteilung $\pi(\theta)$ wird der Bayessche Schaden einer Alternative a bei zwei Zuständen durch folgende Formel berechnet:

$$B'(a) = \pi(\theta_1)s(\theta_1, a) + \pi(\theta_2)s(\theta_2, a).$$

In der Schadensebene von Abbildung 4.3 stellt diese Gleichung eine Gerade mit der Steigung $-\pi(\theta_1)/\pi(\theta_2)$ dar. Alle Bayes-Lösungen liegen auf solchen Geraden. Die Bayes-Lösung mit der minimalen Schadenserwartung findet man, wenn in der Schadensebene eine Gerade dieser Steigung parallel vom Ursprung aus nach rechts oben verschoben wird. Die erste Entscheidung, auf die man dann trifft, ist die Bayes-Lösung.

Tabelle 4.3. Schadenstabelle des Beispiels mit Werten der a priori Verteilung und des Bayes - Optimalitätskriteriums.

	θ_1	θ_2	$B'(a)$
a_1	4	4	4.0
a_2	5	0	2.0
a_3	2	5	3.8
$\pi(\theta)$	0.4	0.6	

Unterschiedliche a priori Verteilungen über den Zuständen spiegeln sich in der graphischen Lösung durch unterschiedliche Geradensteigungen. Abbildung 4.3 läßt erkennen, daß bei hinreichend steilem Verlauf der Geradenschar der Bayes-Lösungen in unserem Beispiel statt a_2 auch a_3 als Bayes-Entscheidung in Frage käme.

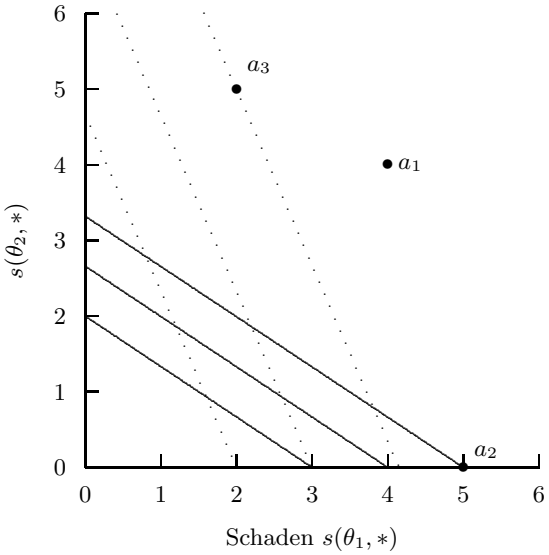

Abb. 4.3. Die Schadensebene aus dem Beispiel in Tabelle 4.3. Die Bayes-Entscheidungen aus dem Beispiel liegen auf Geraden mit der Steigung $-4/6$. Für diese Grundraten ergibt sich damit Alternative a_2 als Bayes-Entscheidung. Wären die Grundraten $\pi(\theta_1) = 0.7$ und $\pi(\theta_2) = 0.3$, ergäbe sich a_3 als Bayes-Entscheidung. Die Abbildung demonstriert auch, daß die zulässigen gemischten Entscheidungen auf der Verbindungslinie von a_2 und a_3 für bestimmte Grundraten Bayes-Entscheidungen darstellen. Dagegen kann Entscheidung a_1 niemals eine Bayes-Entscheidung sein.

Die graphische Darstellung zeigt auch, daß bei einem Bayesschen Vorgehen das Ergebnis nicht davon abhängen kann, ob man die Schäden oder den Regret betrachtet. Wir hatten früher gesehen, daß der Übergang vom Schaden zum Regret nur eine Parallelverschiebung der Punkte in der Schadens- oder Regretebene zur Folge hat. Das Bayessche Vorgehen mit der Verschiebung einer Geraden in der Schadensebene vom Ursprung aus muß daher immer auf die gleiche Alternative stoßen. Nur wenn die Menge der Schadenspunkte in der Schadensebene gedreht würde, könnte sich das Ergebnis des Bayesschen Verfahrens dadurch ändern.

Der erwartete Schaden einer gemischten Entscheidung p über k Alternativen ist nach Gleichung (4.2) gegeben durch

$$s(\theta, p) = \sum_{i=1}^{k} s(\theta, a_i) p_i.$$

Damit ergibt sich der Bayessche Schaden einer solchen Entscheidung über

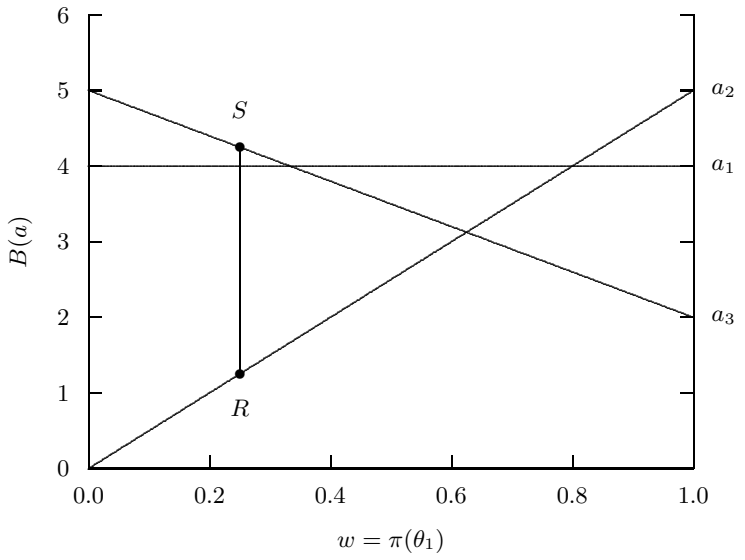

Abb. 4.4. Verlauf des Bayesschen Schadens für die drei Alternativen a_1, a_2 und a_3 in Abhängigkeit von der a priori Wahrscheinlichkeit des Zustands θ_1.

die verschiedenen Zustände durch

$$\begin{aligned} B'(p) &= \mathscr{E}[s(\theta, p)] \\ &= \sum_{\theta} \pi(\theta) \Big(\sum_{i=1}^{k} s(\theta, a_i) p_i \Big). \end{aligned}$$

Die Bestimmung der gemischten Entscheidung mit dem minimalen Bayesschen Schaden ist damit ein Minimierungsproblem mit $(k-1)$ unbekannten Parametern, da der stochastische Vektor $p = (p_1, \ldots, p_k)$ zu bestimmen ist, der wegen $p_1 + \ldots + p_k = 1$ genau $(k-1)$ freie Parameter enthält. In Tabelle 4.4 sind die Schäden aus Tabelle 4.1 enthalten. Statt fester Werte für die a priori Wahrscheinlichkeiten wurde die Variable $w = \pi(\theta_1)$ eingeführt. Abbildung 4.4 zeigt die Schäden der drei Alternativen als Funktion der a priori Wahrscheinlichkeit. Die Abbildung zeigt für jede Alternative eine Gerade, die ihren Schadensverlauf angibt. Die Alternative a_2 ergibt den minimalen Bayesschen Schaden, solange $w < \frac{5}{8}$. Ist $w > \frac{5}{8}$, so wird a_3 zur Bayesschen Entscheidung. Für $w = \frac{5}{8}$ ergeben a_2 und a_3 den gleichen Bayesschen Schaden, jede von beiden kann gewählt werden.

Tabelle 4.4. Schadenstabelle des Beispiels mit variablen Werten der a priori Verteilung.

	θ_1	θ_2	$B'(a)$
a_1	4	4	4.0
a_2	5	0	$5.0w$
a_3	2	5	$5 - 3w$
$\pi(\theta)$	w	$(1-w)$	

Welchen Einfluß hätte nun die Einführung von Mischungen? Jede Mischung zwischen den Alternativen ergäbe eine konvexe Kombination des Schadens der reinen Alternativen. Bei gegebenem Wert für w liegen also die erwarteten Schäden gemischter Entscheidungen auf der senkrechten Verbindungsstrecke zwischen den Schadenswerten der reinen Entscheidungen. Als Beispiel einer solchen Strecke ist die Strecke $\overline{SR}$ eingezeichnet.

Für alle Werte von w liegen die erwarteten Schäden in dem Bereich, der von den minimalen und maximalen Linienabschnitten begrenzt wird. Da für uns nur die Minima von Interesse sind, ergibt sich daraus, daß die

Einführung von Mischungen keinen Gewinn bringen kann, da keine neuen Punkte *unter* den minimalen Linienabschnitten hinzukommen können. Gemischte Entscheidungen können beim Bayesschen Vorgehen keine zusätzliche Schadensreduktion herbeiführen.

4.2.5 Zulässigkeit

Am Beispiel aus Tabelle 4.1 kann man erkennen, daß die zu erwartenden Schäden vom auftretenden Zustand abhängen werden. Falls nun der Schaden bei einer Alternative a_i unter allen Zuständen kleiner ist, als bei einer anderen Alternative a_j, so ist a_i gegenüber a_j stets vorzuziehen.

Definition 4.2. Eine Alternative a_i *dominiert* eine Alternative a_j, wenn für alle θ aus Θ gilt $s(\theta, a_i) \leq s(\theta, a_j)$. Falls darüber hinaus die Ungleichung für einige θ streng gilt, dann *dominiert* die Alternative a_i die Alternative a_j *streng*.

Definition 4.3. Eine Alternative heißt *zulässig*, wenn sie nicht von einer anderen Alternative streng dominiert wird.

In Abbildung 4.1 kann man erkennen, daß die Zulässigkeit einer Entscheidung davon abhängt, welche anderen Entscheidungen noch verfügbar sind. In der Menge der reinen Entscheidungen a_1, a_2, a_3 sind alle Entscheidungen zulässig. In der Menge der gemischten Entscheidungen sind nur Mischungen von a_2 und a_3 zulässig. Die Alternative a_1 wird in dieser Menge von einem Teil der Verbindungslinie zwischen den Punkten a_2 und a_3 streng dominiert.

Die Abbildung 4.2 zeigt eine konvexe Menge von Schadenspunkten in einer Schadensebene. Zulässige Entscheidungen befinden sich, wie oben festgestellt wurde, auf der linken unteren Begrenzungslinie dieser konvexen Menge. In Abbildung 4.3 sieht man, daß Bayessche Schadenslinien immer Geraden negativer Steigung sind und daher die konvexe Teilmenge der Schäden immer am linken unteren Rand treffen werden.

Man kann darüber hinaus sehen, daß auch der Punkt, der in Abbildung 4.2 als Minimax-Lösung auftritt, als Bayes-Lösung auftreten würde, wenn eine geeignete Geradensteigung vorläge. Tatsächlich gelten folgende Aussagen:

- Bayes-Lösungen sind in der Regel zulässig.
- Eine Minimax-Entscheidung ist immer eine Bayes-Entscheidung mit bestimmten a priori Wahrscheinlichkeiten.
- Zulässige Entscheidungen können als Bayes-Entscheidungen bei bestimmten a priori Wahrscheinlichkeiten betrachtet werden.

Vor allem der dritte Punkt ist ein Hinweis darauf, daß die Bayes-Theorie die allgemeinste Theorie statistischer Entscheidungen ist. Aus diesem Grund kann auf die Betrachtung vieler spezieller Optimalitätskriterien verzichtet werden, da sich diese auch im Rahmen der Bayes-Theorie abhandeln lassen (Berger, 1980). Wir gehen hier neben dem Bayes-Kriterium nur auf das Minimax-Kriterium näher ein, da es für psychologische Anwendungen besonders gut geeignet ist, den Standpunkt einer einzelnen betroffenen Person darzustellen. Aus der Sicht der Bayes-Theorie stellt das Minimax-Prinzip eine Vorgehensweise dar, die dann angemessen ist, wenn die Zustände von einem „Gegenspieler“ bestimmt werden, der versucht, den Schaden der betroffenen Person zu maximieren.

Aus Abbildung 4.4 können die Bayesschen Schäden für verschiedene a priori Verteilungen abgelesen werden. Die a priori Verteilung $\pi(\theta_1) = \frac{5}{8}$, $\pi(\theta_2) = \frac{3}{8}$ ergibt das maximale Schadensminimum. Ein „Gegenspieler“, der den Schaden der betroffenen Person maximieren will, würde daher diese a priori Verteilung wählen, da sie die ungünstigste a priori Verteilung darstellt. Dies korrespondiert mit der Minimax-Lösung in den gemischten Entscheidungen, die aus einer Mischung von a_2 und a_3 im gleichen Zahlenverhältnis besteht.

