16. Kapitel

Statistische Entscheidung

Axel Ostmann und Joachim Wutke

1. *Einleitung*

In der empirischen und experimentellen Psychologie werden in der Regel nach der Datenerhebung und der Datenaufbereitung spezielle Verfahren angewandt, um bezüglich der formulierten statistischen Hypothesen zu statistisch gestützten Entscheidungen zu gelangen. Dabei gilt unverändert, was bereits Bredenkamp (1972, S. 9) feststellte: „Der Signifikanztest ist das bei empirisch arbeitenden Psychologen beliebteste und am häufigsten verwendete Verfahren der beurteilenden Statistik." Vielen Anwendern gilt der Signifikanztest, von Bredenkamp (1972, S.159) als „eine Regel, mit der statistische Hypothesen angenommen oder abgelehnt werden" interpretiert, als unproblematisches Standardverfahren. Dabei wird gern übersehen, daß der Signifikanztest nicht das einzige Verfahren zur Unterstützung statistischer Entscheidungen ist, „daß es ‚den' Signifikanztest nicht gibt" (Hager & Westermann, 1983, S.71) und daß jedes statistische Entscheidungsverfahren mit einer Reihe erheblicher Probleme zu rechnen hat. So bemerkt etwa Stegmüller (1973, S. 1): „Es besteht bis zum heutigen Tag eine ungeheure Kluft zwischen logischen und wissenschaftstheoretischen Analysen von Begriffen der Prüfung, der Bestätigung und der Bewährung von Hypothesen auf der einen Seite, und von Fachleuten im Gebiet der mathematischen Statistik angestellten Untersuchungen über diese Themenkreise auf der anderen Seite." Wenngleich wir auf die begrifflichen und wissenschaftstheoretischen Probleme der statistischen Bewährung von Hypothesen nur am Rande eingehen und auf die Arbeiten von Hacking (1965), Lenzen (1974) und Stegmüller (1973) verweisen, wollen wir hier eine Reihe von klassischen Versuchen vorstellen, welche die Kluft zwischen statistischen Entscheidungen und konkreten statistischen Verfahren zu schließen versuchen.

Im Kontext der Einführung statistischer Grundbegriffe (Abschnitt 2) soll auf Probleme hingewiesen werden, die es bezüglich des Zusammenhangs zwischen empirischen Daten und Hypothesen gibt und die sich bezüglich der Begrün-

dung der Wahl eines geeigneten statistischen Modells zur Entscheidungshilfe ergeben. Nach der Vorstellung der ursprünglichen Idee des Signifikanztests von R.A. Fisher (Abschnitt 3) wird das am häufigsten angewandte Entscheidungsverfahren, das Hypothesentesten nach Neyman und Pearson dargestellt (Abschnitt 4). In einigen psychologischen Lehrbüchern zur Statistik wird unter dem Namen „Signifikanztest" ein Verfahren eingeführt, welches man, wie wir später sehen werden, eigentlich als ,,Testen von Nullhypothesen nach Neyman und Pearson" bezeichnen müßte (Abschnitt 5). Wir werden auch dafür eintreten, nicht vom „statistischen Schließen" zu sprechen; mittels eines statistischen Modells schließt man nicht in der Art eines logischen Schlusses, man entscheidet sich. Dieser Tatbestand ist den in Abschnitt 6 vorgestellten Bayesianern allerdings geläufig, denn der sich auf das Bayessche Theorem stützende Ansatz betont insbesondere die Vorläufigkeit aller statistischen Entscheidungen.

Mehrfach bereits gab es „Signifikanztest-Kontroversen" (siehe etwa Morrison und Henkel, 1970; Witte, 1989), und weil es in diesem immerwährenden Streit zwischen Signifikanzlern und Bayesianern oftmals um weltanschauliche Positionen geht, erscheint uns eine Rückbesinnung auf die problemgeschichtliche Ausgangslage der unterschiedlichen statistischen Entscheidungsverfahren wichtiger als die Vorstellung spezieller Techniken.

Aus wissenschaftstheoretischer Sicht ist es das Ziel der experimentellen Forschung in der Psychologie, ,,gesetzesförmige Aussagen von hypothetischem Charakter zu formulieren" (Herrmann 1979, S. 17). Der hypothetische, das heißt vorläufige Charakter der gesetzesförmigen Aussagen in der Psychologie resultiert auch aus der Verwendung statistischer Entscheidungsverfahren, mit denen man zu einer Bewertung der aus den wissenschaftlichen Hypothesen abgeleiteten statistischen Hypothesen kommt. Statistische Hypothesen sind weder verifizierbar noch falsifizierbar, eben deshalb muß man sich - stets nur vorläufig - entscheiden. In der Testtheorie nach Neyman und Pearson gibt es deshalb bei der Entscheidung bezüglich statistischer Hypothesen zwei mögliche Gefahren: Die irrtümliche Annahme einer eigentlich falschen Hypothese und die irrtümliche Verwerfung einer eigentlich richtigen Hypothese - sowohl Annahme wie Verwerfung sind mit der Möglichkeit eines Irrtums behaftet, eine zweifelsfreie Absicherung gegen beide Irrtumsmöglichkeiten gibt es nicht. Für statistische Hypothesen gilt (leider) nicht das von Popper (1966, S. 13) für wissenschaftliche Theorien geforderte Abgrenzungskriterium zur Metaphysik, das Kriterium der Falsifizierbarkeit. „Ein empirisch wissenschaftliches System muß an der Erfahrung scheitern können." Als Gegner der Induktion hielt Popper den Schluß von den durch die Erfahrung bestätigten besonderen Beobachtungsaussagen auf die Theorie für logisch unzulässig. Als Vertreter einer deduktionistischen Methodologie fordern etwa Erdfelder und Bredenkamp (siehe Kapitel 2) eine implikative Beziehung zwischen psychologischer

und statistischer Hypothese; letztere soll aus ersterer deduziert werden können. Da Deduktionen jedoch nicht gehaltserweiternd sind, bleibt die Frage, an welcher Stelle die zum Testen statistischer Hypothesen notwendigen probabilistischen Annahmen eingeführt werden; am konsequentesten wäre wohl die Verwendung probabilistischer Theorien und stochastischer Modelle.

Wer statistische Entscheidungen trifft, sollte sich darüber im klaren sein, daß er sich einer probabilistischen und/oder entscheidungstheoretischen Sicht des Problems der Bewährung von Hypothesen angeschlossen hat. Er muß das Problem wissenschaftlicher Gewißheit bescheidener betrachten, etwa so, wie Lakatos (1974, S. 93) es bezüglich des Übergangs von einem wissenschaftlichen Determinismus Newtonscher oder Kantscher Provenienz zu einem ,,neuzeitlichen Probabilismus“ formuliert hat: ,,Der Probabilismus wurde von einer Reihe von Philosophen entwickelt, die der Ansicht waren, daß wissenschaftliche Theorien, obwohl gleichermaßen unbeweisbar, dennoch verschiedene Grade von Wahrscheinlichkeit in bezug auf die vorhandene Evidenz besäßen. Wissenschaftliche Redlichkeit verlangt demnach weniger, als man gedacht hat: Sie besteht darin, daß man nur hochwahrscheinliche Theorien vorbringt oder, noch bescheidener, daß man für jede wissenschaftliche Theorie die Erfahrungsdaten und die Wahrscheinlichkeit der Theorie im Lichte dieser Daten spezifiziert.“ (Es ist anzumerken, daß der von Lakatos verwendete Begriff der ,,Wahrscheinlichkeit der Theorie“ mit dem mathematischen Wahrscheinlichkeitsbegriff wenig gemein hat). Eine Einführung in die Geschichte des Probabilismus und seine Bedeutung für die moderne Wissenschaft geben Gigerenzer et al. (1991).

Diese Arbeit möchte dem Anwender statistischer Entscheidungsverfahren aufzeigen, welche impliziten und expliziten Annahmen mit jedem konkreten Verfahren verbunden sind, angefangen von Annahmen über den ,,state of the world“, mathematischen Vorannahmen und Annahmen über die Rolle von Hypothesen im Forschungsprozeß. Wir wollen zeigen, daß statistische Entscheidungen kein „triviales Standardritual” am Ende einer Datenerhebung sein sollten. Ob es uns allerdings gelingt, solch hartnäckigen Unsinn aus der Welt zu schaffen wie etwa den, mittels einer ,,Signifikanz“ sei eine Hypothese zu beweisen, ist zweifelhaft - zu viele Versuche vor uns sind bereits gescheitert. Obwohl wir über statistische Entscheidungsverfahren berichten, weisen wir an einigen Stellen darauf hin, daß aus der Konfrontation einer zunächst wissenschaftlichen und dann (ja nicht notwendig) statistischen Hypothese mit den Daten nicht unbedingt die Anwendung eines statistischen Entscheidungsverfahrens resultieren muß; man könnte sich mit anderen, etwa deskriptiven Techniken begnügen. Sorgfältige Analyse klug aggregierter Rohdaten kann mehr Erkenntnisse bringen als ein möglicherweise belangloses signifikantes Ergebnis. Auch muß man sich ja nicht gleich entscheiden; man könnte sich auch zu einer „Urteilsenthaltung (als) adäquateste Reaktion” (Stegmüller 1973,

S. 3) entschließen. Vielleicht lassen eine Vergrößerung der empirischen Indizienbasis oder eine verbesserte Datenerhebung später eine angemessenere Entscheidung zu.

2. Statistische Grundbegriffe

Im Kontext statistischer Entscheidungen, mißverständlicherweise oft auch Infcrenzstatistik genannt, sollen Aussagen getroffen werden, mittels derer Charakteristika der erhobenen empirischen Daten in spezieller Weise verallgemeinert werden sollen. Die - kurz so genannte - Statistik bedient sich dabei eines probabilistischen Kalküls. Sind gewisse methodologische Vorannahmen und wahrscheinlichkeitstheoretische Oberhypothesen bezüglich der Herkunft der Daten gemacht, so kann ein empirisches Datum x als Realisation einer Zufallsvariable X interpretiert werden. **Zufallsvariable** sind in der Regel reellwertige Funktionen auf einem Wahrscheinlichkeitsraum Ω, den wir im folgenden allerdings unspezifiziert lassen wollen (vgl. Steyer, Kap. 15). Der Wahrscheinlichkeitsraum soll uns dazu dienen, das Eintreten eines Datums x mit dem Ziehen eines speziellen Zufalls ω zu unterlegen. Dieser Zufall führt bei der Prozedur, dem Einzelexperiment X im Sinne der Erhebung dieses einen Datums, eben zu $X(\omega) = x$. Der Begriff „Experiment“ wird hier technisch gebraucht und verstanden als „Ziehen eines Datums im Rahmen eines speziellen Zufallsmodells“; die in der empirischen Psychologie übliche Forschungsmethode ist hier nicht gemeint. Der Hinweis, daß mit der Anwendung des Kalküls unterstellt wird, daß Daten durch einen wie auch immer gearteten Zufall generiert sind, ist uns so wichtig, daß wir an vielen Stellen und in den meisten Formeln dieses Abschnittes zur Erinnerung ω mitnotieren. Wie wir später noch begründen werden, hegen wir den Verdacht, daß aus der Unterschlagung dieser zentralen Idee der Zufallsvariablen manche Fehlinterpretation statistischer Aussagen folgen kann. Die Idee, Daten auch als vom Zufall abhängig zu konzipieren, hat etwa zur Konsequenz, auch den Ausgang eines statistischen Tests als vom Zufall abhängig anzunehmen.

Zufallsvariablen werden charakterisiert durch ihre Verteilung, die man als **Verteilungfunktion** F_x mit $F_x(x) = p(X \leq x) = p(\{o;X(o) \leq x\})$ angibt. Wenn möglich, wird die Verteilungsfunktion im stetigen Fall über ihre Dichte und im diskreten Fall über ihre Wahrscheinlichkeitsfunktion angegeben, die jeweils mit f_x bezeichnet wird. In der psychologischen Forschung wird im stetigen Fall oft die Annahme gemacht, ein Datum sei durch eine normalverteilte Zufallsvariable erzeugt, im diskreten Fall nimmt man öfters eine binomialverteilte oder hypergeometrisch verteilte Zufallsvariable an. Innerhalb des probabilistischen Kalküls werden Aussagen darüber, „woher die Daten kommen“, durch eine Spezifikation der Verteilung von X gemacht. Diese Spezifikation

besteht in der Regel in der exakten Angabe einer ganz speziellen Verteilung. Man sagt dann, das Datum x ist aus einer Population mit eben dieser Verteilung gezogen. Zwei Populationen sind gleich, wenn die entsprechenden Zufallsvariablen identisch verteilt sind, man schreibt dann $X \sim Y$.

Im Normalfall interessiert nicht nur ein einzelnes Datum, üblicherweise werden in einer empirischen Untersuchung nacheinander mehrere Daten $x_1,...,x_n$ gewonnen. In der Regel wird unterstellt, daß alle Daten unabhängige Realisierungen derselben Zufallsvariablen X sind; man kann sie als Realisierung einer **Stichprobe** $\mathbf{X = (X_1,...,X_n)}$ ansehen, also ist $\mathbf{X}(\omega) = (X_1(W),...,X_n(O)) = (x_1,...,x_n)$. Dabei ist eine Stichprobe vom Umfang n ein n-Vektor von unabhängigen und identisch verteilten Zufallsvariablen (also $X_i \sim X$, $i \in \{1,...,n\}$). Einen konkreten Datensatz kann man als Vektor $x = (x_1,...,x_n)$ im sogenannten **Stichprobenraum** darstellen.

Ist die Verteilung von X bekannt, so läßt sich die Verteilung einer zugehörigen Stichprobe X errechnen. Normalerweise ist die Ausgangsverteilung, die Verteilung von X, jedoch nicht bekannt. Ist wenigstens bekannt oder wird zumindest als Oberhypothese vorausgesetzt, daß die Verteilung der Zufallsvariablen X Element einer parametrisierten Menge von Verteilungen ist, so spricht man von einem parametrisierten Wahrscheinlichkeitsmodell für X. **Parameter** sind Kennzahlen von Verteilungen. Diese Vorannahme eines parametrisierten Wahrscheinlichkeitsmodells ist zumeist notwendig oder zumindest hilfreich, um trotz Unkenntnis der genauen Verteilung weitere Aussagen treffen zu können. In einem solchen Wahrscheinlichkeitsmodell werden spezielle Verteilungen durch die Angabe eines oder mehrerer Parameter identifiziert, die Verteilung von X also etwa durch einen entsprechenden Parametervektor $\theta = (\theta_i)_{i \in I}$. Eine Menge von Verteilungen heißt n-parametrisch, falls man zur Identifizierung n Parameter benötigt. Die konkrete Wahl der Parameter weist meist eine gewisse Willkür auf oder folgt Konventionen; in der Psychologie wird als typisches Wahrscheinlichkeitsmodell oft angenommen, X sei normalverteilt mit (unbekanntem) Erwartungswert μ und (unbekannter) Varianz σ^2, oder X sei binomialverteilt mit (bekanntem) N und (unbekanntem) p. In beiden Verteilungsfamilien lassen sich auch spezifische Verteilungen durch Angabe zweier Zahlen identifizieren, nur müssen dies eben nicht notwendig Erwartungswert und Varianz sein; zur Identifizierung einer Normalverteilung etwa könnte man auch zwei beliebige, aber verschiedene Quartile oder Centile verwenden, beides Kennwerte, die in der Vergangenheit in der Intelligenzforschung durchaus verwendet wurden.

Bei Vorliegen eines solchen **Wahrscheinlichkeitsmodells** besteht eine naheliegende Standardaufgabe darin, aufgrund von Daten zu beurteilen, welcher spezifische Parametervektor $\theta = (\theta_i)_{i \in I}$ zur Beschreibung von X adäquat ist, zudem sollte spezifiziert werden, mit welchen Unsicherheiten oder möglichen

Fehlern ein solches Urteil verbunden ist. Bei einer solchen Aufgabe, die sich zudem meist nur mit einer kleineren Anzahl von Parametern beschäftigt (in der Psychologie werden häufig nur die ersten zwei Momente, gelegentlich die ersten vier, betrachtet), spricht man von einer Aufgabe der **parametrischen Statistik.**

Die Frage nach einer zu den Daten passenden Beschreibung der Zufallsvariablen X stellt sich natürlich in ähnlicher Weise, wenn kein spezifiziertes Wahrscheinlichkeitsmodell vorliegt, wenn z.B. infolge mangelnder Skalenqualität eine Verteilung aus der Menge gängiger parametrisierter Verteilungen nicht zur Verfügung steht. Hier hätte eine mögliche Parametrisierung die Komplexität der Angabe der gesamten Verteilungsfunktion und nicht nur einzelner Parameter. In diesen Fällen spricht man von einer Aufgabe der **nichtparametrischen Statistik.**

Statistische Hypothesen sind Spezifikationen der Verteilung einer Zufallsvariablen. Sie spezifizieren die Verteilung entweder vollständig oder (meistens) teilweise, etwa durch Angabe der Parameter der Verteilung. Erwartungswert, Varianz, Median oder Interquartilsabstand sind hierfür bekannte Beispiele. In einem parametrischen Wahrscheinlichkeitsmodell ist eine statistische Hypothese deshalb identifizierbar mit einer Teilmenge 0_0 der Menge Θ aller möglichen Parametervektoren $\theta = (\theta_i)_{i \in I}$ des entsprechenden Modells. Eine oft benutzte Annahme ist von der Form: „E(X) = μ und X normalverteilt“. Die Normalverteilungsannahme kann als Wahrscheinlichkeitsmodell oder **Oberhypothese** aufgefaßt werden. Aus der zweiparametrigen Hypothesenmenge Θ mit Erwartungswert und Varianz als Parameter kann die eben formulierte Hypothese mit der Menge 0_0 = {(l.@,),e,~IR+} identifiziert werden. Unter der Annahme der Gültigkeit dieser Hypothese kann man dann auch Aussagen wie etwa

$$p(\omega; X(\omega) \in C) < 0{,}05 \qquad (2.1)$$

erhalten, wobei C eine bestimmte Menge im Stichprobenraum ist. Eine solche Aussage besagt, daß für das gegebene Wahrscheinlichkeitsmodell die Wahrscheinlichkeit dafür, Stichprobendaten aus dem Bereich C zu erhalten, für alle Verteilungen mit dem Erwartungswert μ kleiner als 0,05 ist. Diese Aussage hat eine für die üblichen statistischen Verfahren typische Form, es werden Aussagen über die Wahrscheinlichkeit von Daten gemacht unter der Annahme der Richtigkeit bestimmter Oberhypothesen.

