DIE FALSIFIKATION STATISTISCHER HYPOTHESEN*

MAX ALBERT

SUMMARY. *The Falsification of Statistical Hypotheses.* It is widely held that falsification of statistical hypotheses is impossible. This view is supported by an analysis of the most important theories of statistical testing: these theories are not compatible with falsificationism. On the other hand, falsificationism yields a basically viable solution to the problems of explanation, prediction and theory testing in a deterministic context. The present paper shows how to introduce the falsificationist solution into the realm of statistics. This is done mainly by applying the concept of empirical content to statistical hypotheses. It is shown that empirical content is a substitute for 'power' as defined by Neyman and Pearson. Since the empirical content of a hypothesis is independent of alternative hypotheses, the proposed theory of statistical testing allows for tests of isolated hypotheses.

Key words: Wissenschaftstheorie, Statistik, Falsifikationismus, Testtheorie, Hypothesentests, Signifikanztests, Neyman-Pearson-Theorie.

1. EINLEITUNG

Die vorliegende Arbeit enthält einen Vorschlag zur Lösung eines Problems der statistischen Testtheorie. Man könnte das Problem als Frage formulieren: Sind statistische Hypothesen falsifizierbar? Diese Frage wird normalerweise mit einem klaren „Nein" beantwortet. Akzeptiert man diese Antwort, dann folgt, daß der Falsifikationismus für einen großen Teil der modernen Wissenschaft irrelevant ist. Ich werde jedoch versuchen zu zeigen, daß das Problem der Falsifikation statistischer Hypothesen im Rahmen des Falsifikationismus gelöst werden kann.

Zur Klärung der Problemsituation werde ich die beiden wichtigsten Testtheorien diskutieren, nämlich die Theorie Fishers und die Neyman-Pearson-Theorie (NPT). Die Darstellung beginnt mit der Theorie Fishers, die unter allen Testtheorien dem Falsifikationismus am nächsten steht. Fishers Theorie liefert jedoch keine Anhaltspunkte für die Wahl des richtigen unter den vielen Signifikanztests, die in dem meisen Fällen denkbar sind und die im allgemeinen zu gegensätzlichen Ergebnissen führen. Dadurch ist keine objektive Entscheidung darüber möglich, ob eine Hypothese mit den vorliegenden Daten vereinbar ist oder nicht, so daß die Testtheorie eine aus falsifikationistischer Sicht wesentliche Funktion nicht erfüllen kann.

Eine Kritik an der Uneindeutigkeit des Fisherschen Signifikanztest war auch der Ausgangspunkt für die NPT. Nach dieser Theorie wird der Signifikanztest gewählt, bei dem die Wahrscheinlichkeit, die Hypothese zu widerlegen, wenn eine bestimmte Alternativhypothese wahr ist, so groß wie möglich ist. In den meisten Fällen muß man jedoch mit der Möglichkeit

Journal for General Philosophy of Science 23: 1–32, 1992.

rechnen, daß weder die zu prüfende Hypothese noch die in Betracht gezogene Alternative wahr ist. Nach der NPT kann es keinen Test geben, der diese Möglichkeit berücksichtigt. Das führt dazu, daß auch diese Theorie mit dem Falsifikationismus nicht vereinbar ist.

Der Leitgedanke meines eigenen Vorschlages, den ich Fishers Theorie und der NPT gegenüberstelle, besteht darin, die Idee des empirischen Gehalts auf die statistische Testtheorie anzuwenden. Durch die Poppersche Regel der Maximierung des empirischen Gehalts läßt sich ein bestimmter Signifikanztest auszeichnen, der zur Falsifikation statistischer Hypothesen verwendet werden kann. Der empirische Gehalt tritt an die Stelle der Macht eines Tests; die Definition des empirischen Gehalts greift jedoch nicht auf mögliche Alternativen zu der zu prüfenden Hypothese zurück.

Die Arbeit ist wie folgt aufgebaut. Zunächst stelle ich in Abschnitt 2 den Zusammenhang zwischen dem Falsifikationismus und dem Testproblem in der Statistik her. Dem schließen sich mit den Abschnitten 3 und 4 eine kritische Darstellung der Fisherschen Theorie und der NPT an. Abschnitt 5 behandelt meinen eigenen Vorschlag zur Testtheorie. Abschnitt 6 enthält einige Schlußbemerkungen.

2. FALSIFIKATIONISMUS UND STATISTIK

Der Falsifikationismus läßt sich recht gut anhand des deduktiv-nomologischen Modells der Erklärung und Prognose darstellen (Popper 1984: 31ff, Hempel & Oppenheim 1948). Nach diesem Modell macht eine Erklärung oder Prognose von mindestens einer Gesetzeshypothese (H) Gebrauch, aus der mit Hilfe von singulären Sätzen, den Randbedingungen (R), weitere singuläre Sätze (P), die das zu erklärende oder vorherzusagende Ereignis beschreiben, logisch abgeleitet werden können. Wir unterstellen im folgenden, daß diese singulären Sätze durch Beobachtung überprüft werden können. Aus der Hypothese H folgt somit, daß der durch Beobachtung überprüfbare Satz „R und nicht P" falsch sein muß. Die Menge der möglichen Beobachtungssätze, die von einer Hypothese ausgeschlossen werden und sie daher falsifizieren können, wird von Popper (1984: 83ff) als der empirische Gehalt der Hypothese bezeichnet. Erklärungs- und Prognosekraft einer Hypothese beruhen auf dem empirischen Gehalt. Man wird sich daher Hypothesen mit möglichst hohem empirischem Gehalt wünschen. Gleichzeitig steigt mit dem empirischen Gehalt die Verwundbarkeit oder Falsifizierbarkeit einer Hypothese.