4.3 Entscheidungen aufgrund von Daten

4.3.1 Entscheidungsregeln

Daten werden in das Entscheidungsverfahren durch die Entscheidungsregel eingebracht. Diese Abbildung des Stichprobenraumes $\mathfrak{X}$ auf die Menge der Alternativen A ordnet jedem möglichen Datum oder jedem Datenvektor, falls es sich um mehrdimensionale Daten handelt, eine Alternative a zu. Wird also das Datum x beobachtet, so muß eine Entscheidungsregel δ dem Datum x eine Alternative a zuordnen: $\delta(x) = a$. Da bei festgelegter Entscheidungsregel und gegebenem Datenpunkt die zu wählende Alternative festliegt, besteht das Problem des Entscheidungsverfahrens darin, eine optimale Entscheidungsregel zu finden. Das Optimalitätskriterium kann sich damit nicht mehr auf die Alternativen, sondern muß sich auf die Entscheidungsregeln beziehen. Es gilt, jeder möglichen Entscheidungsregel eine Zahl so zuzuordnen, daß mit dieser Bewertung eine optimale Entscheidungsregel bestimmt werden kann.

Die Anzahl der *möglichen* Entscheidungsregeln hängt davon ab, wieviele verschiedene Werte von X auftreten können und wieviele Alternativen a möglich sind. Bei k Alternativen und m verschiedenen Ausprägungen von X gibt es k^m verschiedene Entscheidungsfunktionen δ in Δ. Von den k^m theoretisch möglichen Entscheidungsfunktionen sind natürlich manche unsinnig, manche ignorieren die Daten, und manche benutzen sie falsch. Das Ziel des Entscheidungsverfahrens ist, die optimale Entscheidungsfunktion zu finden.

Tabelle 4.5 zeigt alle möglichen Entscheidungsregeln für einen Stichprobenraum $\mathfrak{X} = \{x_1, x_2\}$ mit zwei möglichen Ausprägungen als Daten für das Beispiel aus Tabelle 4.1. Es ergeben sich 9 mögliche Entscheidungsregeln. Die Regeln δ_1, δ_2 und δ_3 ignorieren die Werte der Stichprobe, sie führen unabhängig von den Daten immer zur gleichen Entscheidung. Die Regeln δ_4 und δ_5, δ_6 und δ_7, und die beiden Regeln δ_8 und δ_9 sind jeweils „gegenteilig“, sie werten die Daten gegensinnig, so daß nur eine von beiden sinnvoll sein kann.

Tabelle 4.5. Die Menge der möglichen Entscheidungsregeln für das Beispiel aus Tabelle 4.1. Es gibt neun mögliche Zuordnungen von den drei Alternativen zu den zwei Datenpunkten.

	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9
x_1	a_1	a_2	a_3	a_1	a_2	a_1	a_3	a_2	a_3
x_2	a_1	a_2	a_3	a_2	a_1	a_3	a_1	a_3	a_2

Analog zu gemischten Entscheidungen können auch Regeln randomisiert werden. Einem Datum x wird dann nicht eine bestimmte Regel, sondern eine Wahrscheinlichkeitsverteilung über der Menge Δ aller möglichen Regeln zugeordnet. Alternativ dazu kann aber auch jede Regel selbst in der Zuordnung einer Wahrscheinlichkeitsverteilung über den Alternativen zu den Daten bestehen. Es läßt sich zeigen, daß im Falle einer endlichen Anzahl von Alternativen beide Methoden äquivalent sind.

Wir wollen annehmen, die Daten im erwähnten Beispiel kämen durch die Beobachtung der Bewölkung am Himmel zustande:

$$X = \begin{cases} x_1, & \text{falls der Himmel bedeckt ist,} \\ x_2, & \text{falls der Himmel klar ist.} \end{cases}$$

Die bedingten Verteilungen von X unter den beiden Zuständen seien durch die Werte in Tabelle 4.6 gegeben. Für eine bestimmte Entscheidungsregel δ ist der eintretende Schaden eine Zufallsvariable, die sowohl vom Zustand θ als auch vom Datum X abhängt.

Tabelle 4.6. Die angenommenen bedingten Verteilungen der Zufallsvariablen X für die beiden Zustände des Beispiels aus Tabelle 4.1.

	$f(X \mid \theta_1)$	$f(X \mid \theta_2)$
x_1	0.8	0.1
x_2	0.2	0.9

Den Erwartungswert des Schadens bei einem bestimmten Zustand θ unter der Entscheidungsregel δ bezeichnet man als *Risikofunktion*:

$$R(\theta, \delta) = \mathscr{E}[s(\theta, \delta(x))].$$

Bei der Definition der Risikofunktion ist zu beachten, daß die Bildung des Erwartungswertes über die Ausprägungen von X zu erfolgen hat, während bei der Schadenserwartung in Gleichung (4.3) der Erwartungswert über die Verteilung der Zustände gebildet wird. Der Erwartungswert ist hier bezüglich der durch ein θ festgelegten, bedingten Verteilung von X zu bestimmen:

$$R(\theta, \delta) = \sum_x s[\theta, \delta(x)]\, f(x \mid \theta), \tag{4.4}$$

wobei $f(x \mid \theta)$ die bedingte Wahrscheinlichkeit von x bei gegebenem Zustand θ ist (im Fall diskreter Ausprägungen von X).

Tabelle 4.7. Die mit den neun Entscheidungsregeln aus Tabelle 4.5 verbundenen Werte der Risikofunktion.

	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9
θ_1	4.0	5.0	2.0	4.2	4.8	3.6	2.4	4.4	2.6
θ_2	4.0	0.0	5.0	0.4	3.6	4.9	4.1	4.5	0.5

Tabelle 4.7 zeigt die Werte der Risikofunktion für die beiden Zustände θ_1 und θ_2 des Beispiels. Das Problem, beim Vorliegen von Daten die optimale Entscheidungsregel zu finden, ist völlig analog zum Problem der Auswahl einer Entscheidung im Fall ohne Daten. Statt der Schadensfunktion dient im Fall mit Daten die Risikofunktion als Bewertungsgrundlage der Entscheidungsregeln. Abbildung 4.5 zeigt analog zu Abbildung 4.1 die Risikoebene des Beispiels mit den Werten von Tabelle 4.7 als Punktkoordinaten. An der Abbildung kann man erkennen, daß die Regeln δ_1, δ_2 und δ_3 den reinen Entscheidungen a_1, a_2 und a_3 entsprechen. Sie ergeben die gleichen Schadenswerte. Mischungen der Regeln δ_1, δ_2 und δ_3 entsprechen genau den Mischungen der reinen Alternativen a_1, a_2 und a_3. Die Abbildung zeigt auch, daß einige andere Regeln die Mischungen aus δ_1, δ_2 und δ_3 klar dominieren, da sie bei jedem Zustand ein geringeres Risiko ergeben. Damit ist demonstriert, daß ein geeignetes Ausnutzen von validen Daten die Größe des Risikos reduziert. Andererseits zeigen die Regeln δ_5, δ_6 und δ_8, daß eine falsche Verwendung der Daten zu schlechteren Entscheidungen führten, als sie ohne Daten getroffen würden.

4.3.2 Verminderung des Risikos durch Validität

Eine Reduktion des Risikos kann allerdings nur von validen Daten erwartet werden. Sind die Daten nämlich unabhängig von den Zuständen ist also $f(X \mid \theta) = g(X)$ für alle θ, dann ergibt sich für das Risiko einer Regel δ

$$R(\theta, \delta) = \sum_x s[\theta, \delta(x)]\, g(x).$$

Dies stellt aber nur eine konvexe Kombination aller Schäden $s(\theta, a)$ der Alternativen $\delta(x) = a$ dar. Diese Risiken entsprechen den gemischten Entscheidungen, die in Abbildung 4.1 als Dreieck mit den Eckpunkten $s(\theta, a_1)$,

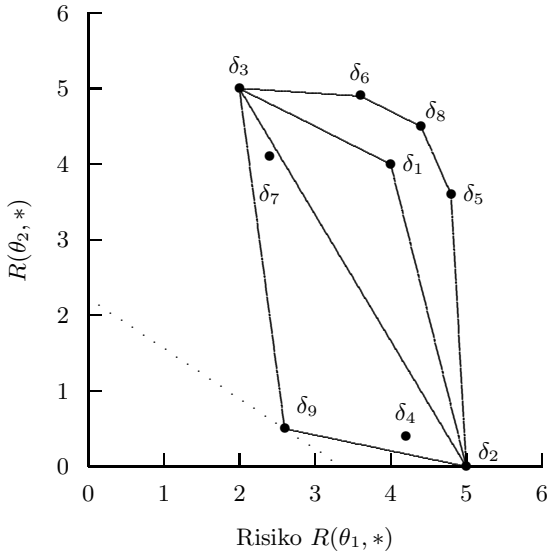

Abb. 4.5. Die Risikoebene aus dem Beispiel in Tabelle 4.7. Die Achsen sind die Risiken bei den beiden Zuständen θ_1 und θ_2. Für jede Entscheidungsregel δ_i ist ein Punkt dargestellt. Alle Punkte innerhalb des Dreiecks können durch gemischte Entscheidungen ohne Berücksichtigung von Daten erreicht werden. Die äußere Hülle der Punktmenge beschreibt die Menge der gemischten Entscheidungs*regeln*. Die zulässigen Regeln liegen auch hier auf der linken unteren Begrenzungslinie. Diese Punkte erreichen für alle Zustände geringere Schadenswerte als die Entscheidungen ohne Daten. Analog zu Abbildung 4.3 ergibt sich hier die Entscheidungsregel δ_9 als Bayes-Lösung.

$s(\theta, a_2)$, $s(\theta, a_3)$ eingezeichnet sind. Eine Reduktion der Schäden wird durch die Gewichtungen mit $g(X)$ nicht möglich. Nur bedingte Wahrscheinlichkeiten $f(X \mid \theta)$, die tatsächlich von den Zuständen θ abhängen, können die Linie der zulässigen Entscheidungsregeln zum Ursprung der Schadensebene hin ausdehnen. Dies wird dadurch möglich, daß die konvexen Kombinationen der Schäden bei einer Gewichtung durch die bedingten Wahrscheinlichkeiten für jeden Zustand separat betrachtet werden können. Damit werden die konvexen Mischungen *achsenweise* gebildet und können bei günstigen Werten der bedingten Wahrscheinlichkeiten zu Risikopunkten führen, die unterhalb bzw. links von den Mischungen der Schadenspunkte liegen.

4.3.3 Zulässige und optimale Entscheidungsregeln

Die Konzepte der Dominanz und Zulässigkeit nach Definition 4.2 und 4.3 können von Alternativen auf Entscheidungsregeln übertragen werden.

Definition 4.4. Eine Entscheidungsregel δ_i *dominiert* eine Entscheidungsregel δ_j, wenn für alle θ aus Θ gilt $R(\theta, \delta_i) \leq R(\theta, \delta_j)$. Falls darüber hinaus für mindestens ein θ die Ungleichung streng gilt, dann *dominiert* δ_i die Regel δ_j *streng*. Eine Entscheidungsregel ist *zulässig*, falls sie von keiner anderen streng dominiert wird.

Wie bereits früher angedeutet, ist das mathematische Problem der Bestimmung einer optimalen Entscheidungsregel äquivalent dem Problem der Bestimmung einer optimalen Alternative im Falle des Vorgehens ohne Daten. Statt der Schadensfunktion wird hier die Risikofunktion als Bewertung benutzt. Wir können also auch hier nach dem Minimax- oder dem Bayes-Prinzip vorgehen. Das Minimax-Kriterium ist

$$M(\delta) = \max_{\theta} R(\theta, \delta).$$

Die Minimax-Regel δ_M minimiert $M(\delta)$. Als *Wert* der Daten kann die Differenz zwischen $M'(a_M)$, dem Minimum der maximalen Schäden im Verfahren ohne Daten, und $M(\delta_M)$, dem Minimum der maximalen Risiken beim Vorliegen von Daten, betrachtet werden.

Das Bayes-Kriterium ist das erwartete Risiko

$$\begin{aligned} B(\delta) &= \mathscr{E}[R(\theta,\delta)] \\ &= \sum_{\theta} R(\theta,\delta)\,\pi(\theta) \end{aligned} \tag{4.5}$$

mit der a priori Wahrscheinlichkeit $\pi(\theta)$ der Zustände. Als Bayes-Regel wird diejenige Entscheidungsregel δ_B bezeichnet, die $B(\delta)$ minimiert. Wie im Fall ohne Daten kann auch hier vom Regret oder vom Schaden ausgegangen werden, was wiederum nur beim Minimax-Verfahren eine Rolle spielt. Auch hier kann das minimale Bayessche Risiko $B(\delta_B)$ mit dem minimalen Bayesschen Schaden $B'(a_B)$ verglichen werden, um die Schadensverminderung durch die Datenerhebung zu beurteilen.