Wir gehen im folgenden davon aus, daß die Wahl des speziellen Wahrscheinlichkeitsmodells begründet und expliziert wird. Dieser Schritt der Wahl einer Oberhypothese wird bedauerlicherweise von vielen Anwendern als vollkommen unproblematisch angesehen, in einer Vielzahl von Fällen wird konven-

tionell auf ein Normalverteilungsmodell zurückgegriffen. Dies kann angemessen sein, muß es aber nicht.

Wir gehen ferner davon aus, daß in geeigneter, auf das gewählte Wahrscheinlichkeitsmodell bezugnehmenderweise Daten $x = (x_1,...,x_n)$ erhoben werden, die sich als Realisation einer Stichprobe der Größe n auffassen lassen. Es ist danach die Frage, wie die erhobenen Daten verwendet werden können, um etwas über die Verteilung der Zufallsvariablen und ihrer Parameter zu erfahren. Schon in der deskriptiven Statistik werden Datensätze x in geeigneter Weise komprimiert und verrechnet, arithmetisches Mittel, Median, Range, (empirische) Varianz und viele andere Kennziffern sind hierfür Beispiele. Zu jeder Kennziffer gehört eine Rechenvorschrift oder Funktion $t = t_n: \mathbb{R}^n \rightarrow \mathbb{R}$, die den erhobenen Daten x eine Kennziffer t(x) zuweist. Mit einer solchen Funktion versucht man, die in den Daten vorhandene Information in geeigneter Weise aufzuschließen. Verknüpft man nun das oben vorgestellte Wahrscheinlichkeitsmodell mit einer solchen Funktion t, so wird der Wert von t abhängig vom Zufall: es entsteht eine sogenannte Stichprobenfunktion oder Statistik $T = T_n$, also

$$T(\omega) = T_n(\omega) = t(X_1(\omega),...,X_n(\omega)) \tag{2.2}$$

Statistiken sind damit ebenfalls Zufallsvariablen und werden wie diese durch ihre Verteilung charakterisiert. Spezifiziert eine Hypothese die Verteilung von X vollständig, so läßt sich auch die Verteilung einer Statistik T_n vollständig bestimmen. Generell gilt, daß Verteilungsspezifikationen für X mit Verteilungsspezifikationen für T_n einhergehen. Ist die Verteilung von X spezifiziert, so können Ausdrucke wie $p(T_n \leq t)$ im Prinzip errechnet werden. Ist eine Hypothese H gegeben, so werden auch Ausdrucke wie $p(T_n \leq t \mid H)$ verwandt; diese Schreibweise wird insbesondere von bayesianisch orientierten Statistikern präferiert. Das Zeichen $\mid$ wird dann nicht wie üblich im strengen Sinne zur Notation einer bedingten Wahrscheinlichkeit verwandt. Eine bedingte Wahrscheinlichkeit wurde ein entsprechendes Wahrscheinlichkeitsmodell voraussetzen, etwa eines, in dem nicht nur X, sondern auch die Hypothese H als Zufallsvariablen vorkommen, und für die zudem eine gemeinsame Verteilung existiert. Wir wollen hier jedoch das Zeichen $\mid$ - etwas ungenau, aber hinreichend präzisierbar - als „unter der Annahme der Gültigkeit beziehungsweise Richtigkeit von“ lesen, wie es die meisten Bayesianer auch tun.

Im Kontext statistischer Entscheidungen werden Statistiken verwendet, um Parameter zu schätzen und Hypothesen zu testen. Dabei wird allerdings oftmals übersehen, daß die Auswahl einer geeigneten Statistik durchaus ein Problem sein kann. Die Statistiklehrbücher machen am Beispiel der Stichprobenvarianz darauf aufmerksam, daß die Übertragung der Rechenvorschrift, mit der man einen Parameter einer Verteilung bestimmt, in eine Rechenvorschrift

für eine Statistik keineswegs ein Verfahren darstellt, das automatisch einen „guten" **Schätzer** erzeugt. Schon Anfänger lernen, daß man bei erwünschter Erwartungstreue in der Formel für die Stichprobenvarianz durch (n-l) dividiert. Die Probleme entstehen dadurch, daß jede Statistik, da sie eine Zufallsvariable ist, ihrer Verteilung nach beurteilt werden muß. Bezüglich der Beantwortung der Frage, was nun eine „gute" Statistik ist, was ein „guter" Schätzer ist und was eine „gute" Testgröße ist, werden üblicherweise eine Reihe von Kriterien herangezogen, von denen die wichtigsten auf R.A. Fisher zurückgehen.

Statistiken sollen der Reduktion von Daten dienen. Eine erste naheliegende Forderung an eine Statistik ist demnach, daß bei dieser Reduktion von Daten in eine Statistik möglichst kein Informationsverlust auftreten soll. Im idealen Fall ließen sich die Daten vollständig aus der Statistik rekonstruieren. Diese Eigenschaft „kein Informationsverlust" wird im Begriff der Suffizienz einer Statistik präzisiert, Fisher hielt dieses Kriterium für zentral. Ist ein parametrisiertes Wahrscheinlichkeitsmodell gegeben, so nennt man eine Statistik T **suffizient,** falls die unter den Werten t der Statistik bedingten Dichten oder Wahrscheinlichkeiten $f_x(x \mid T = t)$ jeweils für alle $\theta \in \Theta$ übereinstimmen. Diese bedingten Größen werden dann bereits durch die Partition des Stichprobenraums in die Bereiche {x;T = t} der Daten gleichen Wertes t der Statistik induziert (vgl. etwa Lindgren 1976, S. 224f.). Deshalb kann man in entsprechender Weise suffiziente Partitionen des Stichprobenraums definieren. Es läßt sich dann zeigen, daß die gesamte aus den Daten verfügbare Information, die zur Identifikation von θ verwendet werden kann, in der Angabe der speziellen Menge der Partition oder - anders ausgedrückt - in der Angabe T = t enthalten ist.

Uns ist es an dieser Stelle wichtig festzustellen, daß durch den übergang vom Datensatz zum Wert einer Statistik die Daten reduziert werden und daß deshalb die Wahl einer geeigneten Statistik begründet werden sollte. Es sollte insbesondere gezeigt werden, daß die gewählte Statistik keinen oder einen geringeren Informationsverlust aufweist als andere, daß die gewählte Statistik also die suffizienteste ist.

Verwendet man die Statistik T als Schätzer für den Parameter θ, so wird diese **Schätzung** bezüglich einer Reihe von Eigenschaften beurteilt. Eine Statistik T bzw. T_n (mit n als Stichprobengröße) heißt:

- **erwartungstreu** oder unverfälscht (unbiased) falls $E(T) = \theta$. Ein bekanntes Beispiel, einen Schätzer erwartungstreu zu machen, ist die Korrektur des Nenners bei der Schätzung der Stichprobenvarianz,
- **asymptotisch erwartungstreu** falls $\lim_n E(T_n) = \theta$,

- **konsistent** falls $\lim_n p(|T_n - \theta| > \varepsilon) = 0$ für alle ε. Die relative Häufigkeit ist ein bekanntes Beispiel für einen konsistenten Schätzer für den Parameter p,
- **effizient,** falls T eine erwartungstreue Statistik ist und unter allen anderen erwartungstreuen die geringste Varianz hat. Schon Anfänger lernen, daß bezüglich des ersten Moments das arithmetische Mittel eine geringere Varianz hat als der Median und demgemäß „effizienter" ist.

Nicht in jedem Wahrscheinlichkeitsmodell stehen für jeden Parameter Schätzer mit allen wünschenswerten Eigenschaften zur Verfügung. So gibt es etwa bei der Familie der hypergeometrischen Verteilungen nicht für jeden Parameter erwartungstreue Schätzer (vgl. etwa Hartung 1985, S.207). In jedem Anwendungsfall sollte überlegt werden, welche Güteeigenschaften eines Schätzers notwendig und wünschenswert sind.

Nun wollte man nicht nur wissen, wie „gut" ein Schätzer ist, man wollte auch wissen, wie „genau" die aus den Daten erhaltene Schätzung eines Parameters ist. Die Präzisierung und Lösung dieser Aufgabe ohne bayesianische Methoden (dort ist der Parameter eine Zufallsvariable) gelang Neyman (1935). Seine Modifikation der Schätzaufgabe besteht darin, mittels zweier Statistiken A und B aus den Daten $(X_1,X_2,...,X_n)(w) = (x_1,x_2,...,x_n)$ Intervallgrenzen $a = A(o)$ und $b = B(\omega)$ zu bestimmen, innerhalb derer der (nichtstochastische, feste aber unbekannte) Parameter θ „wahrscheinlich" liegt; das heißt: Das Paar (A,B) soll die Eigenschaft haben, daß $p(A \leq \theta \leq B) = 1 - \alpha$ mit kleinem α. Die Realisierung [a,b] heißt dann das aufgrund der Daten und der Schätzer A und B bestimmte **Konfidenzintervall** zum vorgegebenen Niveau $1-\alpha$. Das Verfahren zur Bestimmung von Konfidenzintervallen weist große Ähnlichkeit mit dem im nächsten Abschnitt vorgestellten Signifikanztest auf, dem Komplement des Konfidenzintervalls hier entspricht dort der kritische Bereich, in dem H_0 verworfen werden soll, dem Komplement α des (Konfidenz-)Niveaus $1-\alpha$ hier entspricht dort das Signifikanzniveau α; α steht also jedesmal für die maximal tolerierte Irrtumswahrscheinlichkeit.

Verwendet man eine Statistik T zum Testen, so soll eine Entscheidung darüber getroffen werden, welche von mehreren Handlungsmöglichkeiten aufgrund des Wahrscheinlichkeitsmodells und aufgrund der aus den Daten verfügbaren Informationen als adäquat angesehen werden kann. Ein Test besteht (aus vereinfachter heutiger Sicht) aus einer Statistik T und einer Entscheidungsregel $d: \mathbb{R} \rightarrow A$, wobei A die Menge der Handlungsalternativen darstellt. Aus Fisherscher Sicht wurde sich etwa der bekannte t-Test darstellen lassen als t-Verteilung, abhängig von den Freiheitsgraden, und als Entscheidungsregel, unter welchen Bedingungen die Nullhypothese beibehalten wird oder nicht beibehalten („verworfen") wird.

Wird nun T = t realisiert, **so** legt die **Entscheidungsregel d(t)** fest, was ein Entscheider „tun soll", welches Element a aus A er wählen soll. Der Modaloperator „soll" weist darauf hin, daß ein statistischer Entscheider gehalten ist, sich aus Gründen wissenschaftlicher Redlichkeit an die konventionell vereinbarten Entscheidungsregeln zu halten. Die Elemente a von A werden meist als Entscheidung für eine Teilmenge 0_a der Parametermenge Θ des Wahrscheinlichkeitsraumes interpretiert. Da die Teilmenge 0_a eine Hypothese H darstellt, kann man die Handlungsalternative a in diesem Falle mit der Hypothese $H = 0_a$ identifizieren. Die Menge der Handlungsalternativen A kann aber auch weitere, nicht durch Parametermengen identifizierbare Elemente enthalten, etwa ein Element u für die Alternative „keine Entscheidung treffen" oder Elemente für „Modell präzisieren" oder „weitere Daten erheben"; die letzte Alternative ist etwa in der Entscheidungsregel für sequentielle Testverfahren enthalten.

üblicherweise wird unterschieden zwischen einem

- komplementären Alternativtest mit $A = \{H_0,H_1\}$, wobei H_1 als Komplement von H_0 im Parameterraum definiert ist, und einem
- Entscheidbarkeitstest mit $A = \{H_0,u,H_1\}$, wobei H_0 und H_1 sich hier in der Regel nicht zum ganzen Parameterraum ergänzen; deshalb verbleibt hier das Element u für „keine Entscheidung" in der Menge der Handlungsalternativen A.

Hypothesen, die nur aus einem Parameter bestehen, heißen einfache Hypothesen oder Punkthypothesen, Ein gebräuchlicher Anwendungsfall in der Psychologie ist ein komplementärer Alternativtest mit der Punkthypothese $H_0 = (0,)$ (auch „zweiseitiger Test" genannt) oder der unspezifischen Bereichshypothese $H_0 = \{\theta \in \mathbb{R}, \theta \; 18\}$ („einseitiger Test"). Sind alle Hypothesen einfach, so heißt der Test einfach. Man spricht vom „einfachen Alternativtest" für $A = \{H_0,H_1\}$ mit $H_i = \{e_i\}$, i = 0,1, und vom „einfachen Entscheidbarkeitstest" für $A = \{H_0,u,H_1\}$, mit $H_i = \{e_i\}$. Einfache Tests haben für parametrische Wahrscheinlichkeitsmodelle den Charakter der Überprüfung einer **Effektgröße;** diese wird ausgedrückt in der Differenz $8_0 - 8_1$. Wie wir später sehen werden, eröffnet der einfache Entscheidbarkeitstest zusätzlich die Möglichkeit einer Erweiterung zu einem sequentiellen Test.

Oft hat man aus inhaltlichen Erwägungen Grund, eine Hypothese, sie heiße Nullhypothese oder einfach H_0, besonders auszuzeichnen. Der Bereich $K = \{t;d(t) \neq H_0\}$ führt zum Verwerfen von H_0 und heißt **kritischer Bereich;** entsprechend heißt $C = \{x;d(t(x)) \# H_0\}$ kritischer Bereich im Stichprobenraum. Ein Alternativtest (T,d,A) kann damit auch durch (T,K) oder C charakterisiert werden.

Halten wir fest: Ein Test, aufgefaßt als Prozedur des statistischen Entscheidens, ist charakterisiert durch

- die Teststatistik T sowie die Verteilung von T,
- die Menge der Handlungsalternativen A,
- die Entscheidungsregel d.

Im Sinne dieser allgemeinen Formulierung kann man das Schätzen eines Parameters θ auch als Spezialfall des Testens auffassen: A ist im Falle des Schätzens die Menge aller möglicher Parameterwerte bzw. die entsprechende feinste Partition dieser Menge.

Da die Verteilungen vieler Statistiken oftmals nur aufwendig exakt zu berechnen sind, verwenden Anwender und Statistiker seit Beginn der Entwicklung statistischer Entscheidungsverfahren durch Karl Pearson gern in Tabellenform zugängliche Verteilungen oder weichen auf leichter zu handhabende Grenzverteilungen aus, falls diese gute Näherungen darstellen, etwa weil die Stichprobe einen großen Umfang hat. Die χ^2-Verteilung bei Fragen des Goodness-of-fit oder für die Verteilung von Varianzen folgt oft diesem pragmatischen Rationale. Die Entscheidungsregel wird oft so modifiziert, daß bestimmte Grenzverteilungen verwendet werden können, etwa indem man dem Test (T,A,d) eine Statistik hinzufügt, die aus Transformation von T hervorgeht und die eben diese in Tabellen oder in der verwendeten Software verfügbare Grenzverteilung hat. Dazu kommen Überlegungen, wann eine aufgrund der Grenzverteilung getroffene Entscheidung eine hinreichend gute Näherung darstellt. In der Praxis wird diese Frage oft nicht gesondert untersucht; meistens greifen Anwender hierbei auf „Faustregeln“ zurück, die in Lehrbüchern zu finden sind. In der Regel geben solche Faustregeln an, ab welcher Stichprobengröße eine befriedigende Näherung als erreicht betrachtet werden kann.

Wie alle Entscheidungen können auch statistische Entscheidungen falsch sein. Da die Teststatistik eine Zufallsvariable ist, kann man die Wahrscheinlichkeit von Fehlentscheidungen im Prinzip berechnen. Als **a-Fehler** oder Fehler der 1. Art bezeichnet man jene Entscheidung, bei der H_0 nach der Entscheidungsregel des Tests verworfen wird, obwohl H_0 eigentlich zutrifft. Die Wahrscheinlichkeit für dieses Ereignis, oder genauer, die Wahrscheinlichkeit für das Verwerfen von H_0 bei Vorliegen einer Verteilung aus H_0 ist zumindest kleiner gleich

$$\alpha = \sup_{\theta \in H_0} p_\theta(\{\omega; T(\omega) \in K\}) \tag{2.3}$$

Es ist das Verdienst von Neyman und Pearson, in kritischer Abgrenzung zu Fisher darauf aufmerksam gemacht zu haben, daß nicht nur das irrtümliche Verwerfen, sondern auch die irrtümliche Annahme von H_0 einen Fehler darstellt. Dieser Fehler wird als **ß-Fehler** oder Fehler 2.Art bezeichnet. Wissenschaftler sollten sich bemühen, auch für diesen Fehler eine Obergrenze ein-

zuhalten (vergleiche Kapitel 14 dieses Bandes). Ausgehend von dieser Idee zweier möglicher Fehler haben Neyman und Pearson (1933) Gütekriterien vorgeschlagen. Als Hilfsmittel zur Formulierung dieser Kriterien haben sie (für statistische Tests) zunächst eine Gütefunktion (power function) definiert, die wir im folgenden für den Alternativtest vorstellen.