Ich setze voraus, daß im nicht-statistischen oder „deterministischen" Fall Erklärung, Prognose und Falsifikation durch das deduktiv-nomologische Modell grundsätzlich zufriedenstellend behandelt werden. Unter dieser Annahme ist eine möglichst weitgehende Übertragung dieses Modells auf den statistischen Fall wünschenswert. Hempels (1977) Modell der induktiv-statistischen Erklärung zeigt, wie diese Übertragung aussehen könnte. Nach diesem Modell kann eine Hypothese, die einem Ereignis E eine Wahrschein-

lichkeit nahe 1 zuschreibt, das Eintreten von E erklären, da man ja E nach dieser Hypothese auch vorhersagen oder vernünftigerweise erwarten würde. Eine Interpretation dieser Erklärung als induktives Argument mußte mit dem Scheitern der Versuche, eine induktive Logik zu finden, aufgegeben werden. Damit muß man auf eine *methodologische Regel* zurückgreifen, die es erlaubt, von der hohen Wahrscheinlichkeit eines Ereignisses zu der Prognose, das Ereignis werde tatsächlich eintreten, überzugehen (Gadenne 1989). Diese Idee läßt sich offensichtlich auch zur Falsifikation verwenden: Wir könnten uns entschließen, eine Hypothese abzulehnen, wenn etwas geschieht, das nach dieser Hypothese zwar möglich, aber sehr unwahrscheinlich ist. Schon vor Hempel hat Popper (1984: Kap. 8, insbes. 146ff) den Vorschlag gemacht, statistische Hypothesen durch eine solche Regel falsifizierbar zu machen; die vorliegende Arbeit kann als eine Ausarbeitung dieses Vorschlags aufgefaßt werden.

Eine einfache statistische Gesetzeshypothese hat folgende Form:

> H: Immer und überall, wenn eine Situation der Art S herrscht, gehorcht die Größe X der Wahrscheinlichkeitsverteilung F.

Der hier verwendete Gesetzesbegriff macht kontrafaktische Behauptungen möglich, verbindet also Gesetzmäßigkeit mit Notwendigkeit (Popper 1984: 380ff, Hempel 1977: 55f, Armstrong 1983). Dieser Gesetzesbegriff ist im Falle statistischer Hypothesen mit der inzwischen weitverbreiteten, auf Popper zurückgehenden Propensity-Interpretation der Wahrscheinlichkeit verbunden. Die Häufigkeitsinterpretation greift dagegen normalerweise auf den Humeschen Gesetzesbegriff zurück (Armstrong 1983: 29ff, Giere 1973: 478).

Die aus H folgende Hypothese, daß das sogenannte sichere Ereignis stattfinden wird, die Größe X also einen der durch H erlaubten Werte annimmt, bezeiche ich als den deterministischen Teil der statistischen Hypothese H. Der deterministische Teil hat empirischen Gehalt, wenn das sichere Ereignis irgendwelche Möglichkeiten ausschließt. Der deterministische Teil einer statistischen Hypothese ist unproblematisch; das Problem besteht darin, die Hypothese über die Wahrscheinlichkeiten der Elementarereignisse zu prüfen. Dieses Problem soll durch eine methodologische Regel gelöst werden, die es gestattet, von durch Beobachtung nicht prüfbaren Wahrscheinlichkeitsaussagen zu Aussagen überzugehen, die Beobachtbares beschreiben. Wenn man den hier uninteressanten Fall ausschließt, daß der deterministische Teil der Hypothese falsifiziert wird, bleibt nichts anderes übrig, als die Hypothese abzulehnen, wenn gewisse laut Hypothese mögliche Ereignisse eintreten. Die Menge der möglichen Ereignisse, deren Auftreten man zum Anlaß nimmt, die Hypothese abzulehnen, heißt Ablehnungsbereich der Hypothese; die möglichen Ereignisse, die nicht im Ablehnungsbereich liegen, bilden den Akzeptanzbereich. Die Summe der Wahrscheinlichkeiten aller Ereignisse im Ablehnungsbereich wird Irrtumswahrscheinlichkeit oder Signifikanzniveau genannt. Die Irrtumswahrscheinlichkeit ist der Preis, den

man dafür zahlen muß, daß man die Verteilungshypothese testen kann; sie gibt die Wahrscheinlichkeit an, mit der die Hypothese abgelehnt wird, falls sie wahr sein sollte.

Hat man einer Hypothese für jede Stichprobengröße einen Ablehnungsbereich eindeutig zugeordnet, kann man im Rahmen des deduktiv-nomologischen Modells argumentieren. Man lehnt die Hypothese als vorläufig falsifiziert ab, wenn die Beobachtungen in den Ablehnungsbereich fallen; ansonsten akzeptiert man sie vorläufig. Allerdings ist es möglich, daß durch weitere Beobachtungen die Falsifikation aufgehoben wird, weil in jeder vernünftigen Testtheorie der Akzeptanzbereich für zwei Beobachtungen extremere Werte für eine der beiden Beobachtungen zuläßt als der Akzeptanzbereich für eine einzige Beobachtung. *Logisch* herrscht damit eine völlig andere Situation als im deterministischen Bereich. *Methodologisch* ist der Unterschied jedoch gering, weil auch im deterministischen Bereich Falsifikationen fehlbar sind, da Beobachtungsaussagen auf Täuschungen beruhen können. Das Problem der Fehlbarkeit der Beobachtungsaussagen wurde von Popper als das Problem der empirischen Basis bezeichnet und bereits in der ‚Logik der Forschung' in befriedigender Weise gelöst (Popper 1984: 60ff, Andersson 1984). Wenn die Hypothese richtig ist, sollten im statistischen wie im deterministischen Bereich Falsifikationen nur selten vorkommen und sich nicht systematisch reproduzieren lassen. Das ist im statistischen Bereich garantiert, wenn die Irrtumswahrscheinlichkeit gering ist: irrtümliche statistische Falsifikationen lassen sich dann in Analogie zu sogenannten „okkulten Effekten" in der experimentellen Naturwissenschaft behandeln (Popper 1984: 9, 156).

3. DER FISHERSCHE SIGNIFIKANZTEST

Signifikanztests im Sinne von R. A. Fisher sind auf den ersten Blick gut mit dem Falsifikationismus vereinbar. So behauptet Fisher im Zusammenhang mit möglichen Leistungen der Statistik:

> ...we may be able validly to apply a test of significance to discredit a hypothesis the expectations from which are widely at variance with ascertained fact. (1959: 35)

Was Fisher unter Diskreditierung von Hypothesen versteht, stimmt mit dem überein, was weiter oben über vorläufige Falsifikation gesagt wurde. Insbesondere scheint Fisher einen fehlgeschlagenen Widerlegungsversuch nicht für hinreichend gehalten zu haben, um die geprüfte Hypothese zu ‚akzeptieren' (Giere 1977: 25f). Das ist konsequent, da von zwei konkurrierenden Hypothesen durchaus *beide* einen Widerlegungsversuch überstehen könnten. Unter ‚Akzeptanz' einer Hypothese im falsifikationistischen Sinn wird im folgenden nichts anderes verstanden, als daß die Hypothese vorläufig als ‚nicht falsifiziert' betrachtet wird,[1] daß sie also weiterhin als Bestandteil wissenschaftlicher Theorien in Frage kommt (Fisher 1959: 35).