4.3.4 Der Sonderfall von zwei Zuständen und zwei Alternativen

Wir betrachten im folgenden ein Entscheidungsproblem mit zwei Zuständen ($\Theta = \{\theta_1, \theta_2\}$) und zwei Alternativen ($A = \{a_1, a_2\}$). Die Entscheidungsregel δ: $\mathfrak{X} \to A$ nimmt dann nur die beiden Werte a_1 oder a_2 an. Für das Risiko bei den beiden Zuständen erhält man wegen Gleichung (4.4)

$$R(\theta,\delta) = \sum_{x\in\mathfrak{X}} s[\theta,\delta(x)]\, f(x\,|\,\theta).$$

Wir trennen nun die Summation auf in die beiden Teilmengen $\mathfrak{X}^{(1)}$ und $\mathfrak{X}^{(2)}$ von $\mathfrak{X}$, die durch die Werte a_1 und a_2 der Entscheidungsregel δ bestimmt sind: $\mathfrak{X}^{(j)} = \{x\,|\,\delta(x) = a_j\}$ für $j = 1, 2$. Dann erhalten wir

$$R(\theta,\delta) = \sum_{x\in\mathfrak{X}^{(1)}} s(\theta,a_1) f(x\,|\,\theta) + \sum_{x\in\mathfrak{X}^{(2)}} s(\theta,a_2) f(x\,|\,\theta),$$

da für alle $x \in \mathfrak{X}^{(j)}$ gilt $\delta(x) = a_j$. Die Schäden unter den Summenzeichen können nach vorne gezogen werden:

$$R(\theta,\delta) = s(\theta,a_1) \sum_{x\in\mathfrak{X}^{(1)}} f(x\,|\,\theta) \;+\; s(\theta,a_2) \sum_{x\in\mathfrak{X}^{(2)}} f(x\,|\,\theta).$$

Da $\mathfrak{X} = \mathfrak{X}^{(1)} \cup \mathfrak{X}^{(2)}$ müssen sich alle Wahrscheinlichkeiten $f(X\,|\,\theta)$ in den beiden Summen zu 1 addieren und wir können eine der beiden Summen durch das Komplement der anderen zu 1 ersetzen. Für den geplanten Vergleich zweier Entscheidungsregeln ist es vorteilhaft, die Summen für θ_1 und θ_2 unterschiedlich zu ersetzen. Mit

$$\sum_{x\in\mathfrak{X}^{(1)}} f(x\,|\,\theta) = 1 - \sum_{x\in\mathfrak{X}^{(2)}} f(x\,|\,\theta)$$

erhält man für $R(\theta_1,\delta)$

$$R(\theta_1,\delta) = s(\theta_1,a_1) + [s(\theta_1,a_2) - s(\theta_1,a_1)] \sum_{x\in\mathfrak{X}^{(2)}} f(x\,|\,\theta_1). \tag{4.6}$$

Mit

$$\sum_{x\in\mathfrak{X}^{(2)}} f(x\,|\,\theta) = 1 - \sum_{x\in\mathfrak{X}^{(1)}} f(x\,|\,\theta)$$

erhält man für $R(\theta_2,\delta)$

$$R(\theta_2,\delta) = s(\theta_2,a_2) + [s(\theta_2,a_1) - s(\theta_2,a_2)] \sum_{x\in\mathfrak{X}^{(1)}} f(x\,|\,\theta_2). \tag{4.7}$$

Wir betrachten nun die Zulässigkeit zweier Entscheidungsregeln δ_1 und δ_2. Die Regel δ_1 ist nicht zulässig, wenn für die Risiken bei den beiden

Zuständen folgende zwei Ungleichungen gleichzeitig gelten und mindestens eine davon streng gilt:

$$\begin{aligned} R(\theta_1, \delta_1) &\geq R(\theta_1, \delta_2) \\ R(\theta_2, \delta_1) &\geq R(\theta_2, \delta_2). \end{aligned}$$

Die Bedingung

$$R(\theta_1, \delta_1) \geq R(\theta_1, \delta_2)$$

ist gleichbedeutend mit

$$\begin{aligned} & s(\theta_1, a_1) + [s(\theta_1, a_2) - s(\theta_1, a_1)] \sum_{x \in \mathfrak{X}_1^{(2)}} f(x \,|\, \theta_1) \\ \geq \;& s(\theta_1, a_1) + [s(\theta_1, a_2) - s(\theta_1, a_1)] \sum_{x \in \mathfrak{X}_2^{(2)}} f(x \,|\, \theta_1) \end{aligned}$$

mit $\mathfrak{X}_i^{(j)} = \{x \,|\, \delta_i(x) = a_j\}$.

Der Term $s(\theta_1, a_1)$ ist auf beiden Seiten gleich, beeinflußt also die Ungleichung nicht. Das gleiche gilt für $[s(\theta_1, a_2) - s(\theta_1, a_1)]$, so lange dessen Wert positiv ist. Handelt es sich um Schäden, so ist anzunehmen, daß der Schaden $s(\theta_1, a_1)$ kleiner ist als $s(\theta_1, a_2)$, wenn a_1 die für θ_1 korrekte Entscheidung ist. Damit ist anzunehmen, daß $[s(\theta_1, a_2) - s(\theta_1, a_1)]$ positiv ist und es ergibt sich, daß

$$R(\theta_1, \delta_1) \geq R(\theta_1, \delta_2)$$

genau dann gilt, wenn

$$\sum_{x \in \mathfrak{X}_1^{(2)}} f(x \,|\, \theta_1) \geq \sum_{x \in \mathfrak{X}_2^{(2)}} f(x \,|\, \theta_1).$$

Analog zur obigen Ableitung für θ_1 gilt für θ_2

$$R(\theta_2, \delta_1) \geq R(\theta_2, \delta_2)$$

genau dann, wenn

$$\begin{aligned} & s(\theta_2, a_2) + [s(\theta_2, a_1) - s(\theta_2, a_2)] \sum_{x \in \mathfrak{X}_1^{(1)}} f(x \,|\, \theta_2) \\ \geq \;& s(\theta_2, a_2) + [s(\theta_2, a_1) - s(\theta_2, a_2)] \sum_{x \in \mathfrak{X}_2^{(1)}} f(x \,|\, \theta_2), \end{aligned}$$

was wie bei θ_1 genau dann der Fall ist, wenn

$$\sum_{x \in \mathfrak{X}_1^{(1)}} f(x \,|\, \theta_2) \geq \sum_{x \in \mathfrak{X}_2^{(1)}} f(x \,|\, \theta_2).$$

Die Zulässigkeit einer Entscheidungsregel kann also unabhängig von den Schadenswerten festgestellt werden. Dies machen auch die Gleichungen (4.6) und (4.7) deutlich. Diese zeigen nämlich, daß die Risiken $R(\theta, \delta)$ durch eine lineare Transformation aus den Wahrscheinlichkeiten einer Fehlklassifikation hervorgehen. Der Ausdruck

$$\sum_{x \in \mathfrak{X}^{(2)}} f(x \,|\, \theta_1)$$

ist die Wahrscheinlichkeit dafür, die Alternative a_2 zu wählen, wenn der Zustand θ_1 vorliegt, gibt also die Wahrscheinlichkeit einer Fehlklassifikation an. Das Risiko einer Entscheidungsregel δ beim Vorliegen des Zustandes θ_1 ist nach Gleichung (4.6)

$$R(\theta_1, \delta) = u + v\Big(\sum_{x \in \mathfrak{X}^{(2)}} f(x \,|\, \theta_1) \Big)$$

mit $u = s(\theta_1, a_1)$ und $v = [s(\theta_1, a_2) - s(\theta_1, a_1)]$. Unterschiedliche Schadenswerte haben für das Risiko nur den Effekt, daß sie es linear transformieren, wobei die Faktoren v immer positiv sind, da $s(\theta_1, a_2)$ immer größer sein wird als $s(\theta_1, a_1)$. Eine solche lineare Transformation bildet die Menge der zulässigen Entscheidungsregeln immer in sich selbst ab: Entscheidungsregeln, die bei einer Schadensfunktion zulässig sind, sind damit bei allen Schadensfunktionen zulässig. Positive Lineartransformationen können eben die Ordnung einzelner Punkte entlang einer Achse nicht ändern.

Die Gleichungen (4.6) und (4.7) zeigen auch, daß bei Bayes-Entscheidungen die Schadenswerte $s(\theta_1, a_1)$ und $s(\theta_2, a_2)$ der „richtigen“ Entscheidungen keine Rolle spielen, es kommt nur auf die Differenz zwischen den Schäden der falschen und richtigen Entscheidungen bei den beiden Zuständen an. Dies ist Ausdruck der Tatsache, daß im Falle des Bayes-Kriteriums Entscheidungen auf der Basis von Regret und auf der Basis des Risikos immer zu den gleichen Ergebnissen führen. In manchen Darstellungen der Bayes-Theorie wird daher der Schaden einer richtigen Entscheidung gleich von Anfang an als 0 angenommen (Raiffa, 1961).

4.3.5 Konstruktion einer Bayes-Lösung

Die Bayes-Regel minimiert das erwartete Risiko

$$B(\delta) \quad = \quad \sum_{\theta} R(\theta, \delta)\pi(\theta). \tag{4.8}$$

Die Suche dieser Lösung vereinfacht sich, wenn man auf die a posteriori Verteilung übergeht. Wir setzen dafür statt $R(\theta, \delta)$ die Definition des Risikos

$$R(\theta, \delta) \quad = \quad \sum_{x} s[\theta, \delta(x)] f(x \mid \theta)$$

in Gleichung (4.8) ein und erhalten damit

$$B(\delta) \quad = \quad \sum_{\theta} \sum_{x} s[\theta, \delta(x)] f(x \mid \theta)\pi(\theta) \tag{4.9}$$

Aufgrund der Definition der bedingten Wahrscheinlichkeit gilt

$$f(X \mid \theta)\pi(\theta) = P(\theta, X) = h(\theta \mid X)g(X) \tag{4.10}$$

mit den Randwahrscheinlichkeiten

$$\begin{aligned} g(X) \quad &= \quad \sum_{\theta} P(\theta, X) \\ &= \quad \sum_{\theta} f(X \mid \theta)\pi(\theta). \end{aligned}$$

$h(\theta \mid X)$ ist die a posteriori Wahrscheinlichkeit dafür, daß ein bestimmter Zustand θ auftritt, wenn die Realisation der Zufallsvariablen X bereits bekannt ist. Sie kann nach dem Theorem von Bayes aus den a priori Wahrscheinlichkeiten und den bedingten Verteilungen von X berechnet werden:

$$h(\theta \mid X) = \frac{f(X \mid \theta)\pi(\theta)}{\sum_{\theta} f(X \mid \theta)\pi(\theta)}. \tag{4.11}$$

Gleichung (4.10) in (4.9) ergibt

$$B(\delta) = \sum_{x} \sum_{\theta} s[\theta, \delta(x)]\, h(\theta \mid x)g(x),$$

da die Summationen ausgetauscht werden können. Das innere Produkt ist der unter der a posteriori Verteilung $h(\theta \mid X)$ erwartete Schaden bei Entscheidungsregel δ und den Daten x:

$$S_h(\delta, x) = \sum_{\theta} s[\theta, \delta(x)]\, h(\theta \mid x). \tag{4.12}$$

Der Bayessche Schaden $B(\delta)$ ist eine gewichtete Summe von solchen erwarteten Schadenswerten, wobei alle Gewichtungen $g(x)$ positiv und unabhängig von δ sind. Diese Summe wird minimal sein, wenn $S_h(\delta, x)$ so klein wie möglich gemacht wird. Wählt man eine Entscheidungsregel δ, die $S_h(\delta, x)$ minimiert, so erhält man damit auch eine Entscheidungsregel, die das Bayessche Risiko minimiert, also eine Bayes-Lösung.

Bei praktischen Anwendungen kann Gleichung (4.12) benutzt werden, um die Bayes-Regel zu finden. Es genügt dafür, eine Entscheidungsregel zu konstruieren, die zu jedem Datum x diejenige Alternative $\delta(x)$ auswählt, die den minimalen Wert für $S_h(\delta, x)$ ergibt. Mathematisch bedeutet das, daß die Funktion $S_h(\delta, x)$ minimiert wird, indem für jedes $x \in \mathfrak{X}$ ein Funktionswert $\delta(x)$ bestimmt wird. Geht man dagegen von Gleichung (4.8) aus, um die Bayes-Regel zu finden, dann muß diese Funktion in der Menge Δ aller möglichen Entscheidungsregeln minimiert werden, ein Problem, das sehr viel schwieriger zu lösen ist.