Die **Gütefunktion** π (eigentlich n_n) auf einer parametrisierten Verteilungsfamilie gibt an, wie groß die Wahrscheinlichkeit ist, bei Vorliegen einer mit θ identifizierten Verteilung und der Stichprobengröße n gegen die Hypothese H_0 zu entscheiden; anders ausgedruckt: es wird die Größe der kritischen Region angegeben, gemessen in Wahrscheinlichkeit. Für den Fehler 1. Art gilt dann

$$\alpha = \sup_{\theta \in H_0} \pi(\theta) \tag{2.4}$$

Das Vorliegen eines ß-Fehlers oder Fehlers 2.Art hat dann höchstens die Wahrscheinlichkeit

$$\beta = \max_{\theta \in H_1} 1 - \pi(\theta) \tag{2.5}$$

Verwendet man einen komplementären Alternativtest, so ist bei Stetigkeit von π, die im allgemeinen vorausgesetzt werden kann, ß = 1 - α. Die Größen α und ß ergänzen sich hier jeweils zu 1; wird der eine Fehler kleiner, so wird der andere Fehler größer. Wer beide Fehler kontrollieren möchte, sollte entweder keine komplementären Alternativtests benutzen oder müßte sich mit hohen Fehlerwahrscheinlichkeiten zufrieden geben, nur wurde eine Befundsicherung etwa mit α = ß = 0,50 wahrscheinlich wenig Akzeptanz erlangen.

Mittels der Gütefunktion $\pi = r_n$ lassen sich für einen Test folgende **Giiteeigenschaften** definieren; ein Test heißt

- **zum Niveau α,** wenn $\pi(\theta) \leq \alpha$ für alle $\theta \in H_0$ gilt,
- **unverfälscht** (unbiased), wenn $\pi(\theta) \geq \alpha$ für alle $\theta \in H_1$ gilt,
- **konsistent,** wenn $\lim K_n(\theta) = 1$ für alle $\theta \in H_1$ gilt.

Bei der Definition eines Tests werden alle Hypothesen gleichbehandelt, die Zuordnung der Namen H_0 und H_1 ist eigentlich willkürlich, wenngleich sich in der Psychologie gewisse Konventionen durchgesetzt haben; so wird als H_0 oft jene Punkthypothese bezeichnet, die einen ,,Nulleffekt“ postuliert, gelegentlich wird jene Hypothese, die man gern zurückweisen (to nullify) wurde, als H_0 bezeichnet. Leider wird die folgenreiche Entscheidung, welche Hypothese als H_0 betrachtet werden soll, nur selten begründet (siehe etwa Neyman 1950, S. 262f.). Nach Festlegung der Hypothesen wird bezüglich der Eigenschaft ,,zum Niveau α“ die Obergrenze für einen tolerierbaren Fehler der

1. Art festgelegt. Wenn, wie in der Praxis meistens üblich, „zum Niveau cr“ klein gewählt wird, ist der kritische Bereich klein und die Wahrscheinlichkeit für die Entscheidung zugunsten von H_0 groß, weil $1 - \alpha$ groß ist. Viele Anwender bewerten die Kontrolle des Fehlers 2.Art als nachrangig oder sogar völlig Verzichtbar, andere hingegen betonen ihre Wichtigkeit. So fordert etwa Bredenkamp (1980, S.22): **„α,** ß und die Effektgröße müssen vor dem Experiment festgelegt worden sein, so daß auch der benötigte Stichprobenumfang ... geschätzt werden kann.“ Dieser Forderung wird in der Praxis nur selten entsprochen. Ein möglicher Grund für den Verzicht auf die Kontrolle des ß-Fehlers kann allerdings auch sein, daß das irrtümliche Festhalten an H_0 im Forschungskontext als unerheblich bewertet wird: Man hat etwas übersehen, aber das macht nichts.

Üblicherweise wird die Obergrenze des a-Fehlers, das sogenannte **Signifikanzniveau,** niedrig festgelegt. Wenn der maximale Fehler 1. Art festgelegt ist, soll nun auch der ß-Fehler kontrolliert werden, d.h. man strebt an, den ß-Fehler so klein wie möglich zu halten. Dies geschieht durch Auswahl eines sogenannten **besten Tests.** Tests mit selben Niveau α lassen sich nämlich bezüglich ihrer Güte vergleichen. Es sei $H_0 = \{e_0\}$. Die Tests u und v haben die Gütefunktion r_u bzw. x_v. Dann heißt Test u gleichmäßig besser als Test v, falls $n_u(e) \geq r_v@)$ für alle $\theta \in H_1$. Wenn H_1 eine Punkthypothese ist, so gibt das Neyman-Pearson-Lemma Bedingungen für die Existenz eines besten Tests zum vorgegebenen Niveau an (vgl. Lindgren 1976, S.299-302). Dem Anwender begegnet das Problem des besten Tests zum Beispiel, wenn entschieden werden soll, ob zur statistischen Sicherung von Mittelwertsunterschieden der t-Test oder der U-Test angewandt werden soll.

Die Güteeigenschaft der Unverfälschtheit eines Tests verlangt, daß der α-Fehler keinesfalls größer wird als der ß-Fehler. Diese Eigenschaft hat zum Beispiel der einseitige Gaußtest, sie ist aber keineswegs trivial, wie wir weiter unten (Seite 34) an einem Beispiel von Neyman und Pearson sehen können, die dort im Hypothesenraum statt des Mittelwertes (wie beim Gaußtest) die Varianz variieren. Auch der zweiseitige Kolmogoroff-Smirnov-Test ist verfälscht (siehe etwa Büning & Trenkler, 1978, S. 90). Die Summe α + ß beider Fehler ist bei Unverfälschtheit kleiner gleich 1. Der beste denkbare Test wurde natürlich beide Fehler ganz vermeiden, wurde also gleichzeitig $\alpha = 0$ und ß = 0 erfüllen. Dieser ideale Test existiert natürlich in der Praxis nicht, jedoch kann man fordern, daß ein Test sich mit Wachsendern Stichprobenumfang diesem Idealfall stetig annähert. Diese Eigenschaft ist die obengenannte Konsistenz.

Anhand zweier Probleme aus der Geschichte der Mathematik, die mit den Namen Arbuthnot und Bayes verbunden sind, soll ein grundlegender Unterschied zwischen den bayesianischen und den nicht-bayesianischen Zugängen zur Statistik aufgezeigt werden. Entscheidend ist dabei die unterschiedliche

Richtung des statistischen Denkens: Einmal wird dabei ein Bezug von den Hypothesen zu den Daten hergestellt (gelegentlich wird dies „statistischer Schluß“ genannt), das andere Mal wird ein Rückbezug von den Daten zu den Hypothesen vorgenommen (gelegentlich „statistischer Rückschluß“ genannt). Bei der Darstellung beider Probleme wird eine Zufallsvariable X betrachtet, die nur zweier Werte fähig ist: 0 und 1. Mit p_0 sei $p(X = 0)$ bezeichnet.

Arbuthnot stellte sich 1711 die Aufgabe zu zeigen, daß die Wahrscheinlichkeit einer Knabengeburt nicht 1/2 ist. Er will das Ergebnis seiner Untersuchung für ideologische Zwecke nutzen, nämlich für einen Gottesbeweis; aber das soll hier nicht interessieren. In unserer Schreibweise sei $X_i(\omega) = 0$, falls die i-te Beobachtung eine Knabengeburt ist, und $X_i(\omega) = 1$, falls sie eine Mädchengeburt ist. Arbuthnot stellt die Hypothese $p_0 = 1/2$ auf, allerdings nur um zu zeigen, daß diese als extrem unwahrscheinlich abzulehnen ist. Als Datensatz benutzt Arbuthnot Geburtsregister, die sich über einen Zeitraum von 82 Jahren erstrecken. Als Statistik wird die Anzahl der Jahre verwendet, in dem innerhalb dieses Zeitraums ein Knabenüberschuß vorliegt. Nun wurden in allen 82 Jahren mehr Knaben als Mädchen geboren, und die Statistik T nimmt einen solch extremen Wert an, daß Arbuthnot seine Vermutung gestützt glaubt, eben daß die Wahrscheinlichkeit einer Knabengeburt nicht gleich 1/2 ist.

Drei Aspekte erscheinen aus heutiger Sicht an diesem Lösungsansatz unbefriedigend. Zum einen scheint die Auswahl der Statistik abhängig gewesen zu sein vom gewünschten Resultat. Zum zweiten liefert Arbuthnot keine von den Daten unabhängige Begründung für seine Verwerfungsregel. Und schließlich unterzieht er nicht die Hypothese einer strengen Prüfung, an die er eigentlich glaubt, sondern er prüft die Hypothese, die er gern verwerfen möchte. Eine ausführliche Darstellung dieses frühen statistischen Entscheidungsversuchs findet sich bei Hacking (1966) sowie bei Freudenthal und Steiner (1966, S. 180).

Das zweite historische Beispiel geht zurück auf Bayes und ist posthum 1764 veröffentlicht worden. Bayes hat sein Problem weitgehend befriedigend gelöst. Hier beschreibt X den Ausgang eines Wurfexperiments. $X_i() = 0$ bedeute hier, daß der i-te Wurf eines Gegenstandes in den linken Teil eines zweigeteilten Zielfeldes fällt, und $X_i() = 1$ bedeute, daß der i-te Wurf in den rechten Teil des Zieles fallt. Das Ziel kann man sich als Rechteck vorstellen, die Teilung als vertikale Linie, die das Rechteck in zwei Teile schneidet. Nun sei jedoch unbekannt, wo diese Teilung liegt, wo also das linke Feld aufhört und wo das rechte Feld anfängt. Bayes nimmt nun an, daß die Teilung des Zielfeldes in „links” und „rechts“ ebenfalls durch ein Zufallsexperiment zustande genommen ist, formal also durch eine Zufallsvariable Y. Wenn das der Fall ist und wir $p_0 = Y()$ setzen, betrachten wir auch den Parameter p_0 der Ver-

teilung der Zufallsvariablen X als Realisierung einer Zufallsvariablen, nämlich Y. Es ist nun nach dem **Bayesschen Theorem** möglich, aus den Daten $(x_1,...,x_n)$ (den Würfen) und der Verteilung von Y (bestimmt durch die Grenze zwischen links und rechts) auf die Verteilung von X zu schließen. Es ist nämlich mit $x = (x_1...,x_n)$

$$f_Y(p_0) \mid X = \frac{f_Y(p_0) f_{\mathbf{X}}(\mathbf{x} \mid p_0)}{\int f_Y(\theta) f_{\mathbf{X}}(\mathbf{x} \mid \theta) d\theta} \tag{2.6}$$

Da $Y = p_0$ als Hypothese über X aufgefaßt werden kann, läßt sich diese Formel auch wie folgt verkürzt schreiben (- für proportional):

$$p(\text{Hypothese} \mid \text{Daten}) \sim p(\text{Hypothese}) p(\text{Daten} \mid \text{Hypothese})$$

oder nochmals verkürzt als

$$p(H \mid D) \sim p(H) p(D \mid H) \tag{2.7}$$

Der Ausdruck p(H | D) ist bei gegebenen Daten x eine bis auf einen konstanten Faktor gegebene Funktion, die als **Likelihoodfunktion** bezeichnet wird. Für verschiedene Realisierungen der Stichprobe ist p(H | D) eine mehrdimensionale Statistik, denn sie weist nach Normierung jedem Zufall ω eine Funktion zu.

Zum Zwecke statistischer Entscheidungen kann man die Formel des Bayesschen Theorems wie folgt interpretieren: die Verteilung von Y, angebbar durch die Werte der Wahrscheinlichkeitsdichte für Hypothesen der Form $Y = p_0$, wird berücksichtigt durch eine zunächst beliebig angenommene Wahrscheinlichkeit p(H). Der Ausdruck p(D | H) steht für die Wahrscheinlichkeit der Daten unter der Annahme der Gültigkeit der Hypothese und p(H | D) ist die nach dem Vorliegen der Daten erhaltene ,,Rückschluß"-Wahrscheinlichkeit, die aufgrund der Daten revidierte Wahrscheinlichkeit der Hypothese.

Damit haben wir die beiden statistischen ,,Schließ"-Richtungen vorliegen: Während Arbuthnot die Wahrscheinlichkeit der Daten unter seiner Hypothese p(D | H) betrachtet hat (und wegen ihrer Unwahrscheinlichkeit auch seine Hypothese verworfen hat), untersucht Bayes die Wahrscheinlichkeit der Hypothese unter gegebenen Daten, p(H | D). In diesem Konzept wird die Wahrscheinlichkeit als Stützmaß für die Gültigkeit von Hypothesen interpretiert. Zumindest auf den ersten Blick unbefriedigend am Bayesschen Ansatz ist, daß die volle Spezifikation der Verteilung von Y verlangt wird, was zudem voraussetzt, daß man die Verteilung von X als durch einen Zufallsprozeß zustandegekommen aufzufassen bereit ist. In der Folgezeit wurde vor allem die oft verwendete Annahme der Gleichwahrscheinlichkeit für Y problematisiert. Dies war einer der Grunde, warum der Bayessche Ansatz für lange Zeit in Mißkredit geriet, denn wenn der Wertebereich von Y die volle Zahlengerade

(kontinuierlich und unbeschränkt) ist, so ist wegen mangelnder Integrierbarkeit in klassischer Weise eine Gleichwahrscheinlichkeit überhaupt nicht definierbar.

3. *Der Signifikanztest nach Fisher*

Auf den englischen Statistiker R.A. Fisher gehen bedeutende Neuentwicklungen der mathematischen Statistik zurück wie etwa die eben kurz skizzierte Schätztheorie für Parameter oder die Varianzanalyse. Für uns von besonderem Interesse ist der **Signifikanztest,** dessen Rationale bei heutigen statistischen Entscheidungen häufig Verwendung findet. Am Beispiel der Lösung eines sehr alten praktischen Problems kann man aufzeigen, daß wesentliche Momente des statistischen Entscheidens auch schon vor Fishers Präzisierung des Signifikanztests intuitiv plausibel waren. Stigler (1977) beschreibt die Prozedur „the trial of the pyx", mittels derer seit dem 12. Jahrhundert in der königlich britischen Münze täglich nach einem Zufallsprinzip einige Münzen aus der laufenden Serie entnommen und in einer besonderen Büchse (pyx) gesammelt wurden. In einem Vertrag zwischen der Krone und der Münze waren Toleranzen für das mittlere Gewicht einer Münze vereinbart. Stigler vergleicht diese Toleranzen mit der uns heute vertrauten 2σ-Regel. Bei Kontrollen in unregelmäßigen Abständen wurde der Inhalt der Büchse gewogen, da nun aber die zulässige Toleranz für eine einzelne Münze mit n, der Anzahl der Münzen in der Büchse, und nicht etwa mit III;- multipliziert wurde, soll es in knapp 800 Jahren nur zu einer einzigen Strafe wegen Unterschreitung des vereinbarten Gewichtsgekommen sein.

Als erste moderne statistische Testprozedur, die von praktischen Anwendungsfällen unabhängig konzipiert war, gilt der von Karl Pearson im Jahr 1900 entwickelte **„goodness-of-fit-test".** Pearson betrachtete, anders als seine Vorgänger, nicht nur einzelne Abweichungen zwischen einer theoretisch erwarteten und einer empirisch beobachteten Häufigkeitsverteilung, auch nahm er nicht die Summe der Beträge der Abweichungen, vielmehr schlug er die Teststatistik

$$\chi^2_{\text{beob}} = S\,\frac{(m' - m)^2}{m'} \tag{3.1}$$

als Maß für die Beurteilung der Abweichung beider Verteilungen vor (hier in der Originalschreibweise von Pearson mit m' als theoretischer und m als beobachteter Häufigkeit in einem Intervall und S als Summation über alle Intervalle; vgl. Kendall & Stuart, 1973, S.436). Sowohl die Berechnung der χ^2-Verteilung als auch die Tabellierung der Wahrscheinlichkeit von $p(\chi^2 < \chi^2_{\text{beob}})$ für bis zu 20 Intervalle und bis zu 15 Werten von χ^2 durch Pearson waren

förderlich für die rasche Verbreitung dieses Verfahrens, denn die Arbeit der Anwender wurde durch diese Hilfen ungemein erleichtert. Die Testprozedur ging von der Annahme aus, daß bei einem „guten Fit“ die Unterschiede zwischen der theoretischen Verteilung - die hier gleichsam den Status einer Nullhypothese annimmt - und den empirisch erhobenen Daten gering ist, daß also $X2_{beob}$ klein ist und demgemäß p groß. Wenn nunmehr χ^2_{beob} sehr groß ist und deshalb p sehr klein, so ist unter der Annahme der Angemessenheit der theoretischen Verteilung ein sehr seltenes Ereignis eingetreten. Aber ist nicht auch ein extrem kleiner Wert von χ^2_{beob} im Rahmen der üblichen Meßfehlertheorie kaum zu erklären oder ebenfalls ein sehr seltenes Ereignis? Schon Pearson hat den Gedanken verfolgt, daß auch „zu gute“ Daten Zweifel erlauben.