Trotzdem ist Fishers Testtheorie mit dem Falsifikationismus nicht ver-

einbar, wie eine nähere Untersuchung zeigt.

3.1. *Wahl des Ablehnungsbereichs*

Wir betrachten zunächst den Fall, in dem man eine einfache statistische Hypothese, wie sie oben beschrieben wurde, mit einer einzigen Beobachtung der Größe X konfrontiert (Fisher 1970: 43ff). X sei normalverteilt mit Mittelwert Null und Varianz σ^2. Nach Fisher wird man ein Signifikanzniveau von 0.05 oder 0.01 wählen, jedenfalls einen konventionell vorgegebenen kleinen Wert. Der Ablehnungsbereich wird durch pragmatische Überlegungen bestimmt. Allgemein wählt man ein Maß der Diskrepanz zwischen Hypothese und Beobachtug, das alle laut Hypothese möglichen Beobachtungen in eine solche Ordnung bringt, daß eine höhere Diskrepanz stärkere Evidenz gegen die Hypothese bedeutet. Das Diskrepanzmaß beruht also auf einer Vorstellung darüber, welche Beobachtungen in einem bestimmten Experiment tendenziell gegen die Hypothese sprechen. Dieses Diskrepanzmaß ist rein ordinal; ob ein beobachteter Wert tolerierbar ist oder nicht, hängt von der Wahrscheinlichkeit ab, gleichstarke oder stärkere Evidenz gegen die Hypothese zu finden. Ist die Diskrepanz so groß, daß die Wahrscheinlichkeit, eine solche oder eine größere Diskrepanz zu beobachten, höchstens gleich dem Signifikanzniveau ist, wird die Diskrepanz als signifikant auf dem gewählten Niveau bezeichnet und die Hypothese als vorläufig widerlegt betrachtet. Kurz: Ablehnung erfolgt, wenn die Daten eine unwahrscheinlich starke Evidenz gegen die Hypothese liefern (Spielman 1974: insbes. 216ff). Die Frage, welches Signifikanzniveau gewählt werden sollte, ist eher nebensächlich. Statt ein Signifikanzniveau vorzugeben, kann man den sogenannten P-Wert ermitteln; das ist das Signifikanzniveau, bei dem die Hypothese, gegeben die Beobachtungen, gerade abgelehnt würde. Der P-Wert ist ein Maß der Stärke der Evidenz gegen die Hypothese; konventionell vorgegebene Werte, ab denen eine Hypothese abgelehnt werden sollte, sind nichts als unverbindliche Empfehlungen.

Im Falle einer unimodalen symmetrischen Dichtefunktion, wie sie sich aus der Normalverteilung ergibt, wählt man als Maß der Diskrepanz häufig die absolute Abweichung des beobachteten vom theoretischen Mittelwert. Der resultierende Test besteht darin, daß man „die Schwänze der Verteilung abschneidet": Man wählt den Ablehnungsbereich symmetrisch und im Bereich der niedrigsten Dichte so, daß das gewählte Signifikanzniveau gerade eingehalten wird. Diese Überlegungen lassen sich auf den Fall zweier Beobachtungen x_1 und x_2 übertragen. Die Wahl einer Teststatistik und die Wahl eines Ablehnungsbereichs für die Verteilung der Teststatistik bestimmen zusammen den Ablehnungsbereich im Stichprobenraum, dem Definitionsbereich der gemeinsamen Verteilung aller Stichprobenelemente. Ist man etwa vor allem an der Frage interessiert, ob der Mittelwert tatsächlich Null ist, wird man den standardisierten Mittelwert der Beobachtungen $m = (x_1 + x_2) / \sqrt{2}\sigma$ als Teststatistik wählen; m ist standardnormalverteilt

um Null. Man kann jetzt bei der Verteilung von m die Schwänze abschneiden und einen zweiseitigen Ablehnungsbereich mit einem Signifikanzniveau von 0.05 bilden; damit dient der absolute Wert des Stichprobenmittels als Maß der Diskrepanz. Aus der Tabelle entnimmt man, daß die Hypothese abgelehnt werden sollte, wenn m absolut größer als 1.96 ist.

Diesen Ablehnungsbereich kann man sich auch im Stichprobenraum ansehen. Die in Abb. 1a eingezeichnete Gerade durch den Ursprung mit Steigung -1 gibt die zweidimensionalen Stichproben an, deren Mittelwert gerade 0 ist; der um diese Gerade eingezeichnete Streifen zeigt alle Stichproben, für die der standardisierte Mittelwert im angegebenen Intervall liegt. Fällt die Stichprobe aus diesem Streifen heraus, wird die Hypothese abgelehnt (Gauss-Test). Die gemeinsame zweidimensionale Normalverteilung der Stichprobenwerte kann in der abgebildeten Ebene durch die Linien gleicher Dichte dargestellt werden; diese Linien bilden konzentrische Kreise um den Ursprung, wobei die Dichte mit dem Abstand zum Nullpunkt abnimmt. Der gewählte Ablehnungsbereich wirkt im Verhältnis zu dieser symmetrischen Dichtefunktion völlig beliebig.

Selbst wenn der Experimentator aus theoretischen Gründen vor allem am Mittelwert interessiert ist, folgt daraus nicht, daß er allein Abweichungen des beobachteten vom theoretischen Mittelwert ernst nehmen muß. So könnte eine zu hohe Streuung bei akzeptablem Mittelwert daraus resultieren,

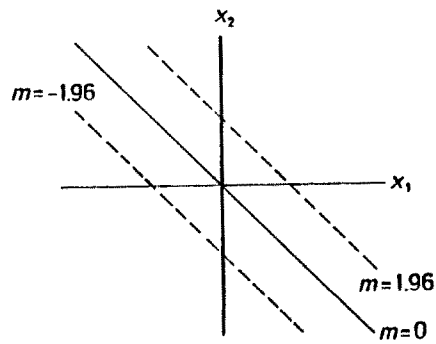

Abb. 1a. Gauss-Test

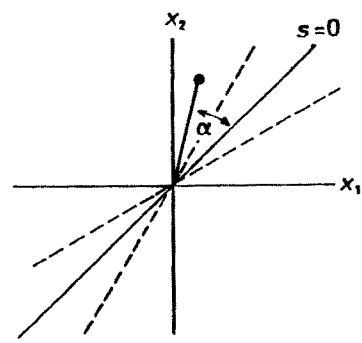

Abb. 1b. t-Test

daß die erste Beobachtung eine Realisation einer Normalverteilung mit Mittelwert +10, die zweite Beobachtung eine Realisation einer Normalverteilung mit Mittelwert −10 darstellt. Die Hypothese eines einheitlichen Mittelwerts wäre also völlig falsch; dies ließe sich aber in diesem Fall am Mittelwert der Beobachtungen nicht erkennen. Das zeigt, daß der Ablehnungsbereich durch Überlegungen des Experimentators bestimmt ist, die – bei Fisher jedenfalls – weitgehend im Dunkeln bleiben; der Ablehnungsbereich ist methodologisch gesehen willkürlich. Fishers Theorie ist mit dem deduktiv-nomologischen Modell unvereinbar, denn dieses Modell fordert, daß angesichts der Beobachtungen die Entscheidung über die Hypothese eindeutig ausfällt. Für eine falsifikationistische Interpretation des Signifikanztests wäre eine methodologische Regel notwendig, die ein kanonisches Maß der Diskrepanz zwischen Hypothese und Beobachtung festlegt.[2]