In dem in Abschnitt 4.3.4 behandelten Fall von zwei Zuständen und zwei Alternativen kommen für jedes x nur die beiden Werte a_1 und a_2 als $\delta(x)$ in Frage. Unter Berücksichtigung der Gleichung (4.12) wählt man demnach a_1, wenn die a posteriori Schadenserwartung für a_1 bei gegebenem Datum x kleiner ist als die a posteriori Schadenserwartung für a_2 bei den gleichen Daten:

$$s(\theta_1, a_1)h(\theta_1 \mid x) + s(\theta_2, a_1)h(\theta_2 \mid x) \leq s(\theta_1, a_2)h(\theta_1 \mid x) + s(\theta_2, a_2)h(\theta_2 \mid x).$$

Diese Ungleichung kann umgeformt werden in

$$[s(\theta_1, a_1) - s(\theta_1, a_2)]\, h(\theta_1 \mid x) \leq [s(\theta_2, a_2) - s(\theta_2, a_1)]\, h(\theta_2 \mid x).$$

Multipliziert man nun beide Seiten mit $g(x)$, erhält man

$$[s(\theta_1, a_1) - s(\theta_1, a_2)]\, h(\theta_1 \mid x)g(x) \leq [s(\theta_2, a_2) - s(\theta_2, a_1)]\, h(\theta_2 \mid x)g(x).$$

Nach dem Theorem von Bayes ist $h(\theta \mid X)g(X) = f(X \mid \theta)\pi(\theta)$, so daß wir erhalten

$$[s(\theta_1, a_1) - s(\theta_1, a_2)]\, f(x \mid \theta_1)\, \pi(\theta_1) \leq [s(\theta_2, a_2) - s(\theta_2, a_1)]\, f(x \mid \theta_2)\pi(\theta_2).$$

Wir können annehmen, daß $s(\theta_1, a_2) > s(\theta_1, a_1)$, denn a_1 soll die für θ_1 „korrekte“ Entscheidung sein. Dividieren wir daher die Ungleichung durch die Differenz $[s(\theta_1, a_1) - s(\theta_1, a_2)]$, so ist ihre Richtung umzukehren. Gleichzeitig dividieren wir durch $f(x \mid \theta_2)$ und durch $\pi(\theta_1)$. Es ergibt sich

$$\frac{f(x \mid \theta_1)}{f(x \mid \theta_2)} \quad \geq \quad \frac{\pi(\theta_2)}{\pi(\theta_1)} \frac{[s(\theta_2, a_2) - s(\theta_2, a_1)]}{[s(\theta_1, a_1) - s(\theta_1, a_2)]}. \tag{4.13}$$

Die Ungleichung bedeutet, daß beim Vorliegen des Datums x die Alternative a_1 genau dann als Wert von $\delta(x)$ definiert wird, wenn der linke Quotient größer als der rechte ist. Man kann auch hier wiederum erkennen, daß Bayes-Entscheidungen nicht direkt von den Schäden korrekter, sondern nur von den Differenzen der Schäden korrekter und fehlerhafter Entscheidungen abhängen.

4.3.6 Likelihoodquotientenregeln

Gleichung (4.13) zeigt auch die Verbindung des Bayes-Kriteriums mit einer anderen Klasse von Entscheidungsregeln, den *Likelihoodquotientenregeln.* Es kann gezeigt werden, daß zulässige Regeln im hier besprochenen Fall immer als Likelihoodquotientenregeln darstellbar sind. Eine *Likelihoodquotientenregel* zum Index q liegt vor, wenn δ folgendermaßen definiert ist:

$$\delta(x) = \begin{cases} a_1 \text{ falls } \dfrac{f(x \mid \theta_1)}{f(x \mid \theta_2)} \geq q \\ a_2 \text{ sonst.} \end{cases}$$

Jede Likelihoodquotientenregel zu einem Index $0 < q < \infty$ ist zulässig, und jede zulässige Regel kann als Likelihoodquotientenregel aufgefaßt werden.

Gleichung (4.13) zeigt, daß die Bayes-Regel eine Likelihoodquotientenregel zum Index

$$q = \frac{\pi(\theta_2)}{\pi(\theta_1)} \frac{[s(\theta_2, a_2) - s(\theta_2, a_1)]}{[s(\theta_1, a_1) - s(\theta_1, a_2)]}$$

ist.

4.3.7 Das Neyman-Pearson-Kriterium

Ein bisher nicht erwähntes Kriterium zur Konstruktion von Entscheidungsregeln ist aus der Statistik bekannt, das *Neyman-Pearson-Kriterium*. Es orientiert sich nicht an den Schäden, sondern nur an den Wahrscheinlichkeiten für Fehlklassifikationen. Effektiv heißt das, daß eine Schadensfunktion mit den Werten $s(\theta_1, a_1) = s(\theta_2, a_2) = 0$ und $s(\theta_1, a_2) = s(\theta_2, a_1) = 1$ angenommen wird.

Das Neyman-Pearson-Kriterium ist folgendes: Man setzt einen Wert α mit $0 < \alpha < 1$ fest und wählt dann diejenige Entscheidungsregel δ, die

$$\sum_{x \in \mathfrak{X}^{(1)}} f(x \,|\, \theta_2) = \beta$$

minimiert, unter der Nebenbedingung, daß

$$\sum_{x \in \mathfrak{X}^{(2)}} f(x \,|\, \theta_1) \leq \alpha.$$

In der statistischen Testtheorie wird α als „Signifikanzniveau“ oder „Wahrscheinlichkeit des Fehlers 1. Art“ und β als „Wahrscheinlichkeit des Fehlers 2. Art“ bzw. $(1 - \beta)$ als „Teststärke“ bezeichnet.

4.3.8 Optimal — für wen?

Die formale Konstruktion einer Entscheidungsregel im Sinne der Entscheidungstheorie hat, wie bereits mehrfach betont wurde, nicht den Zweck, dem Entscheidungsträger vorzuschreiben, welche Entscheidungen zu treffen sind, sondern dient vor allem dazu, alle Argumente und Einflußgrößen eines Entscheidungsproblems offenzulegen. Wesentlich für das Ergebnis eines Entscheidungsverfahrens ist vor allem die Schadenstabelle und das Optimalitätskriterium. Für die Bestimmung der Schadenstabelle gibt es seitens der Entscheidungstheorie keine Einschränkung, außer daß es sich dabei mindestens um eine Ordinal- bzw. im Falle des Bayesschen Vorgehens um eine Intervallskala handeln muß. Ein Entscheidungsverfahren kann natürlich immer nur für den optimal sein, für den auch die Schadenstabelle bestimmt ist. Dabei kann es zu Konflikten zwischen den Interessen einer Institution und denen einer Einzelperson kommen. Das gleiche gilt für das Optimalitätskriterium. Unterschiedliche Optimalitätskriterien werden in der Regel zu unterschiedlichen Entscheidungen führen, und es können für die verschiedenen an einem Entscheidungsverfahren beteiligten Parteien verschiedene Optimalitätskriterien angemessen sein. Man kann sich dies leicht an einem Versicherungsbeispiel klarmachen. Ein Hausbesitzer etwa hat zwei Möglichkeiten: eine Brandversicherung abschließen (a_1), oder dies nicht zu tun (a_2). Es gibt zwei Zustände: das Haus brennt ab (θ_1) oder es brennt nicht ab (θ_2). Tabelle 4.8 enthält eine mögliche Schadenstabelle und angenommene a priori Wahrscheinlichkeiten.

Daneben enthält die Tabelle die Werte sowohl des Minimax- als auch des Bayes-Kriteriums. Die Kosten der Versicherung sind 60.– DM, der Wert des Hauses ist 250000.– DM. Der Versicherungsnehmer wird sich nach dem Minimax-Kriterium richten und die Versicherung abschließen (a_1), da der Schaden von 60.– DM für ihn keine große Beeinträchtigung darstellt, während der Verlust des Hauses ihn ruinieren kann. Auf lange Sicht mit dem Erwartungswert des Schadens von 50.– DM zu rechnen, ist für den Versicherungsnehmer als Einzelperson sinnlos, da er nach einem einzigen Schadensfall in der Regel keinen zweiten „Versuch“ haben wird.

Tabelle 4.8. Schadenstabelle für einen Hausbesitzer, der überlegt, ob es für ihn günstiger sei, eine Brandversicherung abzuschließen oder nicht. M' und B' sind die Werte des Minimax- und des Bayes-Kriteriums (hier ohne Daten).

	θ_1	θ_2	$M'(a)$	$B'(a)$
a_1	60	60	60	60
a_2	250000	0	250000	50
$\pi(\theta)$	0.0002	0.9998		

Anders ist dagegen die Position der Versicherung. Ihre Schadenstabelle zeigt Tabelle 4.9. Sie wird sich nach dem Bayes-Kriterium richten, da sie viele „Versuche“ gleichzeitig macht und deshalb „auf lange Sicht“ kalkulieren kann. Sie wird daher die Versicherung akzeptieren, weil ihre Gewinnerwartung bei a_1 (Akzeptieren der Versicherung) größer ist als bei a_2 (ablehnen der Versicherung).

Tabelle 4.9. Schadenstabelle der Versicherung zum Beispiel der Brandversicherung. Negative Schäden sind als Gewinne zu betrachten.

	θ_1	θ_2	$M'(a)$	$B'(a)$
a_1	249940	-60	249940	-10
a_2	0	0	0	0
$\pi(\theta)$	0.0002	0.9998		

4.3.9 Der Einfluß von Validität und Vorurteilen auf den Informationsgewinn

Mangelnde Validität

Das Bayessche Vorgehen verlangt die Kenntnis der a priori Verteilung $\pi(\theta)$ der Zustände und der bedingten Wahrscheinlichkeiten $f(X \mid \theta)$ der Daten bei gegebenen Zuständen. Damit können mit Hilfe des Theorems von Bayes nach Gleichung (4.11) auch die a posteriori Verteilungen $h(\theta \mid X)$ berechnet werden. Dies ist die Wahrscheinlichkeit des Vorliegens eines bestimmten Zustandes, wenn die Daten X bekannt sind. Hier kann man sehr gut erkennen, welchen Wert die Validität der Daten hat. Sind die Daten nicht valide, gilt also

$$f(X \mid \theta) \quad = \quad g(X),$$

so ist wegen

$$\sum_{\theta} f(X \mid \theta)\pi(\theta) \quad = \quad g(X)$$

die a posteriori Wahrscheinlichkeit gleich der a priori:

$$\begin{aligned} h(\theta \mid X) &= \frac{f(X \mid \theta)\pi(\theta)}{g(X)} \\ &= \frac{g(X)\pi(\theta)}{g(X)} \\ &= \pi(\theta). \end{aligned}$$

Die Erhebung der Daten stellt daher keinen Informationsgewinn bezüglich der Verteilung von θ dar.

Feste Vorurteile

Falls über das Bestehen eines bestimmten Zustandes θ^* weitgehende Sicherheit besteht: $\pi(\theta^*) = 1$, dann kann diese durch neue Daten nicht verändert werden:

$$\begin{aligned} g(X) &= \sum_{\theta} f(X \mid \theta)\pi(\theta) \\ &= f(X \mid \theta^*), \end{aligned}$$

denn aus $\pi(\theta^*) = 1$ folgt $\pi(\theta) = 0$, falls $\theta \neq \theta^*$. Damit ist

$$\begin{aligned} h(\theta \mid X) &= \frac{f(X \mid \theta)\pi(\theta)}{f(X \mid \theta^*)} \\ &= \begin{cases} 1 & \text{falls } \theta = \theta^* \\ 0 & \text{sonst.} \end{cases} \end{aligned}$$

4.3.10 Ein klinisches Beispiel

Im folgenden Beispiel von Birnbaum und Maxwell (1960) wird versucht, eine Gruppe von Patienten einer psychiatrischen Klinik mit Hilfe der Daten einer Eingangsuntersuchung in Behandlungskategorien einzuordnen. Von der Eingangsuntersuchung werden 9 Symptome ausgewählt, deren Vorhandensein bzw. Nichtvorhandensein binär kodiert wird. Der Stichprobenraum besteht damit aus der Menge aller binären 9-dimensionalen Vektoren: $\mathfrak{X} = \{0, 1\}^9$. Jede Komponente zeigt das Vorliegen eines der folgenden Symptome an:

1. Hysterische Symptome,
2. Ängste,
3. schizoide Denkstörungen,
4. Depressionen,
5. Zwangshandlungen oder -gedanken,
6. soziale Unsicherheit,
7. Stimmungsschwankungen bereits vor der gegenwärtigen Krankheit,
8. Impulsivität und/oder Aggressivität,
9. hypochondrische Einstellung gegenüber Krankheit.