In seinem Vorläufer eines Testmodells hat Pearson jedoch weder so etwas wie ein festes Signifikanzniveau noch eine klare Entscheidungsregel vorgegeben, die Entscheidung, ob Empirie und Theorie hinreichend gut übereinstimmen, wurde in das Ermessen des einzelnen Forschers gelegt. Allerdings wollte Pearson die Entscheidung erleichtern durch die Angabe von „odds“, von Wettquotienten gegen den Zufall. Wenn etwa aus der Konfrontation beider Häufigkeitsverteilungen ein p = 0,043 resultierte, so wurde gesagt, die „odds“ dafür, daß die Daten von der theoretischen Verteilung abweichen, sind 23:1.

Die entscheidende Weiterentwicklung dieser Gedanken erfolgte durch Fisher in den zwanziger Jahren (siehe Fisher 1951, 1959) mit dem Signifikanztest. Sein Verfahren wird in Abgrenzung zur Testtheorie von Neyman und Pearson gelegentlich auch „Nullhypothesen-Testen” genannt (letztgenannter Egon S. Pearson war ein Sohn von Karl Pearson; die Beschäftigung mit Grundfragen des statistischen Entscheidens war also Familientradition). Das Rationale des Vorschlags von Fisher läßt sich gut demonstrieren an einem Beispiel, am „problem of the lady tasting tea“ (Fisher, 1951, S. 11; Neyman, 1950, S. 272-94; siehe auch Hagen & Seifert, 1979, S. 190), gleichzeitig lassen sich einige kritische Punkte aufzeigen.

Die Geschichte geht so: Eine ungenannte Lady behauptet, sie hätte ein spezielles geschmackliches Unterscheidungsvermögen, sie könne nämlich sicher herausschmecken, ob in eine Tasse

(a) zuerst die Milch und dann der Tee oder
(b) zuerst der Tee und dann die Milch

eingefüllt worden ist. Fisher, hier in der Rolle eines skeptischen Geschmacksphysiologen, bezweifelt dieses Vermögen und formuliert die folgende Nullhypothese $H_,$: die Lady besitzt das von ihr behauptete Unterscheidungsvermögen nicht. Diese Nullhypothese H_0 gilt es zu testen. Hierzu schlägt Fisher eine spezielle Versuchsanordnung vor: es werden jeweils vier Tassen nach bei-

den Modalitäten (a) und (b) hergestellt; alle 8 Tassen werden dann der Lady in zufälliger Reihenfolge zur Geschmacksprobe vorgesetzt. Die Lady, über diese Versuchsanordnung informiert, soll nun die 8 Tassen in zwei Teilmengen zu je vier Tassen aufteilen, wobei die eine Teilmenge nur solche nach (a) hergestellte und die andere nur solche nach (b) hergestellte Tassen enthalten darf.

Bezüglich der Frage, bei welchem empirischen Ergebnis nun ein skeptischer Wissenschaftler seine Nullhypothese verwirft (und dann vielleicht der Lady glaubt), schlägt Fisher seinen Signifikanztest vor. Dies in einer Art und Weise, die durchaus der Position von Popper in der Beschreibung von Lakatos (1974, S. 90) vergleichbar ist: „Die intellektuelle Redlichkeit besteht nicht darin, daß man versucht, seine Position fest zu verankern oder sie durch Beweis (oder ‚wahrscheinlich machen') zu begründen - die intellektuelle Redlichkeit besteht vielmehr darin, daß man jene Bedingungen genau festlegt, unter denen man gewillt ist, die eigene Position aufzugeben." Fisher setzt das in die folgenden Forderungen um:

(1) Man stelle (zusätzlich zu Vorannahmen, die als bewährt gelten können, nur) eine einzige Hypothese auf, bei deren Verwerfung in der Regel bisheriges Wissen zur Revision anstünde: die **Nullhypothese** H_0.
(2) Man bestimme die unter der Annahme der Gültigkeit von H_0 für diese spezielle Versuchsanordnung resultierende Zufallsvariable X und ihre Verteilung. In unserem Beispiel resultiert für die beiden Modalitäten zur Einordnung der Teetassen in (a) und (b) eine Zufallsvariable X mit den Ausprägungen x = 0,1,2,3 bzw. 4 richtigen Zuordnungen. Wir bezeichnen die Werte x der Zufallsvariablen X „Anzahl richtiger Zuordnungen" im folgenden als „Treffer" sowie die Wahrscheinlichkeiten von X = x mit p_x. Die Wahrscheinlichkeitsverteilung unter der Annahme der Gültigkeit von H_0 ist eine hypergeometrische, die im folgenden tabelliert ist.

Tabelle 3.1: Lady's tea dilemma, 1.Version

Treffer	0	1	2	3	4
p_x	$\frac{1}{70}$	$\frac{16}{70}$	$\frac{36}{70}$	$\frac{16}{70}$	$\frac{1}{70}$

(3) Man lege einen **kritischen Bereich** fest, der diejenigen Realisationen der Zufallsvariablen enthält, die nicht mit der Nullhypothese kompatibel erscheinen. Hier etwa könnte jemand vorschlagen, H_0 sei zu verwerfen, wenn die Lady vier richtige Zuordnungen je Modalität getroffen hat. Nach Zerlegung des Ereignisraums in zwei disjunkte Teilmengen ergäbe sich dann folgende Verwerfungsregel:
- Verwerfe H_0, wenn die Lady jeweils vier richtige Zuordnungen getroffen hat.

Nur wenn die Lady keinen Fehler macht, wird die Nullhypothese verworfen. Dann ist unter der Annahme der Gültigkeit von H_0 ein extremes Ereignis eingetreten, welches eine bedeutsame Abweichung („significant discrepancy“) von H_0 anzeigt. Nur unter der Bedingung, daß sich von ihm kontrolliert und unter seinen Augen etwas extrem Unwahrscheinliches ereignet hat, ist der skeptische Wissenschaftler bereit, seine Hypothese aufzugeben. Mit den ironischen Worten von Neyman (1952, S.43) könnte man es auch so ausdrucken: Wenn H_0 zutrifft, so ist das, was der Wissenschaftler im Falle der Abweichung beobachtet, ein Wunder. Da wir heutzutage aber nicht mehr an Wunder glauben, entschließen wir uns, lieber nicht mehr an die Nullhypothese zu glauben. Doch zurück zum Beispiel; selbst bei Anwendung obiger Verwerfungsregel wäre sich ein skeptischer Wissenschaftler nicht sicher, ob die Verwerfung von H_0 zweifelsfrei richtig wäre, wenn die Lady jeweils vier Treffer gehabt hätte. Denn dies Ereignis ist unter der Annahme der Gültigkeit von H_0 nicht unmöglich, mit einer Wahrscheinlichkeit von $p = 1/70 = 0{,}0143$ ist es eben nur sehr unwahrscheinlich.

(4) Man wähle ein **Signifikanzniveau.** Fisher hat keinen eindeutigen und rational begründeten Vorschlag gemacht, wie genau der durch die Verwerfungsregel festgelegte kritische Bereich bestimmt werden soll, allerdings hat er vorgeschlagen, fairerweise vor einer Untersuchung festzulegen, welche Wahrscheinlichkeit eines Fehlers maximal toleriert werden kann. Diese Wahrscheinlichkeit bestimmt die Größe des kritischen Bereichs, und legt man diesen an den Rand der Zufallsverteilung, ist damit eine Regel zur Verwerfung der H_0 eindeutig formulierbar. Da Wissenschaftler bezüglich der Verwerfung ihrer Hypothese (von deren Richtigkeit sie eigentlich überzeugt waren) skeptischer sein sollten als Menschen im Alltag und sie deshalb ihre Nullhypothese länger beibehalten sollten (Wissenschaftler sind eben nicht so leicht zu überzeugen), schlägt Fisher als übliches Signifikanzniveau 5 % vor (1951, S. 13), in heutiger Schreibweise $\alpha \leq 0{,}05$. Wie Wendt (1966) in Untersuchungen zeigen konnte, haben Menschen im Alltag in vielen Situationen ein „subjektives Verläßlichkeitsniveau“ um 0,20. In der Sprache englischer Wettbüros sind wir im Alltag oft geneigt, uns schon dann für die Verwerfung einer gehegten Vermutung zu entscheiden, wenn die „odds“ 1:4 gegen einen Irrtum stehen, der Wissenschafter entscheidet sich erst dann gegen seine H_0, wenn es 1:19 gegen einen Irrtum steht. Gelegentlich wird dieser Sachverhalt übertragen in das Sprachspiel, ein Wissenschaftler möchte sich höchstens einmal bei 20 Entscheidungen irren, aber diese Aussage ist genaugenommen falsch. In der Psychologie haben sich heute als Konventionen die Signifikanzniveaus $\alpha \leq 0{,}05$, $\alpha \leq 0{,}01$ und $\alpha \leq 0{,}001$ eingebürgert.

(5) Man verwerfe die Nullhypothese genau dann, wenn ein extremes Ereignis eine bedeutsame Abweichung von H_0 anzeigt und wenn die Wahrschein-

lichkeit, die **Verwerfung von H_0** irrtümlich vorzunehmen, höchstens so groß ist wie das Signifikanzniveau.

Im Falle des „Lady's tea dilemma" wurde die Wahl eines Niveaus $\alpha \leq 0{,}05$ bedeuten, daß der Lady im Versuch kein einziger Zuordnungsfehler unterlaufen dürfte. Man könnte nun aber argumentieren, daß die Lady, selbst wenn sie das von ihr behauptete Vermögen hätte, ja auch zufällig, gleichsam „aus Versehen", im Experiment einen Fehler machen könnte. Die Entscheidungsregel des Wissenschaftlers läßt Fehler zu, da dürfte sich die Lady durchaus auch einmal irren. Wollte nun ein großzügiger Wissenschaftler der Lady einen „Zufallsfehler" konzidieren, so hätte er zwei Handlungsmöglichkeiten. Zum einen könnte er seine Nullhypothese bereits verwerfen, wenn die Lady mindestens 3 richtige Zuordnungen trifft, er müßte in diesem Fall aber bereit sein, einen Fehler von $p = 17/70 \approx 0{,}243$ zu tolerieren. Zum zweiten könnte er einen Vorschlag von Fisher aufgreifen und seinen Versuchsplan sensitivieren, indem im Experiment die Anzahl der Geschmacksproben auf 12, je zweimal 6, erhöht wird. Wie zuvor (vgl. Tab.3.1) ist die Wahrscheinlichkeitverteilung eine hypergeometrische, jedoch mit anderen Parametern. Die folgende Tabelle zeigt die entsprechenden „Trefferwahrscheinlichkeiten".

Tabelle 3.2: Lady's tea dilemma, 2.Version

Treffer	0	1	2	3	4	5	6
p_x	$\frac{1}{924}$	$\frac{36}{924}$	$\frac{225}{924}$	$\frac{400}{924}$	$\frac{225}{924}$	$\frac{36}{924}$	$\frac{1}{924}$

In diesem von Fisher als „more sensitive" charakterisierten Versuchsplan können quantitativ kleinere Abweichungen von der Nullhypothese entdeckt werden, und die Verwerfungsregel könnte in diesem Fall lauten:

- Verwerfe H_0, wenn die Lady fünf oder sechs richtige Zuordnungen trifft.

Die Wahrscheinlichkeit dafür, die Nullhypothese irrtümlich zu verwerfen, ist dann $p_5 + p_6 = 37/924 \approx 0{,}04 < \alpha \leq 0{,}05$. Bei Wahl des üblichen Signifikanzniveaus könnte man jetzt der Lady einen „Fehler" konzidieren und dennoch H_0 verwerfen.

Insgesamt erscheint Fishers Vorschlag zur Bewertung der Nullhypothese - einmal abgesehen von einer gewissen Willkürlichkeit bei der Wahl des Signifikanzniveaus und damit der Verwerfungsregel - schlüssig und sowohl dem skeptischen Wissenschaftler als auch der Lady gerecht zu werden. Wie aber ein Einwand von Neyman (1950, S. 272) zeigt, wirft dieses eigentlich triviale Beispiel an unvermuteter Stelle Probleme auf. Denn wenn man die experimentelle Versuchsanordnung zur Überprüfung der Fähigkeit der Lady etwas

modifiziert und ihr nach Zufall Teeproben serviert, die jeweils mit p(a) = p(b) = 0,50 hergestellt wurden, so können die einzelnen Teeproben als Bernoulli-Versuche aufgefaßt werden, und das dem Versuch zugrunde liegende Zufallsmodell ist jetzt ein Urnenmodell „mit Zurücklegen". Statt einer hypergeometrischen Verteilung resultiert nun eine Binomialverteilung als Zufallsverteilung unter H_0. In diesem Fall wurden in der ersten Versuchsanordnung von Fisher vier richtige Zuordnungen der Lady zu einer Irrtumswahrscheinlichkeit von p = 1/16 = 0,0625 führen, falls man aufgrund des empirischen Ergebnisses die Nullhypothesen verwerfen wurde (bei der hypergeometrischen Verteilung war p = 0,0143). Oder anders herum gesagt: Bei vorgegebenen Signifikanzniveau von $\alpha \leq 0{,}05$ wurden hier vier richtige Zuordnungen nicht zur Verwerfung von H_0 führen.

Wie schon Morrison und Henkel (1970) aufgezeigt haben, finden sich in den üblichen Statistiklehrbüchern wenig Hinweise auf diese komplexen Zusammenhänge zwischen experimenteller Versuchsplanung, resultierenden Zufallsverteilungen und den letztendlichen statistischen Entscheidungen. Manche Autoren scheinen der Meinung zu sein, einzig die Wahl des Signifikanzniveaus könne eine Entscheidung bezüglich einer statistischen Hypothese beeinflussen.

Nun sollte am Beispiel von „lady's tea dilemma" nur das Grundprinzip der Signifikanztestung demonstriert werden; in die Praxis der Statistik eingegangen sind natürlich Fishers Neuentwicklungen wie etwa der F-Test, der z-Test oder der „exact-probability-test" sowie seine Weiterentwicklungen des t-Tests oder des χ^2-Tests. Dabei hat seine Bevorzugung von effizienten, erwartungstreuen und möglichst suffizienten Schätzern und seine Präferenz für aus der Normalverteilung abgeleiteten Zufallsverteilungen bis heute Wirkung in der Anwendung statistischer Entscheidungsverfahren. Allerdings hat der Signifikanztest nach Fisher nicht nur von den Bayesianern heftige Kritik erfahren, auch Neyman und Pearson waren überhaupt nicht mit dem Vorgehen von Fisher einverstanden. Von Neyman ist überliefert, daß er den Signifikanztest für „schlechter als nutzlos" hielt. Vor allem zwei Punkte sollen erwähnt werden:

(a) Kritik am Nullhypothesen-Testen: Fisher kennt nur eine Hypothese, demgemäß kennt er nur einen möglichen Fehler, den a-Fehler, und er spricht nur davon, ab wann die Hypothese als widerlegt („disproved") gilt. Aus den in der Einleitung genannten Gründen wurden wir nicht mehr von einer Widerlegung, sondern nur noch von einer Verwerfung dieser Hypothese sprechen. Die meisten statistischen Anwender testen heute (mindestens) zwei konkurrierende Hypothesen simultan und wiirden Stegmüller zustimmen in seiner Behauptung, „isolierte statistische Hypothesen können niemals einer adäquaten theoretischen Beurteilung

unterzogen werden. Wir werden versuchen, die These zu begründen, daß die Beurteilung einer statistischen Hypothese nur erfolgen kann in bezug auf eine Klasse von Alternativhypothesen, die mit der zur Diskussion stehenden Hypothese konkurrieren.“ (1973, S. 23) Wenngleich diese Kritik zutrifft, so muß man Fisher dennoch zugute halten, daß für ihn, der seine Statistik oft zur Suche neuer Anbaumethoden in der Landwirtschaft einsetzte, die Testung seiner Nullhypothese im Forschungsprozeß einen ganz speziellen Stellenwert besaß. Fisher hat Nullhypothesen formuliert, die aus dem aktuellen Wissensstand begründet waren. Im Falle der Lady hätte die Verwerfung von H_0 chemische oder physiologische Forschungen ausgelöst, denn ein solches Unterscheidungsvermögen wäre mit bekannten Theorien nicht vereinbar gewesen. Fisher hatte nach der Verwerfung der Nullhypothese keinen „statistisch gesicherten Effekt”, sondern eine interessante inhaltliche Fragestellung, der er sich hätte widmen müssen. Anwender, die heute übliche Signifikanztests nicht zur Effektsicherung, sondern zur zufallskritischen Absicherung benutzen, handeln häufig in diesem Sinne. Hier wird der Test als „screening“ in einem frühen Stadium der Theoriebildung eingesetzt um abzuschätzen, was einer genaueren Untersuchung lohnt.