3.2. *Zusammengesetzte Hypothesen*

Weitere Probleme ergeben sich im Falle von Tests zusammengesetzter Hypothesen. Eine solche Hypothese wäre zum Beispiel, daß die Größe X normalverteilt ist mit Mittelwert Null und unbekannter Varianz. Diese Hypothese kann man als eine Disjunktion unendlich vieler Hypothesen auffassen, wobei jedes Glied der Disjunktion eine einfache Hypothese ist, wie sie soeben diskutiert wurde. Die einzelnen Glieder der Disjunktion unterscheiden sich nur durch die jeweils unterstellte Varianz der Normalverteilung. Wir nehmen an, es bestünde wie im obigen Fall auch hier ein besonderes Interesse am Mittelwert der Beobachtungen. Das legt es nahe, den t-Test durchzuführen, der als Verallgemeinerung des Gauss-Tests angesehen werden kann. Nach Fisher (1959: 79ff) schätzt man zu diesem Zweck die unbekannte Varianz durch die Stichprobenvarianz s^2 und bildet mit Hilfe dieser Schätzgröße das Analogon zum standardisierten Stichprobenmittelwert, im Falle von $n = 2$ also $t = (x_1 + x_2)/\sqrt{2}s$. Der absolute Wert von t ist das Maß der Diskrepanz zwischen Beobachtung und Hypothese. Die Rechtfertigung für die Betrachtung von t liegt in der Analogie zwischen dem resultierenden Test und dem Test für eine einfache Normalverteilungshypothese. Fisher hat nie ein allgemeines Prinzip für die Konstruktion von Tests zusammengesetzter Hypothesen formuliert; seinen Ausführungen zum t-Test und zum Chi-Quadrat-Unabhängigkeitstest scheint jedoch folgende Vorgehensweise zugrundezuliegen.[3]

(1) Bestimme eine Teststatistik für den Fall, in dem der Parameterwert bekannt ist.
(2) Wähle einen Schätzer für den unbekannten Parameter und ersetze diesen Parameter in der gewählten Teststatistik durch seinen Schätzer.
(3) Bestimme die Verteilung der resultierenden modifizierten Teststatistik unter Berücksichtigung der Verteilung des Schätzers.
(4) Benutze diese modifizierte Teststatistik für den Test der zusammengesetzten Hypothese.

Die Begründung für die Idee, die Teststatistik nicht mit Hilfe eines *Schätzwertes* für den unbekannten Parameter, sondern mit Hilfe eines *Schätzers* zu konstruieren, ist, daß auf diese Weise der Schätzfehler berücksichtigt werde (z.B. Fisher 1929: 55, 1970: 118). Die Teststatistik ist also ein Amalgam aus Prüfgröße und Schätzung. Zwar sind die formalen Eigenschaften einer solchen Größe in den betrachteten Fällen ohne weiteres herleitbar, aber ihre Bedeutung ist doch sehr unklar. Eine genauere Begründung für diese Vorgehensweise ist bei Fisher nicht zu finden.[4]

Betrachten wir nun die Eigenschaften des t-Tests. Die aus der oben beschriebenen Verfahrensweise resultierende Größe t gehorcht einer t-Verteilung mit einem Freiheitsgrad; ein Blick auf die Tabelle ergibt, daß ab einem t-Wert von ± 12.71 die Beobachtung signifikant auf dem Niveau 0.05 ist. Wieder kann man sich den resultierenden Ablehnungsbereich im Stichprobenraum ansehen (Abb. 1b oben). Es handelt sich dabei um einen Doppelkegel, punktsymmetrisch zum Ursprung und symmetrisch zu der eingezeichneten Geraden mit Steigung +1.

Eine geometrische Interpretation zeigt klarer als Fishers Rationalisierung, worauf der t-Test beruht. Wir betrachten dazu den Strahl durch den Ursprung, auf dem die Stichprobe liegt. Dieser Strahl bildet einen Winkel α mit der Geraden durch den Ursprung, die die Steigung 1 hat; auf dieser Geraden liegen alle Stichproben mit einer Standardabweichung von Null. Wenn wir den Winkel α als Zufallsvariable betrachten, so ist auf Grund der Symmetrie der Normalverteilung klar, daß α gleichverteilt im Intervall $[0,2\pi]$ ist. Die Größe t ist gerade gleich dem Kotangens von α. Die Willkürlichkeit dieses Tests ist wieder auffällig. Neyman benutzte die Gleichverteilung von α, um Pseudo-t-Tests zu konstruieren, die darauf beruhen, daß man α ausgehend von einer beliebig gewählten anderen Geraden mißt; er konfrontierte die Fishersche Testtheorie mit dem Problem, ein Kriterium für die Wahl zwischen diesen Tests zu formulieren.[5] Die Auswahl einer anderen Geraden läßt sich auch durch eine orthogonale Transformation der Daten bewerkstelligen: man rotiert das Koordinatenkreuz und mißt den entsprechenden Winkel im neuen Koordinatensystem. Das Ergebnis, Neymans Paradox, ist für Anhänger des Signifikanztestes unerfreulich: Eine logisch äquivalente Reformulierung von Daten und Hypothese mit nachfolgender Anwendung des t-Tests kann zur Ablehnung führen, obwohl eine Anwendung des t-Tests auf die ursprüngliche Hypothese und die ursprünglichen Daten nicht zur Ablehnung führt.

Zwei weitere Probleme bei Fishers Behandlung zusammengesetzter Hypothesen müssen diskutiert werden: zum einen verletzt Fishers Theorie eine wichtige Adäquatheitsbedingung, zum anderen ist Fisher an einem zentralen Punkt seiner Theorie inkonsistent.