Die Stichprobe besteht aus 772 Personen. Sie sollen den vier Gruppen (Zuständen) zugeordnet werden, die in Tabelle 4.10 aufgeführt sind. Die „wahren“ Zuordnungen wurden bei den Personen der Stichprobe durch mehrtägige stationäre Beobachtungen festgestellt. Die in der Tabelle genannten a priori Wahrscheinlichkeiten wurden aus einer größeren Stichprobe der Klinik gewonnen.

Tabelle 4.10. Kategorien und Häufigkeiten in der Patientenstichprobe

Kategorie	$N(\theta)$	$\pi(\theta)$
Neurotisch (θ_1)	341	0.461
Schizophren (θ_2)	174	0.237
Manisch Depressiv (θ_3)	148	0.157
Persönlichkeitsstörung (θ_4)	109	0.145
	772	1.000

Von den $2^9 = 512$ möglichen Datenmustern treten in der Stichprobe von 772 Personen nur 178 verschiedene auf. Aus der Anzahl der Personen in einer diagnostischen Kategorie, die ein bestimmtes Antwortmuster zeigen, und der Gesamtzahl der Personen in dieser Kategorie werden die Wahrscheinlichkeiten $f(X \mid \theta)$ durch die entsprechenden relativen Häufigkeiten geschätzt. So tritt etwa das Muster (0, 0, 0, 1, 0, 1, 0, 0, 0) bei genau 28 Personen auf. Davon entfallen 13 Fälle auf Neurotiker, 8 Fälle auf manisch Depressive, 5 auf Schizophrene und 2 auf Persönlichkeitsgestörte. Diese Zahlen dividiert durch die in Tabelle 4.10 angegebenen Häufigkeiten ergeben die

Werte in Tabelle 4.11 in der Spalte $f(X \mid \theta_i)$. Werden diese Zahlen mit den a priori Wahrscheinlichkeiten der Kategorien aus Tabelle 4.10 multipliziert, so ergeben sich die Werte in der Spalte $f(X \mid \theta_i)\pi(\theta_i)$, die wegen Gleichung (4.10) proportional sind zu den a posteriori Wahrscheinlichkeiten $h(\theta_i \mid x)$ dafür, daß ein Patient mit dem oben dargestellten Antwortmuster in die jeweilige Kategorie fällt.

Tabelle 4.11. Bedingte Wahrscheinlichkeiten eines Datenvektors und Gewichte der einzelnen Kategorien proportional zu den a posteriori Wahrscheinlichkeiten $h(\theta_i \mid x)$ für den Datenvektor $x = (0, 0, 0, 1, 0, 1, 0, 0, 0)$.

	$f(X \mid \theta_i)$	$f(X \mid \theta_i)\pi(\theta_i)$
θ_1	0.0381	0.0175
θ_2	0.0541	0.0085
θ_3	0.0287	0.0068
θ_4	0.0183	0.0026

Die Konstruktion der Entscheidungsregel erfolgt nach der in Abschnitt 4.3.5 dargestellten Methode. Für jeden Datenvektor x wird diejenige Alternative $a_i = \delta(x)$ gesucht, die Gleichung (4.12) minimiert. Die in diesem Beispiel von Birnbaum und Maxwell implizit verwendete Schadensfunktion ist $s(\theta_i, a_i) = 0$, $s(\theta_i, a_j) = 1$, falls $i \neq j$. Die Summe auf der rechten Seite von Gleichung (4.12) besteht im vorliegenden Beispiel für jedes a_i nur aus drei Summanden, nämlich den Werten $h(\theta_j \mid x)$, für $j \neq i$. Der Summand für $j = i$ wird durch den zugehörigen Schadenswert von $s(\theta_i, a_i) = 0$ zu 0. Es gilt aber für jedes $i = 1, \ldots, 4$

$$\sum_{\substack{j=1 \\ j \neq i}}^{4} h(\theta_j \mid x) = 1 - h(\theta_i \mid x).$$

Deshalb kann, um Gleichung (4.12) zu minimieren, einfach dasjenige $a_i = \delta(x)$ gesucht werden, das $h(\theta_i \mid x)$ maximiert. Statt $h(\theta_i \mid x)$ kann wegen der in Gleichung (4.10) dargestellten, bei gegebenem x geltenden Proportionalität zur Funktion $f(x \mid \theta_i)\pi(\theta_i)$ auch einfach dieses Produkt maximiert werden. Dies wird von Birnbaum und Maxwell getan. Tabelle 4.11 enthält die Werte von $f(x \mid \theta_i)\pi(\theta_i)$. Da die erste dieser Zahlen die größte ist, wird dem Datum $x = (0, 0, 0, 1, 0, 1, 0, 0, 0)$ die Kategorie 1 = „Neurotiker“ zugeordnet.

Tabelle 4.12 gibt die relativen Häufigkeiten der mit dieser Methode korrekt und falsch klassifizierten Personen an. Es zeigt sich, daß die Anzahl der Korrekt- und Fehlklassifikationen für die verschiedenen Kategorien stark unterschiedlich ist.

Tabelle 4.12. Relative Klassifikationshäufigkeiten bei den a priori Wahrscheinlichkeiten aus Tabelle 4.10.

	θ_1	θ_2	θ_3	θ_4
a_1	.856	.241	.595	.422
a_2	.038	.706	.034	.073
a_3	.073	.034	.338	.055
a_4	.032	.017	.034	.450
$\pi(\theta_i)$	.461	.237	.157	.145

Will man eine gleichmäßigere Güte der Entscheidungsregel erreichen, dann kann dies durch eine Änderung der Gewichtungen durch die a priori

Verteilungen geschehen. Wählt man statt der in Tabelle 4.10 gezeigten Werte für $\pi(\theta)$ den Wert 0.25 für alle Kategorien, so ergeben sich die Klassifikationshäufigkeiten in Tabelle 4.13.

Tabelle 4.13. Relative Klassifikationshäufigkeiten unter der Annahme a priori gleichwahrscheinlicher Kategorien.

	θ_1	θ_2	θ_3	θ_4
a_1	.388	.057	.140	.072
a_2	.000	.577	.008	.000
a_3	.384	.229	.772	.128
a_4	.228	.137	.080	.800
$\pi(\theta_i)$	.250	.250	.250	.250

Die Wahl von empirisch nicht begründeten a priori Wahrscheinlichkeiten führt insgesamt zu einer Verschlechterung der Klassifikationsleistung. Die Wahrscheinlichkeit einer korrekten Klassifikation ergibt sich aus

$$\sum_{i=1}^{4} P(a_i \mid \theta_i)\pi(\theta_i)$$

Nimmt man die Werte von $\pi(\theta)$ in Tabelle 4.12 als die wahren an, dann ergibt sich eine Wahrscheinlichkeit von 0.680. Wird die Entscheidungsregel mit Hilfe der Werte für $\pi(\theta)$ aus Tabelle 4.13 konstruiert und sind die Werte in Tabelle 4.12 die wahren a priori Verteilungen, dann ergibt sich eine Wahrscheinlichkeit von 0.557 für korrekte Klassifikationen. Dies war zu erwarten, da die Klassifikationsregel nur für diejenige a priori Verteilung optimal ist, für die sie entwickelt wurde. Die hier implizit gewählte Schadensfunktion hat ja die Anzahl korrekter Klassifikationen zum Ziel. Eine unterschiedliche Bewertung der einzelnen Klassifikationsfehler wird dadurch nicht vorgenommen.

4.3.11 „Richtige“ und „falsche“ Diagnosen

Es wurde hier mehrfach darauf hingewiesen, daß jeder diagnostischen Entscheidung eine Schadensfunktion zugrunde liegt. Auch wenn diese nicht explizit offengelegt wird, geht sie in jedem Fall in die Bewertung der Konsequenzen ein. Im Fall von zwei Zuständen und zwei Entscheidungsalternativen wurde gezeigt, daß alle zulässigen Entscheidungsregeln Likelihoodquotientenregeln sind und diese wiederum immer als Bayes-Regeln betrachtet werden können. Wie aus Gleichung (4.14) zu erkennen ist, hängt die Bayes-Entscheidung von zwei Größen ab: Dem Quotienten der Differenzen der Schadenswerte zwischen „richtiger“ und „falscher“ Entscheidung und dem Quotienten der a priori Wahrscheinlichkeiten:

$$q = \frac{\pi(\theta_2)}{\pi(\theta_1)} \frac{[s(\theta_2, a_2) - s(\theta_2, a_1)]}{[s(\theta_1, a_1) - s(\theta_1, a_2)]} \tag{4.14}$$

In ihrem Effekt auf den Likelihoodquotienten sind diese beiden Faktoren austauschbar. Dies wird im Beispiel von Birnbaum und Maxwell (1960) demonstriert. Dort wird versucht, die Klassifikationshäufigkeiten durch eine spezielle Wahl der a priori Wahrscheinlichkeiten möglichst gleichmäßig zu verteilen.

Wir sind immer davon ausgegangen, daß die Alternativen a unabhängig von den Zuständen θ sind. In diesem Fall kann es auf den ersten Blick unklar sein, wann von „richtiger“ und wann von „falscher“ Entscheidung zu sprechen ist. Diese Einteilung ist aber einfach aufgrund der Schadensfunktion zu treffen. Ist die Schadensfunktion konstant für alle Paare (θ, a) gleich, dann ist es tatsächlich sinnlos von „richtig“ und „falsch“ zu sprechen. In diesem

Fall ist auch die Anwendung der Gleichung (4.14) unmöglich. Sie ist auch überflüssig, da ja alle Konsequenzen gleich sind.

Häufig wird $s(\theta_1, a_1) = s(\theta_2, a_2) = 0$ und $s(\theta_1, a_2) = s(\theta_2, a_1) = 1$ gesetzt. Der Effekt ist, daß damit einfach die Anzahl der insgesamt korrekt klassifizierten Personen maximiert wird. „Korrekt“ sind in diesem Fall alle Ergebnisse, für die $s(\theta, a) = 0$, also kein Schaden auftritt. „Falsch“ sind Ergebnisse, die einen Schaden erzeugen, für die also $s(\theta, a) = 1$. Dieses Vorgehen wird auch von Birnbaum und Maxwell (1960) gewählt. Allerdings wird dann der Likelihoodquotient (4.14) durch verschiedene Annahmen über $\pi(\theta)$ verändert.

Von „korrekt“ und „falsch“ zu sprechen ist also nur dort sinnvoll, wo $s(\theta, a) = 0$ für „korrekte“ Kombinationen (θ, a) und $s(\theta, a) > 0$ für „falsche“ Kombinationen (θ, a). In der hier vorliegenden Darstellung wird stets von einer allgemeinen Schadensfunktion ausgegangen, da diese Vorgehensweise flexibler ist als eine dichotome Klassifizierung in „richtig“ und „falsch“.

4.4 Selektionsentscheidungen

Im Fall einer einfachen Selektionsentscheidung liegt als Datum eine Ausprägung eines kontinuierlichen Testwertes X vor, das mit einem Bewährungskriterium Y korreliert ist. Als „geeignet“ werden alle Personen betrachtet, deren Bewährungswert Y über einem bestimmten Kriterium y_c liegt. Die Selektion erfolgt aufgrund der Ausprägung von X. Es werden alle Personen akzeptiert, deren Ausprägung von X über einen Kriteriumswert x_c liegt. Abbildung 4.6 veranschaulicht die Situation. Wir gehen von einer bivariaten Normalverteilung für X und Y aus, die Korrelation $\rho(X, Y)$ zwischen dem Testwert X und dem Kriterium Y ist die *Validität* des Tests. Als *Selektionsquote* wird der Anteil der Personen in der Population bezeichnet, deren Testwert über x_c liegt, die daher als geeignet ausgewählt werden. Als *Bewährungsquote* wird der Anteil von Personen in der Gesamtpopulation bezeichnet, deren Kriteriumswert über y_c liegt, die sich also bewähren.

4.4.1 Validität und Erfolgsquote

Die Leistung eines Tests zur Vorhersage eines Kriteriums wird durch seinen Validitätskoeffizienten beschrieben. Eine Möglichkeit, die Güte eines Tests zu beurteilen, besteht darin anzugeben, wie groß der durch den Test erklärte Varianzanteil der Kriteriumsvariablen ist. Ein Test mit der Validität 0.50 erklärt 25% der Varianz des Kriteriums. Ein wesentliches Merkmal des Zusammenhangs zwischen Validität und erklärtem Varianzanteil ist, daß letzterer mit zunehmender Validität immer stärker wächst, da dieser Zusammenhang quadratisch ist. Mittlere Werte der Validität ergeben daher nur kleine Werte bei den erklärten Varianzanteilen. Erst verhältnismäßig hohe Validitäten können einen zufriedenstellenden Anteil der Kriteriumsvarianz erklären. Um die Hälfte der Kriteriumsvarianz aufzuklären, wird eine Validität von 0.71 benötigt. Eine derart hohe Validität ist aber bei einem psychologischen Test nicht erreichbar. Die meisten Validitäten bekannter Tests liegen zwischen 0.2 und 0.4, in Ausnahmefällen bis zu 0.5. Die damit erklärbaren Varianzanteile bewegen sich also von 4 bis 16 oder maximal 25%.