(b) Defizite der rationalen Begründung: Fisher konzentrierte sich bei der Auswahl der von ihm betrachteten Parameter auf die besten Schätzer, etwa auf das arithmetische Mittel und die Varianz, nur ist nicht zu begründen, daß in jedem Anwendungsfall die besten Schätzer auch die besten Tester sind. Es sind durchaus Fälle denkbar, in denen statt des arithmetischen Mittels der Median oder der Modus oder statt der Varianz der Range oder eine Centildifferenz inhaltlich geeignetere Kennwerte wären. Unbestreitbar sind die Gütekriterien, die Fisher in seiner Schätztheorie entwickelt und benutzt hat, einleuchtend und überzeugend; die Wahl der Parameter und der Prüfstatistiken ist dadurch jedoch nicht eindeutig bestimmt. Zweifelsohne war die Auswahl leicht zu berechnender Parameter und die Tabellierung von Prüfverteilungen sehr „benutzerfreundlich“, nur birgt dies eben auch die Gefahr einer schematischen Anwendung. Auch wenn das vorgegebene Verfahren für den selbst zu entscheidenden Fall gar nicht so gut paßt, etwa wegen anderer Oberhypothesen oder wegen Fragestellungen, für die die gängigen Parameter gar nicht optimal sind, wird das Standardverfahren gerechnet, etwa angemessenere **nicht-parametrische Verfahren** werden meist nicht erwogen. Sowohl die Höhe des Signifikanzniveaus als auch die Plazierung des daraus resultierenden Verwerfungsbereichs an den Rand als auch die hälftige Aufteilung dieses Bereichs bei zweiseitiger Fragestellung sind letztlich konventionell, wenngleich nicht ohne Plausibilität. Obwohl Fisher selbst versucht hat, mit seiner „fiducial probability“ ein quantitatives Maß zur Stützung von Hypothesen zu entwickeln (siehe Hacking 1965, S. 133; Stegmüller 1973,

S.258), so bleibt festzuhalten, daß nach der Durchführung eines Signifikanztests nur ein **klassifikatorischer Stützungsbegriff** Anwendung finden kann: Die Entscheidungen können nur lauten „die H_0 wird verworfen“ oder „die H_0 wird nicht verworfen“.

4. Die Testtheorie von Neyman und Pearson

Mit den Namen Neyman und Pearson wird gerne eine dem Fisherschen Signifikanztest überlegene und dem Bayesianischen Denken entgegengesetzte Theorie vom statistischen Entscheiden verbunden. Diese Sicht ist zwar nicht ganz falsch, jedoch zumindest recht unvollständig. Neyman und Pearson nehmen in ihren Arbeiten oftmals explizit die Ansätze von Fisher auf, waren allerdings stets um mathematische Präzisierung und um Anreicherungen bemüht - wobei sie des öfteren in Gegensatz zu Fisher gerieten. Vor allem Fishers Schätztheorie und seine und anderer Arbeiten zu Stichprobenverteilungen und ihrer Momente war Anreiz zu weiterführenden Arbeiten. Mitte der zwanziger Jahre war erreicht, was Karl Pearson einst programmatisch verlangt hatte, die Statistik war nicht mehr auf normalverteilte oder binomialverteilte Populationen beschränkt. Aber mit der Ausarbeitung vieler neuer statistischer Techniken entstand das Problem, Kriterien dafür zu finden, welches Verfahren bei einer Fragestellung am angemessensten ist. Das ist deshalb keine nur akademische Frage für Theoretiker, weil unterschiedliche Techniken offensichtlich zu verschiedenen Ergebnissen und damit zu unterschiedlichen Entscheidungen führten - und diese Frage geht auch Praktiker an. Fisher etwa vertrat den Standpunkt, die Likelihood beim Schätzen zu verwenden, aber weder die Definition der Likelihood noch ihre theoretischen Eigenschaften noch darauf aufbauende Prinzipien wie „maximum likelihood“ wurden von ihm klar dargelegt. Während E. A. Pearson als Reaktion daraufhin vorschlug, Bruche zwischen Likelihoods zu bilden und das „likelihood-ratio“-Kriterium für Entscheidungen zu benutzen, war für Neyman Fishers Lob der Likelihood inkonsequent, weil Fisher gleichzeitig die „Methoden der inversen Wahrscheinlichkeit“ (die Bayesianischen Methoden der „rückschließenden“ Statistik) vehement ablehnte. Neyman und Pearson betrachteten auch nach dem Entwurf ihrer eigenen Testtheorie im Jahre 1926 immer wieder alternative Zugänge zum statistischen Entscheidungsproblem und versuchten, sie integrativ zu verarbeiten. Ihre Maxime „keep doors open“ galt nicht zuletzt auch dem Bayesianismus (Neyman & Pearson, 1928, 1933b).

Im folgenden werden in etwas modernerer Fassung die wesentlichen Aspekte der Neyman-Pearson-Testtheorie vorgestellt, wie sie hauptsächlich zwischen 1928 und 1933 entstand. Eine neuere und lesenswerte Darstellung findet sich bei Neyman (1950).

In ihrer Testtheorie präzisieren Neyman und Pearson (1926) den Begriff der **statistischen Hypothese** in einer Weise, die der statistisch-formalen Bearbeitung zugänglich ist. Sie unterscheiden zwischen einfachen Hypothesen (also $H = \{\theta\}$) und zusammengesetzten Hypothesen und verlangen statt einer einzelnen (Fisherschen Null-) Hypothese die Angabe eines Alternativenraumes Θ. Damit wird die Prüfung einer einzigen (denkbaren) Hypothese - in einer wohlbestimmten Theorie können ja nicht verschiedenen Hypothesen bezüglich des gleichen Sachverhalts gleichzeitig Gültigkeit beanspruchen - ersetzt durch eine Entscheidung zwischen alternativen Möglichkeiten; diese Entscheidung allerdings kann falsch sein. Konsequenterweise wird der Begriff des statistischen Schließens ersetzt zugunsten des Begriffs der statistischen Entscheidung. Die „Signifikanz" der Daten im Sinne ihrer Unwahrscheinlichkeit unter der bisher für zutreffend gehaltenen Hypothese ist für Neyman und Pearson kein Entscheidungskriterium. Erstens ist bestenfalls nur eine der beiden Irrturnsmöglichkeiten (nämlich der Fehler 1. Art) über die Wahrscheinlichkeit ihres Auftretens kontrolliert, zweitens gibt es kein eindeutiges Kriterium, welcher Bereich mit Wahrscheinlichkeit zur Verwerfung von H_0 führt, und drittens sind wahrscheinlichkeitstheoretische Aussagen über die „Signifikanz" der Daten weder möglich noch sinnvoll, denn seltene Ereignisse geschehen ja auch - wie etwa Lottogewinne oder Erdbeben. Begründete Aussagen sind nur über die Güte der Entscheidungsregel möglich. Deshalb haben Neyman und Pearson die Begriffe des a-Fehlers, des ß-Fehlers sowie die **Teststärkefunktion** π präzisiert und definiert. Darauf aufbauend haben sie Gütekriterien für Tests zur Verfügung gestellt. Leider finden in der psychologischen Literatur Überlegungen bezüglich der **Teststärke** unverändert geringe Beachtung. So berichten Sedlmeier und Gigerenzer (1989), daß in nur 2 von 64 betrachteten Untersuchungen die Teststärke überhaupt erwähnt wurde, bestimmt wurde sie nie. Ferner berichten sie, daß sich seit der klassischen Arbeit von Cohen (1962) nur wenig verändert hat. Interpretiert man die Teststärke im Sinne der Wahrscheinlichkeit, ein signifikantes Ergebnis zu erhalten, wenn wirklich ein Effekt vorhanden ist, so ist die Teststärke für mittelgroße Effekte nach Schätzungen von Sedlmeier und Gigerenzer im Median von 0,46 im Jahre 1960 (Cohen) auf 0,37 im Jahre 1984 abgesackt.

Um eine Hypothese $\theta_0 \in \Theta$ zu testen, formulieren Neyman und Pearson das Problem wie folgt um: Es gilt, im Stichprobenraum Konturen $\Phi(x)$ = konstant zu finden (mögliche Kandidaten sind etwa Iso-Likelihoodflächen), so daß

- 1. die Wahrscheinlichkeit unter $H_0 = 8_0$, daß die Stichprobe $\Sigma = X$ innerhalb dieser Konturen liegt, gleich α ist, und
- 2. diese Wahrscheinlichkeit unter jedem anderen $\theta \in \Theta$ maximal ist.

Die hiermit angestrebte Eigenschaft ist die **Unverfälschtheit** des Tests, die zusammen mit Φ ein Kriterium für die Wahl der kritischen Region im Stich-

Probenraum liefert. So schreibt etwa Neyman in einem Brief an Pearson (Pearson 1966, S. 18), ,,using such contours and rejecting H_0 when Σ is inside Φ = const, we are sure that a true hypothesis is rejected with a frequency less than α, and that if H_0 is false and the true hypothesis is, say H', then most often the observed sample will be inside Φ = const and hence H_0 will be rejected".

In fast allen damals untersuchten Fällen konnten die Konturen der Likelihoodfunktion zur Lösung der obigen Aufgabe benutzt werden. Aber schon für die Hypothese H_0 = N(0;1), das heißt für die Hypothese, daß die Daten aus einer normalverteilten Population mit Mittelwert 0 und Varianz 1 stammen, berichtet Neyman (Pearson 1966, S. 18), daß sich die Likelihoodkonturen nicht verwenden lassen, denn danach wurde N(0;1,1) öfter verworfen als H,. Es war also zu klären, unter welchen Bedingungen sich die Likelihoodfunktion zur Konstruktion eines Tests verwenden läßt.

Für den einfachen Alternativtest $\Theta = \{f_0,8_1\}$ ist sie nach dem Neyman-Pearson-Lemma sogar für die Konstruktion eines besten Tests verwendbar. In unserer Notation lautet das **Neyman-Pearson-Lemma** wie folgt: Der Test mit der kritischen Region $C_k = (x; f(x|e_0) = kf(x|e_1)\}$ ist bester Test, wobei $k \geq 0$ so gewählt werden muß, daß das entsprechende erreicht wird. Oft läßt sich dann die kritische Region auch für eine andere, leichter berechenbare Statistik angeben, deren Verwendung man den Vorzug gibt.

Testet man eine einfache Hypothese gegen eine zusammengesetzte, so muß es einen uniform besten Test, also einen Test, der gegen jede einfache Alternative bester Test ist, nicht geben. Das gilt erst recht, wenn nicht nur H,, sondern auch H_0 zusammengesetzt ist. In Analogie zu der oben genannten kritischen Region kann man Tests mit der kritischen Region

$$C_k = \{x; \sup_{\theta_0 \in H_0} f(x|\theta_0) = k \sup_{\theta_1 \in H_1} f(x|\theta_1)\} \tag{4.1}$$

bilden (deren Eigenschaften jedoch im Einzelfall zu prüfen sind). Da $L(\theta)$: = $f(x|\theta)$ oder in anderer Notation $L(\theta|x)$ die schon genannte Likelihoodfunktion darstellt, kann der Wert k als ein Maß der Angemessenheit alternativer Hypothesen aufgefaßt werden. Ein solcher Test heißt Likelihood-Quotienten-Test. Wir werden ihm in Abwandlung weiter unten wieder begegnen (S. 37).

Bei Neyman und Pearson bleibt die Bestimmung eines adäquaten α-Wertes offen: Dieser kann nicht aus mathematischen Kriterien abgeleitet werden, sondern muß aus forschungslogischer und substanzwissenschaftlicher Sicht begründet werden. Nicht die Statistiker, die Fachwissenschaftler sind bezüglich dieser Frage in die Pflicht genommen. Manche Kritiker sind der Meinung,

daß eine solche Begründung für die Obergrenze eines a-Fehlers bei Fragestellungen der Grundlagenforschung oft schwerfallen wurde, sie lehnen zumindest eine unrelativierte Verwendung von Neyman-Pearson-Tests ab. Andere plädieren für eine strenge Anwendung von Neyman-Pearson-Tests, so etwa in der Psychologie Bredenkamp (1972), der als Verfahren einen um die Angabe einer Effektgröße δ modifizierten Entscheidbarkeitstest vorschlägt:

(1) Der Experimentator spezifiziere als Abweichung von H_0 einen Effekt, d.h. er legt eine für den Forschungszweck „praktisch bedeutsame" Differenz $\delta = 8_1 - 8_0$ fest; damit wird auch der Parameterbereich für eine Indifferentzone festgelegt.
(2) Neben der Kontrolle des a-Fehlers erachtet Bredenkamp die Kontrolle des β-Fehlers für ebenso wichtig. Deshalb wähle man ein kleines β; wenn in der Population ein Effekt in der postulierten Größe vorliegt, wird man ihn mit Wahrscheinlichkeit (I-ß) entdecken können; kleinere Abweichungen sind praktisch wenig bedeutsam.
(3) Danach ergibt sich ein weiterer Vorteil; aus dem postulierten Effekt δ und dem maximal tolerierten Fehler 2.Art kann der notwendige Mindestumfang der zu untersuchenden Stichprobe hergeleitet werden und damit der Aufwand für die Untersuchung abgeschätzt werden.

Natürlich ist die exakte Festlegung von ß und δ im gewissen Sinne willkürlich und die meisten psychologischen Theorien sind für solche genauen Effektprognosen nicht spezifiziert genug. Bredenkamp bemerkt aber zu recht, daß auch eine Nichtfestlegung eine implizite Festlegung ist, daß eine explizite Festlegung Kritik und Wiederholbarkeit erleichtert.

Als Integrationsversuch im Sinne von Neyman und Pearson schlägt Witte (1980, 1989, 1991) ein vierstufiges, streng hierarchisches Prüfverfahren vor, bei dem die jeweils nächste Stufe nur ausgeführt werden sollte, wenn die Prüfung bis dahin erfolgreich verlaufen ist:

(1) Zur „Prüfung der Datenerhebungssituation" schlägt Witte vor, zwei konkurrierende einfache Hypothesen 0_0 und 8_1 aufzustellen (es können auch präzise Intervalle sein, so Witte 1991), sowie „Fehlertoleranzen" ~1 = ß zu wählen (beides etwa $\leq$ 0,05). Mittels eines geeigneten Neyman-Pearson-Tests (etwa dem t-Test bei Mittelwertshypothesen) bestimme man die minimale Stichprobengröße n, bei der beide Fehlertoleranzen eingehalten werden können. Scheitert diese erste Prüfung, etwa am zu kleinen n der eigenen Stichprobe, so sollte nicht weiter geprüft werden, „dann sollte jedoch der beobachtete Effekt publiziert werden".
(2) Im zweiten Schritt der „Hypothesenprüfung" wird die Bildung eines Likelihood-Quotienten vorgeschlagen. Wenn etwa

$$\frac{L(\theta_1 \mid x)}{L(\theta_2 \mid x)} \geq \Phi = \frac{1-\beta}{\alpha} \tag{4.2}$$

so wurde 8_1 vorläufig akzeptiert. Kommt man hier zu keiner Entscheidung, „so ist die Theorie in den gewählten empirischen Bereichen nicht zur Erklärung oder Prognose heranzuziehen“ (Witte 1991).

(3) Bei der „Hypothesenqualifikation“ soll geprüft werden, ob sich die Akzeptierung von 8_1 nicht allein darauf stützt, daß $L(\theta_0 \mid x)$ so klein ist. $L(w, \mid x)$ wird deshalb in Relation gesetzt zur maximalen

$$\frac{L(\theta_1 \mid x)}{\max L(\theta_i \mid x)} \geq Q_c = 1 - \sqrt{(1-\beta)\alpha} \tag{4.3}$$

„Die Idee ist hierbei nicht nur, die Stützung der Hypothese [hier e_1] von einer bestimmten anderen abhängig zu machen [hier ($_0], sondern von der empirisch plausibelsten unter Berücksichtigung von Fehlerschwankungen [hier max $L(\theta_i \mid x)$]“ (Witte 1989) (eckige Klammern hinzugefügt).

(4) Zur letzten Prüfung der „Effektqualifikation“ wird der beobachtete Effekt in der Stichprobe daraufhin beurteilt, „ob er auf Datenebene (nicht im Wahrscheinlichkeitsmodell) von Bedeutung ist: $\eta_{\zeta}^{-} \geqslant 0{,}10$ mag eine denkbare Konvention sein“ (Witte 1989). Scheitert Prüfschritt 3, soll man weiter über die Parameter nachdenken, scheitert Schritt 4, ist die Theorie eventuell zu global formuliert.

Wittes Vorschlag zur Integration mehrerer statistischer Entscheidungsmodelle in ein einziges Verfahren ist von Diepgen (1991) kritisiert worden. Die Verwendung der Testtheorie von Neyman und Pearson nicht als Entscheidungsverfahren, sondern „nur“ zur Bestimmung des notwendigen Stichprobenumfangs hält er für unplausibel, die Deutung der (hypothetischen) Fehlerwahrscheinlichkeiten als „Fehlertoleranzen“ ist ihm suspekt. Ein gewichtiges Argument bringt Diepgen gegen Wittes Vorschlag, in Schritt 2 beim Likelihood-Quotienten eine Grenzgröße Ψ einzuführen und diese auch noch an α und ß zu koppeln. Diepgen vermutet hier wohl nicht zu Unrecht, daß die letztendliche Willkürlichkeit der Festsetzung einer Obergrenze für α und ß in der Neyman-Pearson-Testtheorie hier ersetzt wird durch die Willkürlichkeit der Festsetzung von Ψ, obwohl doch in der Likelihood-Stützungs-Philosophie „hypothetische Wahrscheinlichkeiten für das Treffen richtiger oder falscher Entscheidungen . . . überhaupt nicht der Maßstab sind”. (Diepgen 1991). Willkürlich ist natürlich auch die Festsetzung von Q_c in Schritt 3. Bezüglich des letzten Schrittes zur Effektqualifikation wird bemängelt, daß hier nur Stichprobendaten betrachtet werden, obwohl es in einer Inferenzstrategie nur um Aussagen über Effektstärken in der Population gehen sollte.