Fisher (1959: 49) sah in einem signifikanten t-Wert einen Nachweis dafür, daß keine der Hypothesen in der getesteten Disjunktion akzeptabel sei. Das ist jedoch offensichtlich nicht der Fall. Wenn wir berücksichtigen, daß bei einem Test anhand des Mittelwerts der Akzeptanzbereich für eine ein-

fache Hypothese beliebig breit wird, wenn die Varianz nur hoch genug ist, ergibt sich sofort: Zu jedem Punkt im Ablehnungsbereich des *t*-Tests existieren beliebig viele Normalverteilungen um Null, deren Akzeptanzbereiche diesen Punkt einschließen. Fishers Testtheorie verletzt damit die sogenannte Konsequenzbedingung,[6] die zwei logisch äquivalente Formulierungen besitzt:

(1) Wenn F eine logische Folgerung aus H ist und wenn die Beobachtung B F falzifiziert, dann falsifiziert B auch H (Baird 1983: 106).

(2) Wenn F eine logische Folgerung aus H ist und wenn die Beobachtung B H nicht falsifiziert, dann falsifiziert B auch F nicht (Baird 1983: 115).

Wir haben gefunden, daß beim *t*-Test immer unendlich viele Normalverteilungshypothesen nicht widerlegt sind. Trotzdem kann nach Fisher die zusammengesetzte Hypothese durch den *t*-Test widerlegt werden. Die Fishersche Theorie des Signifikanztests ist daher im Falle zusammengesetzter Hypothesen mit der Konsequenzbedingung unvereinbar, denn aus jeder nicht falsifizierten Normalverteilungshypothese folgt logisch die zusammengesetzte Hypothese.

Man sollte allerdings im Zusammenhang mit der Konsequenzbedingung beachten, daß es sich um ein *methodologisches* Prinzip handelt: es geht um Falsifikation, nicht um Falschheit. Keine statistische Testtheorie ist mit der Konsequenzbedingung völlig vereinbar; das ist eine Folge der Irrtumswahrscheinlichkeit bei der Ablehnung von Hypothesen. Ein einfaches Beispiel illustriert diesen Punkt.

Eine Hypothese H postuliere eine Gleichverteilung über n Elementarereignisse. Diese Hypothese hat dann n gleichartige Folgerungen F_i, die man dadurch erhält, daß man das i-te Elementarereignis herausgreift und die Hypothese formuliert, mit Wahrscheinlichkeit $1/n$ ereigne sich das herausgegriffene Elementarereignis, mit Wahrscheinlichkeit $1-1/n$ ereigne sich dieses Elementarereignis nicht. Wenn n groß ist, würde man bei isolierter Betrachtung der Folgerungen die Folgerung F_i ablehnen, wenn das Elementarereignis i stattfindet. Es wäre jedoch absurd, eine solche Falsifikation auf die Hypothese H übertragen zu wollen, da die Hypothese in diesem Fall durch jede Beobachtung falsifiziert würde.

Eine durchgängige Anwendung der Konsequenzbedingung führt somit zu unsinnigen Ergebnissen. Das gilt allerdings nicht für eine abgeschwächte Form der Konsequenzbedingung, das „disjunktive Testkriterium" (Baird 1983). Das Kriterium fordert, den Test einer zusammengesetzten Hypothese als Test einer Disjunktion einfacher Hypothesen aufzufassen und die zusammengesetzte Hypothese so lange als nicht falsifiziert zu betrachten, wie eine Gliedhypothese gefunden werden kann, die für sich betrachtet nicht falsifiziert wurde. Dieses Kriterium erlaubt es, für unbekannte Parameter einfach den Wert einzusetzen, der mit den Daten am besten vereinbar ist. Das Prinzip wurde zuerst von Bowley & Connor in einer Kritik an Fishers Variante des Chi-Quadrat-Unabhängigkeitstests formuliert:

Now if the question is ... whether or not the observations are consistent with a law of a given form but with parameters at choice (as we might ask whether or not a planet moved in an ellipse, the constants of the ellipse *a priori* unknown), we are free to choose those parameters as to make the discrepancy between the formula and the observations a minimum ... (Bowley & Connor 1923: 3)[7]

M. E. sollte man das disjunktive Testkriterium nicht ohne Not aufgeben, da man mit Hilfe dieses Kriteriums die Testtheorie vollständig auf Tests einfacher Hypothesen gründen kann. Nach diesem Kriterium spielen Schätzungen für den Test keine Rolle. Aus wissenschaftstheoretischer Perspektive ist schon der Status einer eigenständigen Schätztheorie äußerst unklar; diese Unklarheit wird durch Fishers Amalgamierung von Test- und Schätztheorie erhöht. Selbst wenn man nicht von vorneherein die durchaus akzeptable Strategie verfolgt, so wenige falsifikationistische Prinzipien wie möglich aufzugeben, besteht genügend Anlaß, Fishers Theorie zugunsten des disjunktiven Testkriteriums aufzugeben.

Es bleibt zu zeigen, daß Fishers Testtheorie nicht konsistent ist.* Der Nachweis ist einfach. Fisher lehnte es ab, bei Signifikanztests explizit auf Alternativhypothesen bezugzunehmen. Seine Behandlung des Chi-Quadrat-Unabhängigkeitstests trägt dem Rechnung: bei der Schätzung der unbekannten Parameter wird vorausgesetzt, daß die zu testende Hypothese wahr ist. Students t-Test dagegen beruht darauf, daß die unbekannte Varianz durch s^2 geschätzt wird, also durch einen Schätzer, der nur bei unbekanntem Mittelwert akzeptabel ist, obwohl die Hypothese beim t-Test den Mittelwert spezifiziert. Der richtige Schätzer für die unbekannte Varianz bei einem hypothetischen Mittelwert von Null ist das Mittel der quadrierten Stichprobenwerte. Die entsprechend korrigierte Teststatistik ist nicht mehr t-verteilt. Im Gegensatz zur NPT erklärt also Fishers Theorie den t-Test *nicht*.

3.3. *Stoppregeln*

Eines der verwirrendsten Probleme des Signifikanztestes ergibt sich aus der möglichen Verwendung von Stoppregeln (s. z.B. Hacking 1965: 107ff, Berger & Berry 1988). Betrachten wir folgenden Fall. Eine einfache Hypothese soll anhand mehrerer Beobachtungen getestet werden. Wir unterstellen, daß ein geeigneter Signifikanztest ausgewählt wurde, der sich bei jeder Stichprobengröße verwenden läßt. Wir können dann zwei Vorgehensweisen unterscheiden.