Von Taylor und Russel (1939) wird darauf hingewiesen, daß die Bewertung der Güte eines Tests allein aufgrund des erklärten Varianzanteils unbefriedigend ist, da in eine solche Bewertung nur wenige Parameter der Entscheidungssituation eingehen, in der der Test benutzt wird. Für die durch einen Test erreichbare Verbesserung von Selektionsentscheidungen sind nämlich neben der Validität auch die Selektionsquote, also der Anteil der zu selegierenden Personen, und die natürliche Bewährungsquote, der Anteil der Personen, der sich ohne Selektion bewährt, von Bedeutung.

Der Abbildung 4.6 veranschaulicht die Situation. Ihr können folgende Feststellungen entnommen werden:

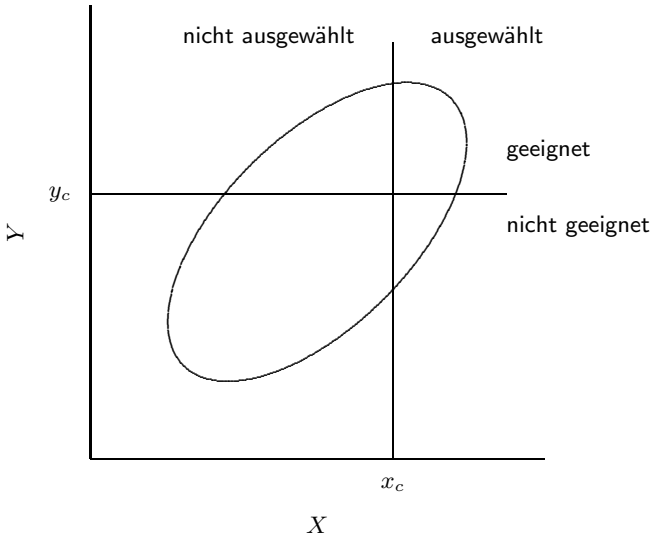

Abb. 4.6. Bei Selektionsentscheidungen wird von einer bivariaten Normalverteilung des Testergebnisses X und des Kriteriums Y ausgegangen. „Geeignet" sind alle Personen, deren Kriteriumswert größer ist als y_c. Als geeignet ausgewählt werden die Personen, deren Testergebnis größer als x_c ist. Die Abbildung zeigt, daß eine Vergrößerung des Anteils der korrekt als geeignet eingestuften Personen gleichzeitig immer auch zu einer Vergößerung des Anteils der Personen führt, die als nicht geeignet bezeichnet werden, obwohl sie tatsächlich geeignet sind. Man kann auch erkennen, daß eine extrem kleine Selektionsquote (also eine hoher Wert von x_c) immer zu sehr hohen Trefferquoten bei der Gruppe der geeigneten führt. Eine große natürliche Bewährungsquote (also ein kleiner Wert von y_c) dagegen macht den Test überflüssig. Ist die Validität sehr klein, dann nimmt die gezeichnete Isodensite die Form eines Kreises an, und der Anteil von korrekt und falsch klassifizierten Personen ist bei allen Selektionsquoten gleich.

1. Je kleiner die Selektionsrate, desto größer ist der Anteil der Geeigneten in der ausgewählten Teilmenge, desto größer ist aber auch der Anteil der Geeigneten in der zurückgewiesenen Teilmenge.
2. Eine hohe Bewährungsquote macht die Testerhebung überflüssig.
3. Je kleiner die Validität (d.h. je kreisförmiger die Isodensite) desto weniger effektiv ist die Selektion.
4. Bei bivariater Normalverteilung in der Gesamtpopulation können die Merkmale X und Y in den ausgewählten Teilgruppen *nicht* mehr normalverteilt sein, da durch die Selektion ein nicht-symmetrischer Teil der Stichprobe selegiert wird.

Wir betrachten folgendes Selektionsbeispiel, das in der Abbildung 4.7 nach Taylor und Russel (1939) dargestellt ist. Die natürliche Bewährungsquote sei 0.5, der kritische Wert y_c liegt also beim Mittelwert von Y. Von den Bewerbern seien 30% auszuwählen, $Q(x_c)$ ist also 0.3. Wählt man nun mit Hilfe des Testwerts X, der eine Validität von 0.5 haben soll, die Personen über x_c, dann werden davon 74% einen Kriteriumswert Y erreichen, der größer ist als y_c. Von den mit Hilfe des Tests ausgewählten Personen werden sich also 74% bewähren, während sich ohne Test nur 50% bewährt hätten. Hat der Test die Validität 0, dann werden sich auch mit Testauswahl nur 50% der ausgewählten Personen bewähren, da dies der natürlichen Bewährungsquote entspricht. Die geringstmögliche Validität führt also zu einer Bewährungsquote von 0.5, die bestmögliche Validität von 1 zu einer Bewährungsquote von 1.0. Die Validität von 0.5 ergibt eine Bewährungsquote von 0.74. Sind statt 30% nur 10% aller Bewerber auszuwählen, so ergibt die gleiche Validität eine Bewährungsquote von 0.84.

Damit ist demonstriert, daß der erklärte Varianzanteil nur eine unzureichende Beschreibung des Testnutzens ist, da er weder die Selektionsquote, noch die natürliche Bewährungsquote berücksichtigt. Es ist auch gezeigt, daß bereits bei geringen Validitäten von 0.2 bis 0.4 eine wertvolle Verbesserung der Bewährungsquote erreicht werden kann, wenn die Selektionsquote

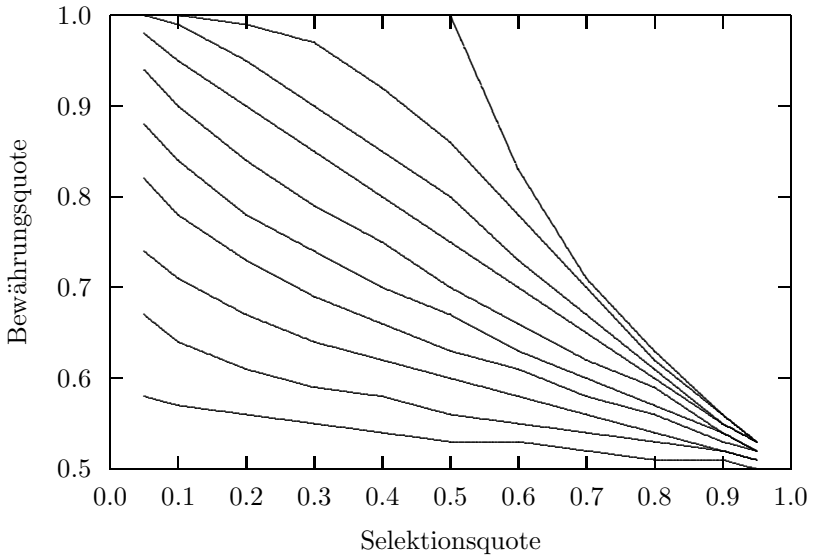

Abb. 4.7. Zusammenhang zwischen Selektionsquote und effektiver Bewährungsquote bei Validitäten von 0.1 bis 1.0 und bei einer natürlichen (d.h. in einer nicht selegierten Population bestehenden) Bewährungsquote von 0.5. Die unterste Kurve gilt für eine Validität von 0.1, die oberste für eine Validität von 1.0. Die Differenz zwischen der natürlichen Bewährungsquote von 0.5 und der auf der Ordinate abgetragenen effektiven Bewährungsquote ist die durch den Test erzielte Verbesserung der Bewährungsquote (nach Taylor & Russel, 1939).

nicht zu klein ist.

4.4.2 Nutzenanalyse

Die folgenden Überlegungen zum Nutzen eines Tests gehen auf Brogden (1946, 1949) und Cronbach und Gleser (1965) zurück. Die wesentliche Annahme ist, daß es einen linearen Zusammenhang zwischen dem Testwert X und dem erwarteten Nutzen gibt, den eine Behandlung bringt, wenn eine Person mit einem bestimmten Testwert x für die Behandlung akzeptiert wird. Hier ist eine Übersicht über alle Annahmen:

1. Die Population von Personen, aus der selegiert wird, ist unbegrenzt. A priori Verteilungen beziehen sich auf diese Population, von der angenommen wird, daß sie bereits durch alle Selektionsverfahren gefiltert wurde, die hier *nicht* zur Diskussion stehen.

2. Für jede Person gibt es zwei Entscheidungsalternativen: annehmen (a_1) oder zurückweisen (a_2).

3. Für jede Person gibt es einen Testwert X. X ist eine normalverteilte Zufallsvariable mit Erwartungswert $\mathscr{E}(X) = 0$ und Varianz $\sigma^2(X) = 1$.

4. Zu jeder Person, für die die Entscheidung „annehmen“ getroffen wird, gibt es einen Nutzenwert U. Die Korrelation $\rho(X, U)$ zwischen dem Testwert X und dem Nutzen U sei positiv.

5. Eine Person, die zurückgewiesen wird (a_2), erzeugt den Nutzen U'. Dieser kann auch als 0 angenommen werden.

6. Die mittleren Testkosten pro Person sind $C > 0$.

7. Die Entscheidungsregel besteht darin, Personen mit hohem Testwert zu akzeptieren und solche mit niedrigem Testwert zurückzuweisen. Der Kriteriumswert x_c wird so gewählt, daß ein bestimmter Anteil $\Phi(x_c)$ von Personen einen Testwert erreicht, der über x_c liegt. Es gilt also

$$\delta(x) = \begin{cases} a_1 & \text{falls } x \geq x_c \\ a_2 & \text{sonst.} \end{cases}$$

Aus Annahme 4 folgt, daß für den erwarteten Nutzen einer akzeptierten Person mit dem Testergebnis x gilt

$$\mathscr{E}(U \mid x, a_1) = \sigma(U)\,\rho(X,U)\,x + U_0.$$

Dabei ist $\sigma(U)$ die Standardabweichung der Zufallsvariablen U, $\rho(X,U)$ die Korrelation von X und U, also die Validität, und U_0 der erwartete Nutzen einer Person mit dem Testwert $x = 0$, also dem Mittelwert, falls diese Person akzeptiert wird. Wegen Annahme 5 gilt ferner

$$\mathscr{E}(U \mid x, a_2) = U'.$$

Für den erwarteten Nutzen einer Person mit dem Testwert x gilt

$$\mathscr{E}(U \mid x) = P(a_1 \mid x)[\sigma(U)\,\rho(X,U)\,x + U_0] + P(a_2 \mid x)\,U'.$$

Wir nehmen an, daß $U' = 0$, womit der zweite Summand wegfällt:

$$\mathscr{E}(U \mid x) = P(a_1 \mid x)(\sigma(U)\,\rho(X,U)\,x + U_0).$$

Für den erwarteten Nutzen einer beliebigen Person gilt

$$\mathscr{E}(U) = \int_{-\infty}^{\infty} \phi(x) P(a_1 \mid x)\,[\sigma(U)\,\rho(X,U)\,x + U_0]\,dx,$$

wobei $\phi(x)$ die Wahrscheinlichkeitsdichte von X ist. Aus Annahme 7 folgt, daß $P(a_1 \mid x) = 0$ für alle $x < x_c$ und $P(a_1 \mid x) = 1$ für $x \geq x_c$. Wir erhalten deshalb

$$\begin{aligned}
\mathscr{E}(U) &= \int_{x_c}^{\infty} \phi(x)\,[\sigma(U)\,\rho(X,U)\,x + U_0]\,dx \\
&= \sigma(U)\,\rho(X,U) \int_{x_c}^{\infty} \phi(x) x\,dx + U_0 \int_{x_c}^{\infty} \phi(x)\,dx \\
&= \sigma(U)\,\rho(X,U) \int_{x_c}^{\infty} \phi(x) x\,dx + U_0 Q(x_c). \qquad (4.15)
\end{aligned}$$

Die letzte Gleichung folgt aus

$$\begin{aligned}
\Phi(x_c) &= \int_{x_c}^{\infty} \phi(x)\,dx. \\
&= P(X \geq x_c).
\end{aligned}$$