Für übliche Anwender erscheint der vierstufige Vorschlag von Witte, der alles enthält, was in der Statistik „gut und teuer" ist, als sehr ansprüchlich und aufwendig. Zudem stimmen wir Diepgen in seiner Beurteilung zu, daß dieses Verfahren die Subjektivität des Forschers nur anders einfließen als als die anderen Modelle, bei Witte müßten immerhin 5 Größen, α, β, Ψ, Q_c und η_c^2, letztlich konventionell oder nur plausibel festgesetzt werden.

Nicht nur für diesen Fall sondern bei statistischen Entscheidungen generell mögen aber nun sowohl die Kosten für die Datenerhebung als auch die Kosten für mögliche Fehlentscheidungen von Interesse sein.

Bei einem Test, der bei der Regelung des Hochschulzugangs helfen soll, kann es sein, daß neben relativ geringen Kosten für die Annahme eigentlich ungeeigneter Bewerber die Maxime tritt, möglichst keinen geeigneten Bewerber abzulehnen. Unter diesen Voraussetzungen wäre man hier bei einer statistischen Entscheidung bereit, eine relativ hohe Wahrscheinlichkeit für den Zugang eigentlich ungeeigneter Bewerber zu akzeptieren. Bei einem Test, der bezüglich der Auswahl für einen Pilotenlehrgang helfen soll, sieht die Interessenlage natürlich ganz anders aus: Hier gilt es nach Möglichkeit zu vermeiden, auch nur einen eigentlich ungeeigneten Bewerber aufzunehmen.

Von Wald (1950) wurde die Theorie von Neyman und Pearson erweitert um die explizite Berücksichtigung solcher Kosten. Schon Neyman und Pearson haben mit dem Entscheidbarkeitstest versucht, Fehlentscheidungen mittels Einführung von Indifferenzzonen zu vermindern. Wald geht nun diesen Weg konsequent weiter und entwickelt allgemeinere Entscheidungsregeln, die auch die Größe der Stichprobe beachten. In diesen sequentiellen Tests werden dann solange Daten erhoben, bis anhand von Daten und Entscheidungsregel zu vertretbaren Experimentalkosten die Kosten möglicher Fehlentscheidungen vertretbar klein sind. Das Risiko $r(d,\theta)$ der Verwendung der Entscheidungsregel bei Vorliegen der durch θ charakterisierten Verteilung ist die Summe des dann erwarteten Verlustes bei Verwendung der Entscheidungsregel d und den zu erwartenden Experimentalkosten. Wald (1950) untersucht insbesondere zwei Familien von Entscheidungsregeln, die er als Bayesianische Regel und als Minimaxregel vorstellt. Bei der bayesianischen Regel ist es das Ziel, das bezüglich einer gegebenen Apriori-Verteilung q erwartete Risiko durch die Wahl von d_b zu minimieren, also

$$E_q r(d_b,.) = \min_d E_q r(d,.) \tag{4.4}$$

Die Minimax-Regel d_{mm} rechnet mit dem ungünstigsten Fall, minimiert also das maximale Risiko, also

$$\max_\theta r(d_{mm},\theta) = \min_d \max_\theta r(d,\theta) \tag{4.5}$$

Die Minimaxregel ist eine Bayesianische Regel für die ungünstigste Apriori-Verteilung; hat man keine Informationen über diese Verteilung oder ist sie unbestimmbar, kann diese Regel in diesem Falle als rational gelten. Diepgen (1987) bemerkt zu den Vorteilen des Ansatzes von Wald: „Die Sequentialstatistik (oder sequentielle Testung) - ähnlich wie die Bayes-Statistik - macht es möglich, auch während des Experiments für die Entscheidungsfindung zu lernen. Als Folge davon . . . ist die Stichprobengröße z.T. über 50 % geringer als die entsprechenden notwendigen Stichprobengrößen bei konventionellen Signifikanztests." Sequentielle Testverfahren haben unter Psychologen nur wenige Anhänger gefunden, vielleicht deshalb, weil Kosten, egal ob für die Datenerhebung oder für mögliche Fehlentscheidungen, in der Grundlagenforschung nur selten diskutiert werden. Diepgen äußert sich hier sehr deutlich: „Besonders ärgerlich, geradezu schon mysteriös, erscheint mir die weitgehende Ignoranz der deutschen Psychologie gegenüber Sequentialstatistik, obwohl diese Statistik zentrale Probleme des herkömmlichen Signifikanztests löst, nämlich die Problematik der ß-Fehler-Kontrolle und der praktischen Bedeutsamkeit, und dies auch noch mit dem ökonomischen Vorteil wesentlich kleinerer Stichproben."

5. Eine Variante des Signifikanztests in der Psychologie

Es gibt einen Typ von Signifikanztestung in der psychologischen Literatur, dem wir wegen der relativen Häufigkeit der Verwendung einen eigenen Abschnitt widmen. Man könnte diese „hybrid theory" (Gigerenzer et al. 1991, S. 106) ironisch als „Nullhypothesentestung nach Neyman und Pearson" oder als „Signifikanztest nach Fisher mit Alibi-H," nennen. Gigerenzer (1986) beschreibt und deutet das Vorgehen wie folgt: „In den anspruchsvolleren Lehrbüchern geht dabei Neyman und Pearsons Theorie, meist anonym, als das ‚Überich' der psychologischen Forschung ein. Hier wird das Ziehen von Zufallsstichproben aus einer definierten Population betont, es wird gelehrt, das Signifikanzniveau und den Stichprobenumfang so festzulegen, daß die erwünschte Macht (l-ß) des Tests erreicht wird, und es werden keine Wahrscheinlichkeitsaussagen über den Bestätigungsgrad von Hypothesen oder über Einzelergebnisse erlaubt. Das Fishersche „Ego" aber bestimmt weitgehend, wie in der realen Forschung vorgegangen wird - wenn auch manchmal mit Schuldgefühlen, die „Regeln" verletzt zu haben. Dort werden kaum Zufallsstichproben gezogen, selten Punktalternativen spezifiziert, um den Fehler 2. Art bestimmbar zu machen, und es werden sogar Aussagen über die Wahrscheinlichkeit getroffen, mit der eine Nullhypothese widerlegt sei." Und vielleicht, um zu ergänzen, spielt jene üble Forschungspraxis, bei der nur „Signifikanzen" zählen, hier die Rolle des „ES", des Antriebes. Das von Gigerenzer zutreffend charakterisierte Verfahren sieht dann etwa so aus:

(1) Es wird nur eine H_0 als Punkthypothese spezifiziert. Beide falschen Etymologien des Begriffs Nullhypothese kommen hier zusammen: Meist wird damit ein „Null-Effekt" behauptet, eigentlich soll sie verworfen, nullifiziert werden.
(2) Die Alternativhypothese H_1 wird als Rest des Hypothesenraumes nicht näher spezifiziert, man erwartet „irgendeinen Effekt". Gigerenzer bemerkt richtig, „eine Vorgehensweise, der Neyman und Pearson nie zugestimmt haben".
(3) Gelegentlich wird ein Signifikanzniveau vorher festgelegt, meistens werden „Signifikanzen" hinterher, z.B. durch ** angezeigt. Hager und Westermann (1982) stellten fest, daß nur in 7 von 76 Untersuchungen das Signifikanzniveau vorher festgelegt war.
(4) Überlegungen zur Kontrolle des ß-Fehlers, zur Größe des erwarteten Effekts, zur erwarteten Richtung, zur Stichprobengröße oder zur Teststärke werden nicht berichtet und wohl auch nicht angestellt.

Eine ähnliche Auflistung dieser üblichen Anwendung gibt Wottawa (1990). Er betont zudem: „Das prinzipielle Problem dieses Vorgehens wird deutlich, wenn man sich vor Augen hält, daß (unter den üblichen Stetigkeitsannahmen) die H_0 unabhängig von allen empirischen Befunden für die jeweilige Population nicht gültig ist, und zwar als Folge ihrer punktförmigen Formulierung."

Bei vielen Anwendern beliebt ist auch die Aufwertung einer statistischen Entscheidung in den Status eines logischen Schlusses. So schreibt etwa Wendt (1983, S. 485): „Logisch gesehen folgt die Neyman-Pearsonsche Entscheidungstheorie dem Schlußprinzip des **Modus tollens** ..." Im Sinne dieser „Logik" sollen aus H_0 gewisse Stichprobendaten „folgen" („aus H_0 folgt D_0", so Wendt); wenn sich diese Daten aber nicht ergeben („nun aber nicht D_0"), so darf man als Konklusion schließen, daß H_0 nicht gilt („also nicht H_0"); man darf dann sogar schließen, daß nunmehr H_0 gilt („da aber nicht H_0 folgt H_1" - dieser letzte „Schluß" gehört übrigens nicht mehr zum modus tollens). Nun muß aber leider festgestellt werden, daß die Interpretation einer statistischen Entscheidungsregel als Syllogismus aus mehreren Gründen unsinnig ist: Daten „folgen" nicht aus Hypothesen, sondern sie lassen sich unter der Annahme der Gültigkeit auch der Oberhypothesen mathematisch ableiten. Daten bekommen nicht den Wahrheitswert „falsch", sondern sie sind höchstens im Lichte einer Hypothese sehr unwahrscheinlich; wir „schließen" nicht auf „nicht H_0" und dann auf H_1 - wir entscheiden uns nach vorher festgelegten Regeln. Bei manchen Autoren drängt sich der Verdacht auf, daß sie bezüglich ihrer - möglicherweise falschen - statistischen Entscheidungen durch Unterlegung einer solchen Schluß-„Logik" Unfehlbarkeit suggerieren wollen.

Ob das eben beschriebene „Gegen-He-Testen" (mit anschließendem Bericht der „signifikanten Effekte") überhaupt forschungslogisch sinnvoll ist, kann in

den meisten Fällen bezweifelt werden. Zum einen: „auch ganz geringe Abweichungen können zu einem statistisch signifikanten Ergebnis führen, wenn nur die Zahl n der Beobachtungen genügend groß ist“ (Hager und Westermann 1982). Zum anderen: „Das Finden einer statistischen Signifikanz ist das vielleicht unwichtigste Attribut eines guten Experiments; sie reicht nie hin, um behaupten zu können, daß eine Theorie sich in brauchbarer Weise bewährt hat, daß eine sinnvolle empirische Tatsache festgestellt worden ist oder daß ein Experimentalbericht veröffentlicht werden sollte.“ (Lykken 1968)

Auf ein methodologisches Paradox speziell dieser Art von Hypothesenüberprüfung hat bereits Meehl (1967) hingewiesen: Während in der Physik durch verbesserte experimentelle Anordnungen und Erhöhung der Meßgenauigkeit die bislang als bewährt geltenden Theorien unter Druck geraten und modifiziert oder ersetzt werden müssen, passiert in der Psychologie, die den Signifikanztest solcherart einsetzt, eher das Gegenteil: Die Theorien zeigen die Tendenz, sich zu immunisieren. Sie stehen bei dieser Art Nullhypothesentestung ja auch eigentlich nie ernstlich zur Disposition. Lakatos (1974, S. 170, Fußnote) nimmt diesen Sachverhalt zum Anlaß, polemisch zu fragen, „ob die Funktion von statistischen Techniken in den Sozialwissenschaften nicht vor allem darin besteht, daß sie einen Mechanismus liefern, der Scheinbestätigung und den Anschein von ‚wissenschaftlichem Fortschritt’ an Stellen produziert, wo sich in Wirklichkeit nur pseudointellektueller Mist anhäuft“.

Die Diskussion um das Nullhypothesentesten ist alt und seit langem heftig, schon Roozeboom (1960) konnte feststellen, daß die Methode wegen ihrer Unangemessenheit nachdrücklich kritisiert worden ist - nur, warum hat sich nichts geändert? Eine Reihe von möglichen Gründen zählt Gigerenzer (1986) auf, er vermutet komplizierte Wechselbeziehungen zwischen der Forschungspraxis und dieser Art von Signifikanztest, denn es zeigt sich, daß „der Forscher weniger für ein sorgfältig angelegtes experimentelles Design verstärkt wird als für den Exorzismus von Nullhypothesen“ und daß „die verbreitete falsche Interpretation von ‚signifikant‘ als ‚replizierbar‘ dazu verführt, Experimente eher nicht zu wiederholen“. Nicht ganz schuldlos an dieser Verwechslung von „signifikant“ mit „replizierbar“ oder mit „damit konnte gezeigt werden“ oder gar mit „nunmehr ist bewiesen“ dürfte die fatale Metapher vom modus tollens sein. Aber nicht nur die Forschungspraxis, auch die Ausbildungspraxis scheint am Beharrungsvermögen dieses Verfahrens beteiligt zu sein, denn in den meisten Statistiklehrbüchern wird eine kritische Diskussion praktisch nicht erwähnt (Leiser 1982), statistische Entscheidungsverfahren werden als im Grunde unproblematisch angesehen, meistens werden sie in Grundkursen im ersten Semester angeboten. Oft wird hier auch der Eindruck erweckt, die angewandten Verfahren selbst wurden entscheiden (insbesondere bei EDV-gestützten Verrechnungen, etwa durch SPSS), die Rolle des für seine Entscheidung verantwortlichen Forschers bleibt unerwähnt.

Wie bereits gesagt tendieren viele Anwender dazu, erst einmal abzuwarten, „welche von der Fülle getesteter Haupteffekte, Interaktionen oder Korrelationen signifikant werden“ (Gigerenzer 1986) - „Erklärungen“ für die Signifikanzen werden posthoc gesucht. Wenn ein Jäger mit einem Schrotgewehr auf einen Schwarm Vögel zielt und uns hinterher erzählt, er habe genau die heruntergefallenen auch treffen wollen, so wäre dies Jägerlatein.

Viele dieser Anwender neigen auch „zur üblichen Fehlinterpretation des Signifikanztests“, nämlich dazu, „p als Maß für Signifikanz zu nehmen” (Bakan, 1966). Wie schon Witte (1980, S.53) feststellte, sagt das Signifikanzniveau nichts über einen Effekt aus. Eine Fehlinterpretation besteht oftmals darin, nach Berechnung der Prüfgröße einer Tabelle den korrespondierenden Wahrscheinlichkeitswert zu entnehmen und eine der folgenden Zuordnungen zu treffen: signifikant (*), hochsignifikant (**) oder höchst signifikant (***). Damit wird das Modell des statistischen Tests überinterpretiert, p wird als komparatives Stützmaß angesehen, „hochsignifikant“ ist signifikanter als „signifikant”, und solcherart nachgewiesenen Effekten kann man noch mehr trauen. Die „*“ legen die Assoziation nahe, die Befunde sollten durch Michelin-Sterne ausgezeichnet werden. Aber mehr als ein klassifikatorischer Stützungsbegriff kann aus den Theorien von Fisher (verwerfen/nicht verwerfen) oder Neyman und Pearson (annehmen/eventuell nicht entscheiden/ablehnen) nicht abgeleitet werden. Wer Aussagen des Typs „ist besser gestützt als“ anstrebt, muß sich anderer Konzepte bedienen. Man könnte in diesem Fall etwa unter Verwendung des Likelihoodkonzepts und der Likelihoodregel die Likelihoods der Daten unter der Annahme der Richtigkeit zweier konkurrierender Hypothesen vergleichen. Unter Beachtung der Axiome von Koopman ließe sich dann ein **komparativer Stützungsbegriff** entwickeln (siehe Stegmüller 1973, S. 84).

Wir wollen dem Signifikanztest nun nicht jedwede Berechtigung im Forschungsprozeß absprechen, in einem frühen Stadium der Theorienbildung als „screening”-Prozedur und/oder zur zufallskritischen Absicherung kann er wertvolle Dienste leisten. Auf diesen letzten Aspekt weist etwa Mittenecker (1983) hin: „Erkenntnis im Alltag (geht) in gleicher Weise vor, nur unter Anwendung von schlecht oder nicht kontrollierten ‚Statistiken‘: Das Herausbilden von Vorurteilen und ‚Aberglauben‘ war und ist das jahrtausendealte Ergebnis vieler Faktoren, unter anderem sicher aber auch des Nichtbeachtens des Zufalls in Stichproben und Beobachtungen.“ Es sollte sich allerdings in der Psychologie einbürgern, daß „Signifikanz“ für eine Hypothese kein hinreichendes Kriterium (vielleicht ein notwendiges) dafür sein kann, als Aussage „in den Bestand des Wissens, der eine Wissenschaft ausmacht, aufgenommen zu werden“ (Wendt 1983, S. 471). Auf die Gefahr für eine Wissenschaft, wenn einzig der Signifikanztest in der Rolle von Ockhams Rasiermesser entscheidet, was in den Bestand aufgenommen wird und was nicht, hat bereits Popper (1966, S. 198) hingewiesen. In einer Fußnote beklagt er sich über Carnap; die-

ser hatte das Poppersche Konzept „Grad der Bewährung“ uminterpretiert als „Grad der Bestätigung“ und damit als ein Synonym für eine Wahrscheinlichkeit. Nur hatte Popper mit seinem „Grad der Bewährung“ für eine Hypothese viel mehr gemeint als eine wahrscheinlichkeitstheoretische Betrachtung der Relation zu den Daten. Zur Bewährung gehörte neben dieser Betrachtung auch die Aufklärung des Zusammenhangs zwischen Hypothese und Theorie, die begriffliche Eindeutigkeit der Hypothese sowie die Bewährung in bezug auf inhaltlich echt konkurrierende Hypothesen. Nur solche Hypothesen sollten als vorläufig bewährt gelten, die einen fairen Wettkampf bestanden hatten.