(1) *Normales Experiment:* Es wird eine vorher festgelegte Zahl n von Beobachtungen gemacht. Dann wird gemäß dem ausgewählten Signifikanztest der P-Wert berechnet. Die Hypothese soll als falsifiziert betrachtet werden, wenn ein P-Wert von α erreicht oder unterschritten wird.

(2) *Stoppregelexperiment:* Es werden solange Beobachtungen gesammelt, bis *gemäß dem unter (1) verwendeten Test* für die Gesamtheit der

* Hier muß ich mich korrigieren: Fisher ist nicht inkonsistent. Der Test, der sich bei Verwendung des richtigen Schätzers ergibt, ist äquivalent zum t-Test.

bisherigen Beobachtungen ein P-Wert von α erreicht oder unterschritten ist. Die Zahl der Beobachtungen, die Stichprobengröße, ist also eine Zufallsvariable.

Wir nehmen an, daß tatsächlich genau n Beobachtungen gemacht werden und daß mit der letzten Beobachtung gerade ein P-Wert von α erreicht oder unterschritten wird, während vor der letzten Beobachtung der P-Wert immer über α lag. Für das normale Experiment ergibt sich eine klare Widerlegung der Hypothese. Die Frage ist: Muß man das Ergebnis anders interpretieren, wenn man erfährt, daß in Wirklichkeit das Stoppregelexperiment durchgeführt wurde? Schließlich hätte unter diesen Umständen der P-Wert schon vor dem Experiment festgestanden und kann also intuitiv kaum eine unwahrscheinlich starke Evidenz gegen die Hypothese anzeigen.

Wir müssen hier zwei verschiedene Probleme unterscheiden. Zum einen scheint man das Ergebnis jedes Tests, der auf einer fixen Stichprobengröße beruht, durch *heimliche* Verwendung einer Stoppregel manipulieren zu können. Es ist ein bekanntes Ergebnis, daß man mit Hilfe der oben angegebenen Stoppregel eine wahre Hypothese garantiert widerlegen kann, wenn man nur ausdauernd genug ist: die Wahrscheinlichkeit ist 1, daß ein beliebiger vorgegebener P-Wert mit einer endlichen Stichprobe erreicht oder unterschritten wird, wenn die Hypothese wahr ist.[8]

Zum anderen ist nicht klar, welches die relevante Verteilung ist, die für einen Test herangezogen werden soll, wenn *bekannt* ist, daß eine Stoppregel verwendet wurde. Soll man die verwendete Stoppregel berücksichtigen, indem man die Verteilung der Stichprobengröße bei vorgegebenem P-Wert betrachtet, oder ignorieren, indem man den Signifikanztest durchführt, den man auch bei fixer Stichprobengröße durchführen würde? Die beiden Tests führen in der Regel zu verschiedenen Schlußfolgerungen.

Da mir die Fishersche Theorie keine Lösungen für diese Probleme zu bieten scheint, will ich sie von vorneherein aus einer falsifikationistischen Perspektive diskutieren.

Manipulierbarkeit: Das Problem der Manipulierbarkeit taucht m.E. dann nicht auf, wenn man mit Popper verlangt, daß Falsifikationen reproduzierbar sind. Eine Falsifikation erfordert dann eine Hypothese mit geringem theoretischen Gehalt – falsifizierende Hypothese genannt –, die beschreibt, wie man die in Frage stehende Hypothese widerlegen kann: letztere ist widerlegt, wenn sich die falsifizierende Hypothese bei Überprüfungen, also Wiederholungen des falsifizierenden Experiments, bewährt (Popper 1984: 54). Im deterministischen Kontext gibt die falsifizierende Hypothese ein Experiment an, mit dem man mit Sicherheit eine Widerlegung erzielt. Im statistischen Kontext müssen wir verlangen, daß durch das Experiment sehr häufig oder in der Regel eine Widerlegung erzielt wird. Wenn nun jemand heimlich eine Stoppregel benutzt und dadurch die Hypothese mit n Versuchen auf dem Niveau α widerlegt, wird er als falsifizierende Hypothese die Behauptung aufstellen müssen, mit höchstens n Versuchen sei *in der Regel* eine Falsifikation der in Frage stehenden Hypothese auf dem an-

gegebenen Niveau zu erreichen. Diese falsifizierende Hypothese ist natürlich selbst keine exakte statistische Hypothese – eine exakte Wahrscheinlichkeit für Falsifikationen könnte sich nur aus einer Alternativhypothese ergeben. Sie kann aber trotzdem sehr leicht überprüft werden: die Falsifikation wird sich nur sehr selten und nicht in der Regel reproduzieren lassen, wenn die angeblich falsifizierte Hypothese wahr ist. Eine Manipulierbarkeit ergibt sich also nur dort, wo man es versäumt, Experimente zu replizieren.

Berücksichtigung der Stoppregel: Auch wenn die Manipulierbarkeit kein Problem darstellt, bleibt unklar, wie die Ergebnisse eines *offen* durchgeführten Stoppregelexperiments zu deuten sind. Wenn man die tatsächlich verwendete Stoppregel berücksichtigen will, hängt die Interpretation der Ergebnisse davon ab, was sich der Experimentator während der Stichprobenerhebung denkt. An den Daten läßt sich ja nicht ablesen, ob ein normales Experiment oder ein Stoppregelexperiment durchgeführt wurde. Eine Berücksichtigung der Stoppregel wäre mit dem Falsifikationismus unvereinbar, denn der fordert, daß die Kenntnis von Daten und Hypothese ausreichen muß, um festzustellen zu können, ob eine Falsifikation vorliegt. Damit ist es für eine falsifikationistische Interpretation von Signifikanztests notwendig, ein Argument zu finden, das es erlaubt, Stoppregeln zu ignorieren. Die folgende logische Analyse liefert die Grundlagen für ein solches Argument, das ich dann im Rahmen meines eigenen Vorschlags zur Testtheorie formulieren werde.