$Q(x_c)$ ist der Anteil der Personen an der Gesamtpopulation, der akzeptiert wird, wenn x_c der Kriteriumswert ist. Wird kein Test durchgeführt, dann ist der erwartete Nutzen $U_0\pi$ wobei π die a priori Wahrscheinlichkeit dafür ist, akzeptiert zu werden. Für jedes π gibt es natürlich ein x_c, so daß $Q(x_c) = \pi$. In diesem Fall ist $U_0\pi = U_0 Q(x_c)$. Zieht man diesen auch ohne Test zu erwartenden Nutzen und die Testkosten C von Gleichung (4.15) ab, erhält man den Nettonutzen U_Δ:

$$\begin{aligned}
U_\Delta &= \mathscr{E}(U) - U_0\pi - C \\
&= \sigma(U)\,\rho(X,U) \int_{x_c}^{\infty} \phi(x) x\,dx - C \qquad (4.16) \\
&= \sigma(U)\,\rho(X,U)\,\mathscr{E}(X \mid X \geq x_c) - C. \qquad (4.17)
\end{aligned}$$

Die Vereinfachung zu Gleichung (4.17) ist möglich, da das Integral genau der Erwartungswert von X unter der Bedingung $X \geq x_c$ ist. Für beliebige Verteilungen von X ist damit gezeigt, daß der Nettonutzen eines Tests eine *lineare* Funktion der Validität, der Standardabweichung des Kriteriums in einer nicht mit Hilfe des Tests selegierten Population und des bedingten Erwartungswertes des Testergebnisses in der selegierten Population ist. Kann man normalverteilte Testwerte annehmen, so gilt

$$U_\Delta = \sigma(U)\,\rho(X,U)\frac{\phi(x_c)}{Q(x_c)} - C, \qquad (4.18)$$

da für die Wahrscheinlichkeitsdichte der Normalverteilung gilt

$$\mathscr{E}(X \mid X \geq x_c) = \frac{\phi(x_c)}{Q(x_c)}.$$

Gleichung (4.18) legt eine neue Interpretation des Validitätskoeffizienten nahe: Er gibt den Anteil vom maximal möglichen Nutzen an, der durch einen Test mit der gegebenen Validität erreicht werden kann. Der maximal mögliche Nutzen wird im wesentlichen durch die Standardabweichung des Nutzens in der nicht selegierten Population $\sigma(U)$ und durch die Selektionsquote $Q(x_c)$ bestimmt.

4.4.3 Anwendungen

Die hier beschriebenen Selektionsmethoden wurden bis zum Anfang der 80er Jahre wenig angewandt, obwohl sie seit Brogden (1946, 1949) bekannt sind. Von Schmidt, Hunter, McKenzie & Muldrow (1979) werden dafür drei Gründe genannt:

1. Man glaubte, die Annahme normalverteilter Testwerte sei essentiell für die Theorie, empirisch aber ebensowenig gerechtfertigt wie die Homoskedastizität der Testwertvarianzen und die Linearität des Zusammenhangs zwischen Testwert und Nutzen. Die obige Ableitung zeigt dagegen, daß die Normalverteilung der Testwerte nicht essentiell ist. Darüber hinaus zeigen empirische Untersuchungen, daß Homoskedastizität und Linearität bei hinreichend großen Stichproben auch gefunden wird (Schmidt et al., 1979). Da diese Forderungen ja Forderungen über die Population und nicht über die Stichproben sind, ist damit gezeigt, daß dieser Einwand nicht stichhaltig ist.

2. Man ging lange Zeit davon aus, daß die Validität eines Tests sehr stark situationsabhängig sei. In zahlreichen empirischen Arbeiten wurden große Schwankungen der Validität eines Tests gefunden, auch wenn die Kriteriumsanforderungen nur geringfügig geändert wurden (Ghiselli, 1966). In mehreren Metaanalysen von Validitätsstudien wurde aber überzeugend nachgewiesen, daß diese Schwankungen der Validitätsschätzungen Artefakte einer zu kleinen Anzahl von Beobachtungen sind (Schmidt, Hunter, Pearlman, Hirsh, Sackett, Schmitt, Tenopyr, Kehoe & Zedeck, 1985; Schmidt, Hunter & Raju, 1988). Testvaliditäten sind durchaus auf ähnliche Aufgabenanforderungen generalisierbar.

3. Es schien lange Zeit sehr schwierig, die Standardabweichung $\sigma(U)$ des Nutzens in einer nicht ausgelesenen Population zu schätzen, was Voraussetzung zur Berechnung des Nutzens nach Gleichung (4.18) ist. Die im folgenden dargestellte Methode von Schmidt et al. (1979) zeigt aber, daß dies durchaus mit vertretbarem Aufwand möglich ist.

Die Methode von Schmidt et al. (1979) zur Schätzung der Standardabweichung des Nutzens $\sigma(U)$ besteht in einer Befragung von Supervisoren über die Leistung ihrer Untergebenen. Sie sollen den Wert der Leistung eines mittelmäßigen Mitarbeiters und der eines Mitarbeiters angeben, der besser als etwa 85% der anderen ist. Aus der Differenz dieser beiden Angaben wird $\sigma(U)$ berechnet. Um die Reliabilität der Befragung zu erhöhen, werden Hilfestellungen gegeben. Diese können etwa darin bestehen anzugeben, wieviel es vermutlich kosten würde, die äquivalente Leistung durch eine Fremdfirma ausführen zu lassen. Von Schmidt et al. (1979) wurden in einer Vorstudie auf diese Weise 62 Supervisoren von Buchprüfern („budget analysts“) befragt. Als Mittelwert für $\sigma(U)$ ergab sich bezogen auf ein Jahr \$11327, der Standardschätzfehler des Mittelwerts lag bei \$1120. Damit liegen 90% der Antworten zwischen \$9480 und \$13175. Die Methode erlaubt also eine zuverlässige Erhebung der Standardabweichung von Nutzenwerten.

In der Studie von Schmidt et al. (1979) wird berechnet, welchen Effekt die Benutzung eines Eignungstest bei der Neueinstellung von Programmierern in den öffentlichen Dienst der USA hätte. Als Test wird der „Programmer Aptitude Test“ (PAT) verwendet. Validitätsschätzungen ergaben für diesen Test eine Validität von 0.76 als Mittelwert aus allen bekannten Validitätsstudien. Die Schätzung von $\sigma(U)$ wurde nach dem oben beschriebenen Verfahren durch Supervisoren in mehreren Institutionen durchgeführt. Als

Schätzwert ergab sich $\sigma(U) = \$10413$, gemittelt über alle Supervisoren und alle Laufbahnebenen der einzustellenden Programmierer.

Um den echten Nutzengewinn durch den PAT beurteilen zu können, wurde dieser nicht mit einer Zufallsauswahl, sondern mit einem hypothetischen Verfahren verglichen, dessen Validität als von 0 verschieden angenommen wurde. Als Anzahl neueinzustellender Programmierer wurde aus den Beschäftigungszahlen von 1970 die Zahl von 618 für den öffentlichen Dienst der USA und von 10210 Personen für die Gesamtwirtschaft der USA geschätzt, ein Wert der inzwischen sicher erheblich höher anzusetzen ist. Die von Schmidt et al. (1979) benutzte Formel zur Berechnug des Gesamtnutzens ist

$$t\,N\,[\rho(X,U) - \rho(Y,U)]\,\sigma(U)\,\frac{\phi(x_c)}{Q(x_c)}\;-\;N\,\frac{(C_X - C_Y)}{Q(x_c)}. \tag{4.19}$$

Dabei ist t die mittlere Anzahl von Jahren, während der ein neu eingestellter Programmierer eine Stelle beibehält (geschätzt mit 9.69 Jahre). N ist die Anzahl eingestellter Personen (618 für den öffentlichen Dienst und 10210 für alle Arbeitgeber der USA). $\rho(X,U)$ ist die Validität des PAT (0.76) und $\rho(Y,U)$ die Validität des Einstellungsverfahrens ohne Test. C_X sind die Testkosten pro Person (als \$10 angenommen) und C_Y die Kosten des Einstellungsverfahrens ohne Test (als \$0 angenommen). Die Größen $\phi(x_c)$ und $Q(x_c)$ sind definiert wie in Gleichung (4.18).

Die genauen Ergebnisse können bei Schmidt et al. (1979) nachgelesen werden. Hier nur einige Beispiele: Ist die Selektionsrate $Q(x_c) = 0.05$ und die Validität $\rho(Y,U)$ des Verfahrens ohne Test gleich 0, dann ist der über 9.69 Jahre durch eine 1-jährige Anwendung des PAT erzielte Produktivitätszuwachs für den öffentlichen Dienst gleich 97.2 Millionen Dollar. Ist dagegen die Selektionsrate gleich 0.8 und $\rho(Y,U) = 0.5$, dann ist der Produktivitätszuwachs 5.6 Millionen Dollar. Diese Zahlen gelten nur für die 618 Neueinstellungen des öffentlichen Dienstes. Die Werte für die Gesamtwirtschaft bei 10210 Neueinstellungen pro Jahr erhöhen sich entsprechend. Um den Produktivitätszuwachs pro eingestellter Person zu berechnen, brauchen die Zahlen nur durch 618 dividiert werden. Im ersten Fall ergibt sich \$157281 pro Person, im zweiten \$9061 pro Person.

In einer Literaturübersicht wird von Schmidt und Hunter (1983) gezeigt, daß die Nutzenstreuung $\sigma(U)$ auch als Funktion des mittleren erzeugten Geldwertes oder des mittleren Gehalts einer Berufsgruppe geschätzt werden kann. In früheren Untersuchungen wurde gefunden, daß $\sigma(U)$ mit 24% bis 34% des mittleren erzeugten Geldwerts der entsprechenden Berufsgruppe angenommen werden kann. Die Literaturanalyse von Schmidt und Hunter (1983) ergab einen Durchschnittswert von 18.5% des mittleren erzeugten Geldwerts. Diese Werte gelten nicht für Akkordarbeit. Dort ist die Streuung geringer, sie macht etwa 15% des mittleren erzeugten Geldwerts aus. Da 57% des Bruttosozialprodukts der USA für Löhne und Gehälter aufgewandt werden, könne diese Anteile auf die mittleren Gehälter umgerechnet werden. Es ergeben sich als Schätzwerte für $\sigma(U)$ bei nicht nach Stückzahlen entlohnten Tätigkeiten 42% bis 60% des mittleren Gehalts.

Die Ergebnisse von Schmidt, Hunter und Mitarbeitern können also benutzt werden, um eine einfache Abschätzung der Standardabweichung $\sigma(U)$ zu erhalten. Für nicht nach Akkord bezahlte Tätigkeiten ist 40% des mittleren Gehalts der entsprechenden Tätigkeit als konservativer Schätzwert geeignet.

Empirische Untersuchungen des Nutzens von psychologischen Selektionsmethoden bei der Personalauslese werden von Schmidt, Hunter, Outerbridge und Trattner (1986) berichtet. Sie beziehen sich auf Einstellungen im öffentlichen Dienst der USA. Die Leistung der Probanden wurde durch mehrere Maße gemessen, darunter Arbeitsproben und Beurteilungen von Supervisoren. Insgesamt wurden 347 Personen untersucht, die aufgrund von Testergebnissen eingestellt worden waren, und 326 Personen, die ohne Testdaten eingestellt worden waren. Im Mittel war die Leistung der Personengruppe mit Tests bei den verschiedenen Kriteriumsmaßen 0.487 Standardabweichungen besser als die Personengruppe ohne Tests. Setzt man $\sigma(U)$

mit 40% des mittleren Gehalts an, dann ergibt dies einen Netto-Testnutzen von \$1758 pro Person und Jahr bei der untersten Gehaltsstufe bis \$15893 pro Person und Jahr bei der höchsten Gehaltsstufe. Umgerechnet auf die mittlere Verweildauer der Personen in einer Stellung von etwa 13 Jahren und auf durchschnittlich 225731 Neueinstellungen pro Jahr im öffentlichen Dienst der USA ergibt sich damit ein Testnutzen von etwa 7.8 Milliarden Dollar bei einer Testanwendung von einem Jahr. Kann ein einmal konstruierter Test über mehr als ein Jahr benutzt werden, vervielfacht sich der Nutzen mit der Anzahl der Jahre.

Einige weitere Verfeinerungen von Gleichung (4.19) werden von Gerpott (1989) vorgeschlagen. Unter anderem bezieht er die Verzinsung des Nutzens, den Effekt der Ablehnung von Einstellungsangeboten und die Bezugnahme auf andere Personalauslesemaßnahmen mit ein. Er berichtet auch ein Anwendungsbeispiel, in dem die Erfolgswirkungen eines verbesserten, strukturierten Auswahlgesprächsverfahrens untersucht wurden.

Übungsaufgaben

1. Skizzieren Sie die verschiedenen Elemente einer Entscheidungssituation, vor der ein Psychologe steht, der über die Erteilung der Fahrerlaubnis an einen Autofahrer zu gutachten hat, dem wegen Trunkenheit am Steuer die Fahrerlaubnis entzogen war. Welche globalen statistischen Daten werden benötigt?