6. Hypothesenprüfung nach Bayes

Spätestens seit der klassischen Arbeit von Edwards, Lindman und Savage (1963) gelten die „Bayesianer“ neben der Signifikanztestung nach Fisher und der Testtheorie nach Neyman und Pearson als dritte große statistische Schule. Das Bayessche Verfahren zur Prüfung von Hypothesen endet jedoch nicht notwendig mit einer statistischen Entscheidung für oder gegen eine Hypothese, oft genügt die Angabe von Aposteriori-Wahrscheinlichkeiten oder von Wettquotientcn, den uns schon von Karl Pearson bekannten odds. Die bayesianische Statistik befaßt sich mit „der graduellen Stützung oder Schwächung von Hypothesen auf Grund von Wahrscheinlichkeiten. Die beiden kategorialen Entscheidungsmöglichkeiten ‚für’ oder ‚gegen’ eine Hypothese, ‚Akzeptierung’ oder ‚Verwerfung’ einer Hypothese, die bei einem Test im Vordergrund stehen, spielen im Bayes-Ansatz eher eine sekundäre Rolle“ (Kleiter 1981, S. 13). In einer klassisch einfachen Schreibweise, in der „basic form” (Edwards et al. 1963, S. 198), kann man das Theorem von Bayes, für diesen Zweck interpretiert als Satz zur Revision subjektiver Wahrscheinlichkeiten, schreiben als

$$p(H \mid D) = \frac{p(H)\, p(D \mid H)}{p(D)} \qquad (6.1)$$

Natürlich liegt das für Hypothesenprüfung interpretierte Theorem von Bayes in mannigfaltiger Form auch für stetige Fälle von Hypothesen, von Daten oder von Prüfstatistiken vor (siehe etwa Kleiter 1981). Die problematischste Größe in obiger Formel ist p(H), die Apriori-Wahrscheinlichkeit einer Hypothese vor der Betrachtung neuer Daten. Zunächst ist anzumerken, daß für einen Forscher in die Schätzung von p(H) drei Informationsquellen eingehen können:

(1) Wenn keinerlei empirische Indizien bezüglich der Hypothese bekannt sind, geht in p(H) nur die subjektive Evidenz des Forschers in seine Hypothese ein (oder eine Wette, siehe weiter unten).

(2) Wenn empirische Indizien aus eigenen oder fremden Quellen vorliegen, können diese in die Schätzung von p(H) eingehen.
(3) Wenn in einer Voruntersuchung bereits eine Hypothesenbewertung mittels des Bayesschen Theorems vorgenommen wurde, so kann die daraus resultierende Aposteriori-Wahrscheinlichkeit als neue Apriori-Wahrscheinlichkeit in die Nachfolgeuntersuchung eingehen. Diese Revision kann beliebig oft wiederholt werden, und dies wird von Bayesianern auch als gewichtiger Vorteil ihres Verfahrens genannt, erspart es ihnen doch die Problematik von Metaanalysen, wie sie im Gefolge von Signifikanztests üblich sind.
(4) Wenn zu einer Hypothese mehrere stochastisch unabhängige Datensätze vorliegen und diese Daten auch sonst unter vergleichbaren Bedingungen erhoben wurden, kann in einer simultanen Verrechnung die Apriori-Wahrscheinlichkeit revidiert werden (Beispiele hierzu siehe etwa Kleiter, 1981, S. 127):

$$p(D_1,...,D_n \mid H) = p(D_1 \mid H)p(D_2 \mid H)...p(D_n \mid H) \tag{6.2}$$

p(H | D) als Aposteriori-Wahrscheinlichkeit oder „inverse" Wahrscheinlichkeit ist die vom Forscher „im Lichte der Daten" revidierte Wahrscheinlichkeit, p(H | D) ist die Größe, in die nach Bayesianischer Auffassung das „Lernen aus Erfahrung" eingeht. „Das Bayes-Theorem . . . ist eine Regel, wie die Unsicherheit über eine statistische Hypothese auf Grund einlangender Daten rational revidiert wird" (Kleiter, 1981, S. 14). Die Größe D bezeichnet hier, anders als etwa bei der Signifikanztestung, nicht nur den Wert eines Parameters oder den Wert einer statistischen Prüfgröße für diesen Parameter, sondern D kann die in allen Daten enthaltene Information in Form der gesammten Verteilung bezeichnen.

Auch dieser Sachverhalt, bei der Hypothesenprüfung nicht notwendig nur auf wenige Parameter zurückgreifen zu müssen, wird von Bayesianern als gewichtiger Vorteil ihres Verfahrens angeführt. p(D | H) ist die schon von Fisher (1959, S.68) bekannte Likelihood, die „Wahrscheinlichkeit" der Daten unter der Hypothese.

Einer der zentralen Kritikpunkte am bayesianischen Ansatz betrifft den zugrundeliegenden „subjektiven" Wahrscheinlichkeitsbegriff, im Gegensatz zum sonst in der Statistik üblichen objektiven und meistens frequentistischen Wahrscheinlichkeitsbegriff. Allerdings, „daß es sich . . . bei dem Begriffspaar personelle (d.h. subjektive) Wahrscheinlichkeit und statistische Wahrscheinlichkeit . . . um zwei voneinander verschiedene, wissenschaftlich wichtige und exakt durchführbare Deutungen des Wahrscheinlichkeitskalküls handelt, wird keinesfalls allgemein anerkannt" (Stegmüller 1973, S. 27). Wie schon zwischen Fisher und Neyman und Pearson, so tun sich zwischen Bayesianer und Nicht-

Bayesianern schier unüberwindliche weltanschauliche Differenzen bezüglich dieses Themas auf, die etwa Stegmüller (1973, S.41) dadurch zu überwinden sucht, daß er in Anlehnung an Braithwaite vorschlägt, Wahrscheinlichkeiten weder als frequentistisches noch als subjektives Konzept einzuführen, sondern als theoretischen Begriff, als eine nicht explizit definierbare theoretische Disposition physikalischer Systeme. Wir wollen diesen Grundlagenstreit hier nicht weiter verfolgen, können allerdings konstatieren, daß der Bayessche Ansatz wegen dieses Subjektivismus entschiedene Ablehnung erfährt. So bemerkt etwa Bredenkamp (1972, S.149): "... diese Kombination (von persönlichen Hypothesenwahrscheinlichkeiten und statistischen Wahrscheinlichkeiten) ist nur sinnvoll, wenn die subjektiven Wahrscheinlichkeiten den Axiomen und Theoremen der mathematischen Wahrscheinlichkeitslehre genügen. Oftmals summieren sich die persönlichen Wahrscheinlichkeiten nicht zu 1." Dazu soll allerdings bemerkt werden, daß uns nichts daran hindert, für ein normatives subjektives Wahrscheinlichkeitskalkül ein von der Axiomatik der Wahrscheinlichkeitsräume begründet abweichendes Kalkül einzuführen, wie es etwa neuerdings unter dem Label „belief functions" (vgl. Shafer 1982) nicht ohne Erfolg geschieht.

Rützel (1979) vermutet, die Apriori-Wahrscheinlichkeiten seien als subjektive Überzeugungsstärken nicht neutral und es gäbe keine intersubjektiv eindeutigen Kriterien zu ihrer Festsetzung. Er wurde deshalb bayesianisches Hypothesentesten als „privates Hypothesentesten" bezeichnen, und als solches „ist es unverbindlich und für den öffentlichen wissenschaftlichen Gebrauch abzulehnen". Auch Witte (1980, S. 40) sieht diese Gefahr: „Zum einen werden subjektive Wahrscheinlichkeiten a priori bezüglich der Wahrscheinlichkeit der Hypothese zugelassen, zum anderen wird die Testtheorie ausgebaut zu einer Entscheidungstheorie, d.h. es werden den möglichen Handlungen Nutzenwerte zugeordnet... Beide Erweiterungen bringen wegen der Subjektivität der einzuführenden Größen erhebliche Schwierigkeiten mit sich. Sie erschweren den wissenschaftlichen Diskurs."

Tholey (1982) vermutet als ein Motiv für diese Art der Ablehnung des Bayesianismus „offenbar eine völlig irrationale . . . Angst vor allem, was den Anschein des Subjektivismus hat". Die Bayesianer führen gegen diese Vorwürfe, mittels eines subjektiven Wahrscheinlichkeitsbegriffs ließen sich statistische Hypothesenprüfungen nicht rational vornehmen, gewichtige Gegenargumente ins Feld. Zum einen wenden sie ein, daß es relativ gleichgültig sei, welche Aprioris man zunächst einsetzt, denn nach dem „principle of stable estimation" (siehe etwa Wendt 1983, S. 501) werden auch sehr schlecht geschätzte Aprioris durch einigermaßen aussagekräftige Daten sehr schnell zu erheblich besseren Hypothesenwahrscheinlichkeiten $p(H \mid D)$ korrigiert. Zum anderen verweisen Bayesianer gern auf die Effizienz ihres Verfahrens im Vergleich zu Menschen im Alltag, die sich als konservative Informationsverarbeiter erwei-

sen. So kommen Coombs, Daves und Tversky (1975, S. 176f.) bezugnehmend auf Untersuchungen von Ward Edwards zu dem Schluß, „Menschen sind konservative Wahrscheinlichkeitsschätzer in dem Sinne, daß ihre Schätzungen um ein Beträchtliches weniger extrem sind als die nach der Bayesschen Regel berechneten".

Wir möchten hier auf ein interessantes Phänomen hinweisen: Für Fisher war der den Signifikanztest nutzende skeptische Wissenschaftler „konservativ", er hielt länger an seiner Nullhypothese fest als ein Mensch im Alltag; für Bayesianer sind die Menschen im Alltag „konservativ", sie nutzen die in den Daten enthaltene Information nicht so effektiv wie das Bayessche Theorem.

Als weiteres Gegenargument gegen den Vorwurf des Subjektivismus führen die Bayesianer an, daß mit p(H) nicht unbedingt der subjektive, private Glaubensgrad eines Forschers gemeint ist, daß man vielmehr den hier verwendeten subjektiven Wahrscheinlichkeitsbegriff in einem normativen Sinn interpretieren kann, denn „subjektive Wahrscheinlichkeit bedeutet hier nicht den empirisch zu ermittelnden Glaubensgrad einer Person, sondern denjenigen Glaubensgrad, den eine rationale Person besitzen sollte, . . . man bezieht sich meist auf den Begriff eines fairen und rationalen Wettsystems". (Tholey 1981). Ein Beispiel für diese Idee der Operationalisierung durch Wetten gibt Kleiter (1981, S. 117); diese Darstellung entspricht der „odd-form" des Theorems von Bayes, eine Form, die auch Edwards et al. (1963, S.218) diskutieren. Es lägen zwei konkurrierende Hypothesen H_i vor, i = 1,2. Aus der Anwendung des Bayesschen Theorems

$$p(H_i) \mid D = \frac{p(H_i)p(D \mid H_i)}{p(D)} \qquad \textbf{(6.3)}$$

und nach Betrachtung der beiden Quotienten

$$\lambda_0 = p(H_1) \,/\, p(H_2) \qquad \textbf{(6.4)}$$

$$\lambda_n = p(H_1 \mid D) \,/\, p(H_2 \mid D) \qquad \textbf{(6.5)}$$

läßt sich als Quotient der Bayes-Theoreme für H_1 und H_2 schreiben:

$$\lambda_n = \lambda_0 L \qquad (6.6)$$

mit L als Likelihood-Quotient $p(D \mid H_1)/p(D \mid H_2)$, der angibt, um wieviel wahrscheinlicher die Daten unter H_1 sind als unter $H_,$. Man kann also sagen: die „posterior odds" sind gleich dem Produkt aus „prior odds" mal „likelihood ratio". Man geht davon aus, „daß der Wert (Nutzen) der Konsequenz zahlenmäßig auf einer Intervallskala bestimmt ist, . . . daß man auch den Wetten einen Wert zuordnen kann und daß dieser Wert eine Reihe von Eigenschaften aufweist, insbesondere, daß er unter bestimmten Voraussetzungen additiv ist".

Solche Axiomatisierungen von Nutzen- und Wettsystemen liegen vor, etwa bei Kleiter (1981, S. 42) oder bei Luce und Krantz (1971), die ein System von bedingten Entscheidungen axiomatisieren. In Krantz, Luce, Suppes und Tversky (1971, S.373) geben sie ein Beispiel für solche bedingten Entscheidungen: angenommen, jemand hat sich entschlossen, an einem bestimmten Tag von New York nach Boston zu fahren und hat dafür das Flugzeug, den Zug oder das Auto ins Auge gefaßt. Jede Möglichkeit zieht nun aber bestimmte Risiken mit bestimmten Konsequenzen nach sich, und diese können sich von Fahrt zu Fahrt ändern. Wichtig ist, daß jede Wahl zwischen den Fahrmöglichkeiten die möglichen Risiken einschränkt auf jene, die dem gewählten Transportmittel eigen sind; die Bewertung jeder bedingten Entscheidung braucht auch nur jene Risiken in Rechnung zu stellen. Wer etwa ans Flugzeug denkt, muß folgende mögliche Ausgänge beachten: pünktliche Ankunft, unpünktliche Ankunft, Streichung des Fluges, Absturz usw. Einige dieser möglichen Ausgänge können nur für einen einzigen Flug zutreffen, andere können typisch sein für alle Fahrten mit diesem Transportmittel, etwa die Streichung eines Fluges. Andererseits schließt die Wahl eines Transportmittels manche Risiken aus, ein Flugzeug kann etwa nicht unterwegs liegenbleiben wie ein Auto. Dieses Beispiel mag als Illustration der Situation eines Bayesianers dienen, der sich zwischen verschiedenen Hypothesen, verbunden mit verschiedenen Aprioris und Aposterioris, und verschiedenen Entscheidungskonsequenzen bewegt.

Wir können die Debatte über die Angemessenheit des Bayesschen Ansatzes nicht beenden, allerdings ist Witte (1980, S.40) zuzustimmen, der zu bedenken gibt: „Gleichzeitig jedoch können (die subjektiven Wahrscheinlichkeiten und die Nutzenwerte für die Handlungen) sehr nützliche Erweiterungen sein, weil man die gesamte Subjektivität des Forschers abbilden kann in eine Entscheidungsstrategie, denn diese subjektiven Elemente sind nicht völlig zu eliminieren.“ Richtig, denn mit der Wahl eines Versuchsplans, eines Zufallsmodells einer Prüfstatistik und einer Entscheidungsregel gehen auch bei Signifikanztestern willkürliche und letztlich subjektive Momente ein, die nur deshalb als solche wenig auffallen, weil sie als Konventionen von den meisten Anwendern geteilt werden.

7. Schlußbemerkungen

Wir haben drei klassische Zugänge zum Problem der Hypothesenprüfung mittels statistischer Modelle diskutiert und dabei dem in der Psychologie häufig verwandten Signifikanztest besonderes Augenmerk geschenkt. Nur, ist es in einer empirischen Wissenschaft überhaupt notwendig, soviel Statistik zu treiben? Bredenkamp (1980, S. 23) meint ja, denn “ . . . leider ist es so, daß in der Psychologie mit wenigen Ausnahmen nicht auf Statistik verzichtet werden kann“. Den Grund sieht Witte (1989) im Entwicklungsstand der Psychologie:

„So wie viele Theorien konstruiert sind, ist der klassische Signifikanztest die einzig mögliche Überprüfungsmethode und besser als reine Spekulation, weil man zumindest Zufallsschwankungen nicht als Stützung interpretiert.“ Nicht nur statistische Entscheidungen werden befürwortet, speziell der Signifikanztest wird präferiert. So schreibt etwa Westermann (1987, S. 124): „Selbstverständlich können außer dem Signifikanztest auch andere Ansätze zur Generierung von Kriterien über empirische Hypothesen herangezogen werden. Insbesondere ist hier an die Likelihood-Tests und die Bayes-Statistik zu denken. Der Signifikanztest ist jedoch m. E. beiden Alternativen vorzuziehen.“

Eindeutige Stellungnahmen also, aber für welchen Signifikanztest? Für den klassischen Nullhypothesentest nach Fisher? Damit ließen sich keine Effekte sichern. Für den modifizierten Neyman-Pearson-Test, wie etwa Bredenkamp ihn vorgeschlagen hat? Oder für jenen seltsamen Zwitter, der sich in der Psychologie so großer Beliebtheit erfreut? Methodiker plädieren immer wieder für ein Testen nach Neyman und Pearson, denn „um einen Experimentalaufbau angemessen beurteilen zu können, ist (es) notwendig, α- und β-Fehler anzugeben sowie ein Effektmaß“ (Witte 1989). Dabei sollten α und ß als kleine Werte festgelegt werden (Bredenkamp 1972, S. 149). Es bleibt zunächst unerklärlich, warum viele Anwender diesen methodischen Maximen nicht folgen wollen. Dabei eröffnet gerade der modifizierte Signifikanztest eine weitere methodologische Möglichkeit, denn wie nicht nur Bredenkamp (1980, Vorwort S. VI) fordert, sind „experimentelle Untersuchungen entgegen den üblichen Gepflogenheiten von vorneherein so zu planen, daß psychologische Hypothesen falsifiziert werden können. Dieser Gesichtspunkt ist deshalb so wichtig, weil die Nullhypothese häufig der wissenschaftlichen Hypothese widerspricht, so daß auf ihr Zutreffen erkannt werden können muß, damit eine psychologische Hypothese, die statistisch überprüft wird, überhaupt falsifizierbar ist.” Und, wie erinnerlich, kann H_0 in seinem Modell mit Wahrscheinlichkeit $1-\beta$ angenommen werden. Wir würden allerdings vorschlagen, den Begriff der Falsifikation der Logik zu überlassen und hier lieber davon zu sprechen, die psychologische Hypothese zu verwerfen. Die Frage ist nur, wieviele Anwender sich überhaupt dem von Popper begründeten Falsifikationismus verpflichtet fühlen, die meisten scheinen lieber für ihre wissenschaftliche Hypothese (meist umgesetzt als statistische H_1) kämpfen zu wollen als gegen sie, und nicht wenige Anwender sind wohl eigentlich „implizite Verifikationisten“, die mit jeder „Signifikanz“ den Fundus gesicherten Wissens erweitern wollen - diesen Anwendern kommt der Zwitter gerade recht. Wie aber Gigerenzer et al. (1991, S. 107) richtig bemerken, haben weder Fisher noch Neyman und Pearson je davon gesprochen, daß eine Nullhypothese etwa auf dem 0,01 -Niveau zurückgewiesen werden könnte.