Die bei Berücksichtigung der Stoppregel herangezogene Verteilungshypothese bezieht sich zunächst auf eine unendliche Folge von Beobachtungen; die Hypothese gibt für jedes Glied der Folge die Wahrscheinlichkeit an, daß für dieses Glied zum ersten Mal in der Folge die in der Stoppregel festgelegte Abbruchbedingung erfüllt ist. Eine solche Folge ist jedoch nicht beobachtbar. Für eine gegebene Menge von n Beobachtungen ist noch nicht einmal klar, zu welcher unendlichen Folge die Beobachtungen gehören. Selbst wenn man annähme, daß die Beobachtungen die ersten n Glieder einer Folge darstellen, steht damit noch nicht fest, welche der Beobachtungen das erste (welche das zweite, etc.) Element dieser Folge ist. Im Gegensatz zur Häufigkeitsinterpretation geht die Propensity-Interpretation bei der Definition der Wahrscheinlichkeiten nicht von Beobachtungsfolgen aus; damit läßt sich keine Anordnung von Beobachtungen als kanonische Anordnung auszeichnen, auch die Anordnung nach der zeitlichen Abfolge nicht.

Ohne weitere Informationen kann man also eine Hypothese, wie sie sich für Stoppregelexperimente ergibt, nicht testen. Die fehlende Information ist die Zuordnung einer Menge von Beobachtungen zu den Elementen der Folge, auf die sich die Hypothese bezieht. Diese Information kann nur durch die Einführung einer zusätzlichen Randbedingung in der Wenn-Komponente der Hypothese gegeben werden. Diese Randbedingung muß festlegen, wie man die relevante Folge von Beobachtungen mit Hilfe der Einzelexperimente erzeugt. Wird dagegen keine solche zusätzliche Rand-

bedingung eingeführt, kann man beliebige Beobachtungen in beliebiger Reihenfolge zu einer Folge zusammenstellen und dann prüfen, bei welchem Element dieser „künstlichen" Folge die Abbruchbedingung erfüllt war. Damit existiert jedoch bei einer gewissen Menge von Beobachtungen zu jeder Folge, in der die Abbruchbedingung überraschend früh erfüllt war, eine, in der sie überraschend spät – wenn überhaupt – erfüllt war. Solche unterschiedlichen Folgen lassen sich durch Umordnung der Elemente erzeugen. Nur durch die zusätzliche Randbedingung kann die Verteilungshypothese des Stoppregelexperiments in eine prüfbare Gesetzeshypothese überführt werden.

Während die Hypothese im normalen Experiment durch n beliebige Beobachtungen geprüft werden kann, solange diese nur aus korrekt durchgeführten Experimenten stammen, läßt sich die Hypothese im Stoppregelexperiment nur durch Beobachtungen prüfen, die aus korrekt durchgeführten Experimenten *innerhalb eines korrekt durchgeführten Gesamtexperiments zur Erzeugung einer Beobachtungsfolge* stammen. Die Hypothese im Stoppregelexperiment ist eine Abschwächung der ursprünglichen Hypothese. Dieses Ergebnis werde ich später heranziehen, um die Berücksichtigung von Stoppregeln auszuschließen.

3.4. *Folgerungen für eine falsifikationistische Testtheorie*

Das deduktiv-nomologische Modell erfordert, daß bei gegebenem Signifikanzniveau der Ablehnungsbereich eindeutig bestimmt ist. Läßt sich dagegen der Ablehnungsbereich immer so wählen, daß er wahlweise das beobachtete Ereignis enthält oder nicht, bleibt als einzige rationale Vorgehensweise, irgendeinen Ablehnungsbereich zu wählen, ihn aber wenigstens vor der Beobachtung festzulegen, wie dies Gillies (1973) vorschlägt. Damit bliebe aber ein großer Teil des Falsifikationismus auf der Strecke. Die statistische Hypothese ließe sich zu einer Erklärung oder Prognose im herkömmlichen Sinn nicht mehr verwenden; die Entscheidung über eine Hypothese ließe sich nicht mehr durch Gegenüberstellung von Hypothese und vorhandenen Daten treffen, da die Daten *per se* nichts über die Hypothese sagen. Somit wäre es unmöglich, Hypothesen zu konstruieren, die den bekannten Daten besser entsprechen als bereits widerlegte Hypothesen.

In Fishers Theorie fehlt eine Regel zur eindeutigen Bestimmung des Ablehnungsbereiches. Das Testproblem kann allerdings aus falsifikationistischer Sicht auf den Fall einfacher Hypothesen reduziert werden, indem man das disjunktive Testkriterium einführt. Eine falsifikationistische Testtheorie müßte zusätzlich (i) eine Regel für die Wahl des Ablehnungsbereiches und (ii) einen Grund für die Ignorierung von Stoppregeln liefern. Der folgende Abschnitt wendet sich dem ersten Problem zu; er geht der Frage nach, ob die NPT geeignet ist, das Problem im Sinne des Falsifikationismus zu lösen.

4. DIE NEYMAN-PEARSON-THEORIE

Die Tatsache, daß Fishers Testtheorie keine klare Regel für die Wahl des Ablehnungsbereichs liefert, nahmen Neyman und Pearson zum Anlaß, die Fishersche Testtheorie in ganz entscheidender Weise zu modifizieren. Im folgenden wird zunächst die Bedeutung des sogenannten Fehlers 2. Art und dann die Rolle von Alternativhypothesen erörtert. Dem schließt sich eine Diskussion der entscheidungstheoretischen Interpretation der NPT an.

4.1. *Der Fehler zweiter Art*

In Fishers Theorie dient die ausgewählte Teststatistik als Maß der Abweichung der beobachteten Häufigkeitsverteilung von der theoretischen Verteilung. Man könnte auch von einer Abweichung der gezogenen von einer in irgendeinem Sinne „typischen" Stichprobe sprechen. In der NPT wird die jeweils interessante Abweichung bestimmt, indem man Alternativen zu der zu testenden Hypothese in Betracht zieht. Besteht die einzige denkbare Alternativhypothese z.B. darin, daß die Verteilung dieselbe Varianz, aber einen höheren Mittelwert aufweist, so ist es einleuchtend, daß „untypische" Varianzen weniger interessant sind als zu hohe Mittelwerte. Diese Idee scheint grundsätzlich auch mit Fishers Vorstellungen vereinbar zu sein. Er schreibt:

> In choosing the grounds upon which a general hypothesis should be rejected, the experimenter will rightly consider all points on which, in the light of current knowledge, the hypothesis may be imperfectly accurate, and will select tests, so far as possible, sensitive to these possible faults, rather than to others. (Fisher 1959: 47)

Allerdings hält Fisher (1959: 70) im Falle *präzis* formulierter Alternativen Signifikanztests für ein ungeeignetes Mittel der Auswahl; er vertritt für solche Problemstellungen die Likelihood-Theorie (vgl. vor allem Hacking 1965).