2. In welchem Element eines Entscheidungsproblems steckt psychologische Information? In welcher Form?

3. Wir betrachten ein Spiel, in dem die Menge der Zustände zwei Elemente θ_1 und θ_2 enthält, die in jeder Runde von einem intelligenten Gegenspieler willkürlich festgelegt werden können. Es gibt zwei Handlungsalternativen a_1 und a_2. Die Schadenstabelle sei

	θ_1	θ_2
a_1	-1	1
a_2	1	-1

   Warum ist bei mehrfacher Wiederholung dieses Spiels eine randomisierte Strategie einer reinen vorzuziehen?

4. Tabelle 4.2 gibt die Werte des Minimax-Optimalitätskriteriums für die Schadensfunktion aus Tabelle 4.1 an. Man berechne aus Tabelle 4.1 die dazugehörige Regretfunktion und für diese die Werte des Minimax-Optimalitätskriteriums. Welche Alternative ist zu wählen, wenn die Entscheidung nicht auf der Schadenstabelle 4.1, sondern auf der Regretfunktion aufgebaut wird?

5. Aufgabe 8-9 aus Lindgren (1976) Seite 373: Bestimmen Sie aus der folgenden Schadenstabelle die (reinen) Minimax-Lösungen für die Schadens- und die korrespondierende Regretfunktion.

	θ_1	θ_2	θ_3
a_1	2	–3	–1
a_2	4	0	5
a_3	1	1	–1
a_4	0	2	–2

6. Seien $\Theta = \Delta = \mathbb{R}$ und θ und a damit reellwertige Parameter. Die Schadensfunktion sei definiert durch

$$s(\theta, a) = (\theta - a)^2.$$

Was ist die Risikofunktion einer für θ erwartungstreuen Schätzfunktion δ?

Hinweis: Eine Schätzfunktion δ ist *erwartungstreu* für den Parameter θ, wenn für alle θ aus Θ gilt $\mathscr{E}_{x\,|\,\theta}[\delta(x)] = \theta$.

7. Wie kommt die Zahl k^m der möglichen Entscheidungsfunktionen zustande?

8. Zeigen Sie, daß eine Entscheidungsregel, die eine nicht zulässige Alternative als Funktionswert hat, selbst auch nicht zulässig ist.

9. Beim Minimax-Verfahren im Fall mit Daten gibt es zwei Möglichkeiten, statt von den Schäden vom Regret auszugehen. Man kann den Regret sofort von den Schäden berechnen oder erst von den Risiken. Zeigen Sie, daß beide Methoden zum gleichen Ergebnis führen.

10. Welchen Wert hat $S_h(\delta, x)$ von Gleichung (4.12), wenn alle Werte der Schadensfunktion konstant gleich s sind?

11. Das Beispiel von Birnbaum und Maxwell (1960) in Abschnitt 4.3.10 verwendet zur Klassifikation einen 9-dimensionalen Datenvektor. Konstruieren Sie für dieses Problem eine optimale Entscheidungsregel ohne Verwendung individueller Daten, und berechnen Sie den Zuwachs in der Wahrscheinlichkeit einer korrekten Klassifikation, der durch die Verwendung von Daten erreicht wird.

Literatur

Andersen, E. B. (1973a). Conditional Inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, **26**, 31–44.

Andersen, E. B. (1973b). A goodness of fit test for the Rasch model. *Psychometrika*, **38**, 123–140.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, **42**, 69–81.

Baker, F. B. (1992). *Item response theory: parameter estimation techiques.* New York: Marcel Dekker.

Bamberg, G. (1972). *Statistische Entscheidungstheorie*. Würzburg: Physica-Verlag.

Bauer, H. (1991). *Wahrscheinlichkeitstheorie* (4. Aufl.). Berlin: De Gruyter.

Berger, J. O. (1980). *Statistical decision theory, foundations, concepts, and methods*. New York: Springer-Verlag.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Birnbaum, A. & Maxwell, A. E. (1960). Classification procedures based on Bayes's formula. *Applied Statistics*, **9**, 152–169 (Nachdruck in Cronbach & Gleser, 1965).

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. *Psychometrika*, **46**, 443–459.

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, **35**, 197–197.

Brémaud, P. (1988). *An introduction to probabilistic modeling*. New York: Springer-Verlag.

Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, **37**, 65–76.

Brogden, H. E. (1949). A new coefficient: application to biserial correlation and to estimation of selective efficiency. *Psychometrika*, **14**, 169–182.

Chernoff, H. & Moses, L. E. (1959). *Elementary decision theory*. New York: Wiley.

Chung, K. L. (1975). *Elementary probability theory with stochastic processes*. New York: Springer-Verlag.

Colonius, H. (1980). Representation and uniqueness of the Bradley-Terry-Luce model for pair comparisons. *British Journal of Mathematical and Statistical Psychology*, **33**, 99–103.

Cronbach, L. J. & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

von Davier, M. (1994). WINMIRA: A windows 3.x program for analyses with the Rasch model, with latent class analysis, and with the mixed Rasch model. Kiel: Intitute for Science Education (IPN).

von Davier, M. & Rost, J. (1995). Polytomous Mixed Rasch Models. In G. Fischer & I. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments and Applications* (S. 371–379). New York: Springer.

Fischer, G. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Fischer, G. H. (1987). Applying the principles of specific objectivity and generalizability to the measurement of change. *Psychometrika*, **52**, 565–587.

Fischer, G. H. (1988). Spezifische Objektivität: Eine wissenschaftstheoretische Grundlage des Rasch-Modells. In K. D. Kubinger (Hrsg.), *Moderne Testtheorie.* Weinheim: Psychologie Verlags Union.

Fishburn, P. C. (1973). Binary choice probabilities: On the varieties of stochastic transitivity. *Journal of Mathematical Psychology*, **10**, 327–352.

Fisseni, H.-J. (1990). *Lehrbuch der psychologischen Diagnostik.* Göttingen: Hogrefe.

Gerpott, T.J. (1989). Ökonomische Spurenelemente in der Personalwirtschaftslehre: Ansätze zur Bestimmung ökonomischer Erfolgswirkungen von Personalauswahlverfahren. *Zeitschrift für Betriebswirtschaft*, **59**, 888–912.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests.* New York: Wiley.

Gnedenko, B. W. (1968). *Lehrbuch der Wahrscheinlichkeitsrechnung* (5. Aufl.). Berlin: Akademie-Verlag.

Goldberg, S. (1969). *Die Wahrscheinlichkeit*, 2. Aufl. Braunschweig: Vieweg.

Hamerle, A. (1982). *Latent-Trait-Modelle.* Weinheim: Beltz.

Hamerle, A. & Tutz, G. (1980). Goodness of fit tests for probabilistic measurement models. *Journal of Mathematical Psychology*, **21**, 153–167.

Hogg, R. V. & Craig, A. T. (1978). *Introduction to mathematical statistics.* 4th. ed., London: Collier Macmillan Publishers.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, **96**, 72–98.

Irtel, H. (1987). On specific objectivity as a concept in measurement. In E.E. Roskam und R. Suck (Eds.), *Progress in mathematical psychology* (pp. 35–45). Amsterdam: Elsevier Science Publishers.

Irtel, H. (1993). The uniqueness structure of simple latent trait models. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology.* New York: Springer-Verlag.

Irtel, H. (1995). An extension of the concept of specific objectivity. *Psychometrika*, **60**, 115–118.

Irtel, H. & Schmalhofer, F. (1982). Psychodiagnostik auf Ordinalskalenniveau: Meßtheoretische Grundlagen, Modelltest und Parameterschätzung. *Archiv für Psychologie*, **134**, 197–218.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, **36**, 109–133.

Klauer, K. C. (1990). Asymptotic properties of the ML estimator of the ability parameter when item parameters are known. *Methodika*, **4**, 23–36.

Knoche, N. (1990). *Modelle der empirischen Pädagogik.* Mannheim: BI-Wissenschaftsverlag.

Krantz, D.H., Luce, R.D., Suppes, P. & Tversky, A. (1971). *Foundations of measurement*, Vol. I. New York: Academic Press.

Kubinger, K. (1984). Nutzentheoretische Beurteilung differentialdiagnostischer Entscheidungen. *Diagnostica*, **30**, 249–266.

Lind, D. (1994). *Probabilistische Testmodelle in der empirischen Pädagogik.* Mannheim: BI-Wissenschaftsverlag.

Lindgren, B. W. (1971). *Elements of decision theory.* New York: Macmillan.

Lindgren, B. W. (1976). *Statistical theory* (3rd ed.). New York: Macmillan.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, **1**, 1–27.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* Paris: Mouton.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, **3**, 1–18.

Pfanzagl, J. (1971). *Theory of measurement.* Würzburg: Physica.

Raiffa, H. (1961). Statistical decision theory approach to item selection for dichotomous test and criterion variables. In H. Salomon (Ed.) *Studies*

in item analysis and prediction (pp. 187–220). Stanford: Stanford University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, **14**, 58–94.

Rey, E.-R. (1988). Klinische Diagnostik. In R. S. Jäger (Hrsg.), *Psychologische Diagnostik: ein Lehrbuch* (S. 437–451). München: Psychologie Verlags Union.

Rost, J. (1988). *Quantitative und qualitative probabilistische Testtheorie.* Bern: Huber.

Rost, J. & von Davier, M. (1995). Mixture distribution Rasch Models. In G. Fischer & I. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments and Applications* (S. 257–268). New York: Springer.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C. & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, **64**, 609–626.

Schmidt, F. L. & Hunter, J. E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology*, **68**, 407–414.

Schmidt, F. L., Hunter, J. E., Pearlman, K., Hirsh, H. R., Sackett, P. R., Schmitt, N., Tenopyr, M. L., Kehoe, J. & Zedeck, S. (1985). Forty questions about validity generalization and meta-analysis with commentaries. *Personnel Psychology*, **38**, 697–798.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N. & Trattner, M. H. (1986). The economic impact of job selection methods on size, productivity, and payroll costs of the federal workforce: An empirically based demonstration. *Personnel Psychology*, **39**, 1–29.

Schmidt, F. L., Hunter, J. E. & Raju, N. S. (1988). Validity generalization and situational specifity: A second look at the 75% rule and Fisher's z Transformation. *Journal of Applied Psychology*, **73**, 665–672.

Schuler, H., Frier, D. & Kauffmann, M. (1993). *Personalauswahl im europäischen Vergleich.* Stuttgart: Verlag für angewandte Psychologie.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, **3**, 25–60.

Steyer, R. & Eid, M. (1993). *Messen und Testen.* Heidelberg: Springer-Verlag.

Stumpf, H. (1988). Zweite Testabnahme im besonderen Auswahlverfahren – Teilnehmerstruktur und Ergebnisse von Item- und Testanalysen. In G. Trost (Hrsg.), *Test für medizinische Studiengänge* (S. 2–21). Bonn: Institut für Test- und Begabungsforschung der Studienstiftung des deutschen Volkes.

Tack, W. H. (1976). Diagnostik als Entscheidungshilfe. In K. Pawlik (Hrsg.), *Diagnose der Diagnostik* (S. 103–130). Stuttgart: Klett.

Tack, W. H. (1980). Zur Theorie psychometrischer Verfahren: Formalisierung der Erfassung von Situationsabhängigkeit und Veränderung. *Zeitschrift für Differentielle und Diagnostische Psychologie*, **1**, 87–106.

Taylor, H. C. & Russel, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology*, **23**, 565–578.

Tent, L. & Stelzl, I. (1993). *Pädagogisch-psychologische Diagnostik. Band 1: Theoretische und methodische Grundlagen.* Göttingen: Hogrefe.

Tutz, G. (1986). Bradley-Terry-Luce models with an ordered response. *Journal of Mathematical Psychology.* **30**, 306–316.

Tutz, G. (1989). *Latent Trait-Modelle für ordinale Beobachtungen.* Heidelberg: Springer-Verlag.

Wainer, H. (Ed.) (1990). *Computerized adaptive testing.* Hillsdale, NJ: Lawrence Erlbaum.

Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology.* San Francisco: W. H. Freeman.

Williams, R. H. & Zimmerman, D. W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement*, **37**, 679–689.

Wottawa, H. (1980). *Grundriß der Testtheorie.* München: Juventa Verlag.

Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, **40**, 395–412.

Zimmerman, D. W. (1976). Test theory with minimal assumptions. *Educational and Psychological Measurement*, **36**, 85–96.

Zimmerman, D. W. & Williams, R. H. (1977). The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology*, **16**, 135–152.

Namensverzeichnis

Sachverzeichnis