Eine ebenfalls von Methodikern erhobene Forderung, „bei der Prüfung von Theorien und Hypothesen . . . sowohl die wissenschaftliche Signifikanz in

Form von Effektgrößen wie die statistische Signifikanz in Form von Wahrscheinlichkeitsaussagen zu berücksichtigen“ (Westermann & Hager, 1984), findet ebenfalls kaum Gehör. Schon Bakan (1966) stellte die Frage, „how much of a difference makes a difference for what?“, die Frage nämlich, wie groß ist der Effekt und wie groß ist der Gewinn für die Theorie. Viele Begriffe (praktische Signifikanz, praktische Bedeutsamkeit, wissenschaftliche Signifikanz, wissenschaftliche Bedeutsamkeit, „material significance“) für diesen Aspekt und etliche Verfahren zur numerischen Abschätzung liegen vor (siehe etwa Bredenkamp 1970; Hager 1983), nur werden sie selten eingesetzt. Hager und Westermann (1982) vermuten als Grund: „Es scheint wohl zwischen den Herausgebern und Lesern ein Gentlemans Agreement zu bestehen, daß nie Maße der erklärten Varianz verlangt werden, sondern stets nur Signifikanztests. Warum? Vielleicht zum Teil deshalb, weil Maße der erklärten Varianz für alle so peinlich wären.“ Wenn dem so wäre, stünde es schlecht um die von Lakatos angeführte intellektuelle Redlichkeit, denn dann ginge es vornehmlich um die Immunisierung der eigenen Theorien.

Wir haben versucht darzulegen, daß statistische Entscheidungen stets dort sinnvoll sind, wo in begründeter und problematisierbarer Weise ein statistisches Modell an die Daten gelegt werden kann und wo in begründeter und problematisierbarer Weise Entscheidungen getroffen werden können. Insbesondere Befürwortern der Signifikanztestung sei angeraten, einige Punkte zu bedenken:

Das Problem, welcher Parameter ein inhaltlich angemessener Indikator für ein psychologisches Konstrukt ist, muß wissenschaftlich, nicht statistisch begründet werden.

In das Modell des Signifikanztests geht Hintergrundwissen in Form akzeptierter statistischer Oberhypothesen und stochastischer Unabhängigkeitsannahmen ein. Diese sind zwar zum Teil selbst wieder statistisch prüfbar, zum Beispiel die Annahme der stochastischen Unabhängigkeit mit dem Runs-Test und die Annahme der Normalverteilung mit dem Kolmogoroff-Smirnoff-Test (nur gehen in diese Verfahren natürlich auch wieder Oberhypothesen ein). Viele Anwender enden mit ihren Überlegungen dort, wo schon „Student“ angefangen hatte: Wenn man schon eine Verteilungsannahme machen muß, dann nimmt man wenigstens eine Verteilung, die gut bekannt und tabelliert ist. In problematischen Fällen, in denen die übliche Normalverteilungsannahme offensichtlich verletzt scheint, spricht wenig dagegen, erst einmal einige ihrer Parameter zu schätzen und gegenüber diesen Vorannahmen bayesianisch hinzuzulernen.

Die Wahl des statistischen Zufallsmodells, gegen das man sich zufallskritisch absichern will, ist nicht trivial. Wie am Einwand von Neyman bezüglich der hypergeometrischen und der Binomialverteilung aufgezeigt, gibt es komplexe - und leider nur wenig beachtete - Zusammenhänge

zwischen experimenteller Versuchsanordnung, resultierenden Zufallsvariablen und statistischen Entscheidungen.

- Die statistische Entscheidung, die nach den vorher festzulegenden Entscheidungsregeln getroffen wird, ist nicht nur bezüglich ihrer möglichen Irrtümern und deren möglichen Kosten zu problematisieren, sondern auch bezüglich der Theorie, aus der die wissenschaftlichen Hypothesen abgeleitet sind. So bemerkt etwa Deppe (1977, S. 162), daß es zwischen wissenschaftlichen Hypothesen und statistischen Hypothesen (d.h. solchen, die das beobachtbare Verhalten von Zufallsvariablen zum Gegenstand haben) zwei Übersetzungen geben muß: „Die Übersetzung der wissenschaftlichen Hypothesen in statistische, und die Rückübersetzung des Ergebnisses der Prüfung statistischer Hypothesen in ein die wissenschaftlichen Hypothesen betreffendes Ergebnis." Dieses Problem der Rückübersetzung der statistischen Entscheidungen in die Theorie wurde deutlich bei der Diskussion der vorher zu postulierenden oder zumindest hinterher zu berichtenden Effektstärken.

Die Problematik der überwiegend mit parametrischen Verfahren arbeitenden Tests haben wir aufgezeigt. Es fallt auf, daß in der Psychologie die oftmals naheliegenden nichtparametrischen Verfahren weitgehend ungenutzt bleiben. Tests für zentrale Tendenzmaße und Dispersionsmaße stehen ausreichend zur Verfügung (siehe etwa Büning und Trenkler 1978; Bortz, Lienert und Boehnke 1990). Mit Ausnahme des Kolmogoroff-Smirnoff-Tests haben die gebräuchlichsten nichtparametrischen Teststatistiken die Normalverteilung oder die χ^2-Verteilung als Asymptote; allerdings lassen sich für manche Tests Güteeigenschaften nicht herleiten. Bei vielen Fragestellungen wäre es hilfreich, die Möglichkeiten der Schätztheorie zu nutzen und insbesondere Konfidenzintervalle anzugeben, eine Forderung, die Rozeboom schon 1960 stellte: "... wann immer möglich, sollte der grundlegende statistische Bericht in Form von Konfidenzintervallen erfolgen."

Wir hatten auf die strukturellen Ähnlichkeiten zwischen der Angabe solcher Intervalle und der Signifikanztestung nach Fisher hingewiesen. Bei vielen Publikationen wäre zudem zu fragen, ob der „basic statistical report" (Rozeboom) nicht ohnehin besser in Form einer wohlaufbereiteten deskriptiven Statistik erfolgen sollte. Leider ist die Ansicht weitverbreitet, die ritualisierte Angabe von (***) zur Charakterisierung von Signifikanzen sei in jedem Fall informativer als eine kluge Darstellung von Konfidenzintervallen oder Rohdaten, aber der Bericht von Rohdaten scheint unmodern geworden zu sein. „Der Streit um den Signifikanztest wird eigentlich stellvertretend geführt als Streit über Theorienbildung. . . . Wer eine ‚bessere' schließende Statistik will, muß präzisere Theorien formulieren", schreibt Witte (1989) über die „im wahrsten Sinne des Wortes provinziell geführte Signifikanztestkontroverse" in der Zeitschrift für Sozialpsychologie, und mit Witte können wir resümieren,

„seit der Kontroverse um 1970 haben sich keine neuen Diskussionspunkte ergeben".

Literatur

Bakan, D. (1966). The test of significance in psychological research. Psychologicul ***Bulletin, 66, 423-37.***

Bickel, P.J. (1969). A distribution-free version of the Smirnov 2-sample test in the p-variate case, ***Annals of Mathematical Statistics, 40, 1-23.***

Bortz, J., Lienert, G.A. & Boehnke, K. ***(1990). Verteilungsfreie Methoden in der Biostatistik.*** Berlin: Springer.

Bredenkamp, J. (1970). über Maße der praktischen Signifikanz. ***Zeitschrift für Psychologie, 177, 310-8.***

Bredenkamp, J. (1972). ***Der Signifikanztest in der psychologischen Forschung.*** Frankfurt: Akademische Verlagsanstalt.

Bredenkamp, J. (1980). ***Theorie und Planung psychologischer Experimente.*** Darmstadt: Steinkopff.

Büning, H. & Trenkler, G. (1979). Nichtparametrische ***statistische Methoden.*** Berlin: De Gruyter.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. ***Journal of Abnormal and Social Psychology, 65,*** 145-53.

Coombs, C. H., Daves, R. M. & Tversky, A. (1975). ***Mutbemutische Psychologie.*** Weinheim: Beltz.

David, F. ***N.*** (ed.) (1966). ***Research papers in statistics. Festschrift*** *for* ***J. Neyman.*** London: Wiley.

Deppe, W. ***(1977). Formale Modelle in der Psychologie.*** Stuttgart: Kohlhammer.

Diepgen, R. (1991). Inkonsistentes zur Signifikanzproblematik. Ein Kommentar zu Witte (1989). ***Psychologische Rundschau, 42, 29-33.***

Edwards, W., Lindman, H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research. ***Psychological Review, 70,*** 193-242.

Fisher, R. A. (6 1951). *The design of experiments.* Edinburgh: Oliver and Boyd.

Fisher, R.A. ***(2 1959). Statistical methods and scientific inference.*** London: Oliver and Boyd.

Freudenthal, H. & Steiner, H. G. (1966). Aus der Geschichte der Wahrscheinlichkeitstheorie und der mathematischen Statistik. In H. Behnke (Hrsg.), ***Grundzüge der Mathematik Band 4: Praktische Methoden und Anwendungen (S.*** 149-95). Göttingen: Vandenhoeck und Ruprecht.

Gigerenzer, G. (1986). Wissenschaftliche Erkenntnis und die Funktion der Inferenz-Statistik. Anmerkungen zu E. Leiser. ***Zeitschrift für Sozialpsychologie,*** 17, 183-9.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. (1991). ***The Empire of Chance. How probability changed science and everyday life.*** Cambridge: Cambridge University Press.

Haagen, K. & Seifert, H. G. (1979). ***Methoden der Statistik für Psychologen.*** Band II. Stuttgart: Kohlhammer.

Hacking, I. (1965). ***Logic of Statisticul Inference.*** Cambridge: University Press.

Hager, W. (1983). über univariate Maße der wissenschaftlichen Signifikanz. ***Zeitschrift für Psychologie, 191, 295-309.***

Hager, W. & Westermann, R. (1982). Die Elle - 10 Jahre danach. ***Zeitschrift für*** So***zialpsychologie, 13, 250-2.***

Hager, W. & Westermann, R. (1983). Planung und Auswertung von Experimenten. In J. Bredenkamp & H. Feger (Hrsg.). ***Hypothesenprüfung. Enzyklopädie der Psychologie, Serie 1, Band 5*** (S.24-238). Göttingen: Hogrefe.

Hartung, J. (^{3}1985). ***Statistik, Lehr- und Handbuch der angewandten Statistik.*** Miinchen: R. Oldenburg.

Herrmann, T. ***(1979). Psychologie als Problem.*** Stuttgart: Klett-Cotta.

Kendall, M. G. & Quart, A. ('1973). ***The advanced theory of statistics. Vol. 2: Inference and relationship.*** London: Griffin.

Kleiter, ***G.*** D. (1981). ***Bayes-Statistik.*** Berlin: de Gruyter.

Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. (1971). ***Foundations of measurement.*** New York: Academic Press.

Lakatos, I. (1974). Falsifikation und die Methodologie wissenschaftlicher Forschungsprogramme. In I. Lakatos & A. Musgrave (Hrsg.), ***Kritik und Erkenntnisfortschritt*** (S. 89-190). Vieweg: Braunschweig.

Leiser, E. (1982). Wie funktioniert sozialwissenschaftliche Statistik?. ***Zeitschrift für Sozialpsychologie, 13, 125-39.***

Lenzen, W. (1974). ***Theorien der Bestätigung wissenschaftlicher Hypothesen.*** Stuttgart: Frommann-Holzboog.

Lindgren, B.W. (^{3}1976). ***Statistical Theory.*** New York: Macmillan.

Luce, R.D. & Krantz, D.H. (1971). Conditional expected utility. ***Econometrica, 39,*** 253-71.

Lykken, D.T. (1968). Statistical significance in psychological research. ***Psychological Bulletin, 70,*** 151-9.

Meehl, P.E. (1967). Theory testing in psychology and physics: a methodological paradox. ***Philosophy of Science, 34,*** 103-15.

Mittenecker, E. (1983). Anmerkungen und Stellungnahme zu E. Leiser: Wie funktioniert die sozialwissenschaftliche Statistik? Zeitschrift ***für Sozialpsychologie, 14, 68-*** 71.

Morrison, D. E. & Henkel, R. E. (Hrsg.) (1970). ***The significance controversy.*** Chicago: Aldine.

Morrison, D. E. & Henkel, R. E. (1970). Significance test in behavioral research: skeptical conclusions and beyond. In D.E. Morrison & R.E. Henkel (Hrsg.), ***The significance controversy.*** Chicago: Aldine.

Neyman, J. (1935). On the problem of confidence intervals, ***Annals of Mathematical Statistics, 6,*** 111.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability, ***Philosophical Transactions, A 236, 333-80.***

Neyman, J. (1943). On the problem of testing hypotheses. ***Annals of Mathematical Statistics, 14, 238-52.***

Neyman, J. (1950). ***First Course in Probability and Statistics. New*** York: Holt.

Neyman, J. (^{2}1952). ***Lectures and conferences on mathematical statistics and probability.*** Washington: Department of Agriculture.

Neyman, J. und Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. ***Biometrika, 2OA, 175-240, 263-94.***

Neyman, J. und Pearson, E.S. (1933). On the problem of the most efficient of statistical hypotheses. ***Philosophical Transactions, A 231, 289-337.***

Neyman, J. und Pearson, E.S. (1933). The testing of statistical hypotheses in relation to probabilities a priori, ***Proceedings of the Cambridge Philosophical Society, 29,*** 492-510.

Pearson, E. ***S.*** (1966). ***The Neyman-Pearson Story: 1926-34.*** In F. N. David (Hrsg.), Research papers in statistics (S. 1-23). London: Wiley.

Popper, K. R. (^{2}1966). ***Logik der Forschung.*** Tübingen: Mohr.

Rozeboom, W. W. (1960). The fallacy of the nullhypothesis significance test. ***Psychological Bulletin, 57,*** 416-28.

Rützel, E. (1979). Bayessches Hypothesentesten und warum die Bayesianer Bias-ianer heißen sollten. ***Archiv für Psychologie, 131, 211-32.***

Shafer, G. (1982). Belief functions and parametric models. ***Journal of the Royal*** Sta***tistical Society, B 44, 322-52.***

Sigel, ***S.*** (1956). ***Nonparametric Statistics for the Behavioral Sciences. New*** York: McGraw-Hill 1956.

Stegmüller, W. (1969). ***Wissenschaftliche Erklärung und Begründung.*** Berlin: Springer.

Stegmüller, W. ***(1973). Personelle und statistische Wahrscheinlichkeit. 2. Halbband: Statistisches Schließen, statistische Begründung, statistische Analyse.*** Berlin: Springer.

Stigler, S. (1977). Eight centuries of sampling inspection: the trial of the pyx. Journal ***of the American Statistical Association, 72, 493-500.***

Tholey, P. (1982). Significanztest und Bayessche Hypothesenprüfung. ***Archiv für Psychologie, 134, 319-42.***

von Mises, R. (1942). On the correct use of Bayes' formula. ***Annals of Mathematical Statistics, 13,*** 156-63.

von Mises, R. (1964). ***Mathematical Theory of Probability and Statistics. New*** York: Academic Press..

Wald, A. (1947). ***Sequential Analysis. New*** York: Wiley.

Wald, A. (1950). ***Statistical Decision Functions. New*** York: Wiley.

Wendt, D. (1966). Versuche zur Erfassung des subjektiven Verläßlichkeitsniveaus. ***Zeitschrift für Psychologie, 172, 40-81.***

Wendt, D. (1983). Statistische Entscheidungstheorie und Bayes-Statistik. In J. Bredenkamp & H. Feger (Hrsg.), ***Hypothesenprüfung. Enzyklopädie der Psychologie, Themenbereich B, Serie*** 1, ***Bund 5*** (S.471-529). Göttingen: Hogrefe.

Westermann, R. (1987). ***Strukturalistische Theorienkonzeption und empirische Forschung in der Psychologie.*** Berlin: Springer.

Westermann, R. & Hager, W. (1984). Zur Verwendung von Effektgrößen in der theorie-orientierten Sozialforschung. ***Zeitschrift für Sozialpsychologie, 15,*** 159-66.

Witte, E. ***(1980). Signifikanztest und statistische Inferenz.*** Stuttgart: Enke.

Witte, E. (1989). Die „letzte" Signifikanztestkontroverse und daraus abzuleitende Konsequenzen. ***Psychologische*** Rundschau, 40, 76-84.

Witte, E. (1991). Antworten auf die „Bemerkungen" von Diepgen. ***Psychologische Rundschau, 42, 34-7.***

Wottawa, H. (1990). Einige Überlegungen zu (Fehl-)Entwicklungen der psychologischen Methodenlehren. ***Psychologische*** Rundschau, ***41,*** 84-107.