Nach der NPT bestimmt man den Ablehnungsbereich einer Hypothese H_1 so, daß die Wahrscheinlichkeit maximal ist, H_1 auch wirklich abzulehnen, wenn eine bestimmte Alternativhypothese H_2 wahr sein sollte. Diese Wahrscheinlichkeit heißt „Macht" des Tests. Die Macht eines Tests wird zumeist als Komplement der Wahrscheinlichkeit β des sogenannten Fehlers 2. Art eingeführt. Das ist die Wahrscheinlichkeit, H_1 zu akzeptieren, obwohl H_2 wahr ist. Offensichtlich ist es gleich, ob man nun die Macht $(1\text{-}\beta)$ maximiert oder die Wahrscheinlichkeit eines Fehlers 2. Art (β) minimiert.

Man kann die NPT übersichtlich darstellen, indem man die Entscheidungsmöglichkeiten und die möglichen Zustände in einem Schema (Abb. 2) aufzeichnet.

Die zu testende Hypothese H_1 kann wahr oder falsch sein; ein Test von H_1 kann dazu führen, daß H_1 vorläufig akzeptiert oder abgelehnt wird. Wenn man die Möglichkeit falscher Beobachtung einmal ausschließt, so ist klar, daß der Fehler 1. Art (irrtümliche Ablehnung einer wahren Hypothese) nur bei statistischen Hypothesen vorkommt. Bei einer Irrtumswahr-

	Oberhyp. akzeptiert H_1 akzeptiert H_2 abgelehnt	 H_1 abgelehnt H_2 akzeptiert	Oberhyp. abgelehnt H_1 abgelehnt H_2 abgelehnt
H_1 wahr H_2 falsch	korrekt $(1-\alpha)$	F. 1. Art (α)	Fehler
H_1 falsch H_2 wahr	F. 2. Art (β)	korrekt $(1-\beta)$	(0)
H_1 falsch H_2 falsch	Fehler 3. Art (1)		korrekt (0)

Abb. 2. Die Fehler 1., 2., und 3. Art in der NPT.

scheinlichkeit von α ist die Wahrscheinlichkeit, daß eine wahre Hypothese H_1 akzeptiert wird, gleich $1-\alpha$. Sollte H_1 tatsächlich falsch sein, läßt sich zwischen korrekter Ablehnung und irrtümlicher Annahme (Fehler 2. Art) unterscheiden. Der Fehler 2. Art kann natürlich genauso bei deterministischen Hypothesen auftreten. Dieser Fehler ist es gerade, der zu der bekannten Asymmetrie zwischen Falsifikation und Verifikation führt.

Ohne weitere Voraussetzungen kann man über die Wahrscheinlichkeit eines Fehlers 2. Art nichts sagen. Die NPT beruht darauf, daß eine solche Voraussetzung gemacht wird: Für jeden Test wird eine Oberhypothese (Stegmüller 1973) vorausgesetzt, die besagt, daß entweder H_1 oder eine Alternativhypothese H_2 wahr ist. Unter dieser Voraussetzung impliziert die Falschheit von H_1 die Wahrheit von H_2; dem Fehler 2. Art läßt sich jetzt für jeden Ablehnungsbereich eine Wahrscheinlichkeit β zuweisen. Hat man die Oberhypothese akzeptiert, erscheint es vernünftig, den Ablehnungsbereich so zu wählen, daß bei gegebener Wahrscheinlichkeit für den Fehler 1. Art der Fehler 2. Art möglichst unwahrscheinlich wird. Die Oberhypothese impliziert ja, daß dies die beiden einzigen möglichen Fehler sind.

In vielen Fällen wird man jedoch nicht bereit sein, die Oberhypothese als durch Beobachtung unerschütterlich zu betrachten. Man muß also zusätzlich die Möglichkeit betrachten, daß die Oberhypothese falsch sein könnte. Als weitere Entscheidungsmöglichkeit kann man die Oberhypothese ablehnen, wenn man weder H_1 noch H_2 akzeptieren will. Über die Wahrscheinlichkeit eines Fehlers 3. Art (irrtümliche Akzeptanz der Oberhypothese)[9] läßt sich nichts sagen – außer in einem Fall: wenn man, wie die NPT es vorschreibt, diese Fehlermöglichkeit vollkommen ignoriert. Dann ist die Wahrscheinlichkeit, die Oberhypothese irrtümlich zu akzeptieren, wenn sie falsch ist, gleich 1.

Die Auswahl einer in irgendeinem Sinne besten Hypothese aus einer gegebenen Menge bezeichnet man üblicherweise als Schätzung. Die NPT reduziert Testen auf eine spezielle Form des Schätzens (vgl. Neyman 1965: 448). Wenn man im deterministischen Bereich den Falsifikationismus akzeptiert, wird man sich im statistischen Bereich auf eine solche Resig-

nationslösung nicht einlassen wollen, bevor gezeigt wurde, daß die Voraussetzung einer Oberhypothese logisch unausweichlich ist.

4.2. *Die Rolle von Alternativhypothesen*

Immerhin läßt sich zugunsten der NPT einwenden, daß unter Zuhilfenahme der Alternativhypothese der Ablehnungsbereich eindeutig festgelegt wird. Was spricht dagegen, diesen Ablehnungsbereich im Sinne des deduktiv-nomologischen Modells als empirischen Gehalt zu interpretieren? Der entscheidende Einwand gegen eine solche Betrachtung läßt sich am ehesten verdeutlichen, wenn man unterstellt, die zu testende Hypothese sei wahr. Der Akzeptanzbereich und damit die Vorhersagen dieser *wahren* Hypothese hingen dann davon ab, welche *falsche* Hypothese man als Alternative heranzieht. Was ist für diesen Zweck die richtige falsche Alternative? Die NPT ist den gleichen Einwänden ausgesetzt wie Fishers Theorie: die Wahl der Alternativen und damit die Wahl des Ablehnungsbereichs erfolgt praktisch willkürlich. Selbst wenn man für Testzwecke mit dem so gewählten Ablehnungsbereich zufrieden ist, kann man ihn nicht für Prognosen oder Erklärungen verwenden.

Gegen diese Position könnte man einwenden, daß die Alternativhypothesen nicht notwendigerweise willkürlich festgelegt werden müssen. Zur willkürlichen Auswahl gibt es zwei Alternativen: konventionelle Festsetzung der Alternativhypothesen oder Heranziehung substantieller, also theoretisch interessanter Alternativhypothesen. Beide Möglichkeiten bieten jedoch keinen echten Ausweg.

Konventionell festgelegte Alternativhypothesen sind nur eine Camouflage für eine konventionelle Festlegung des richtigen Signifikanztests. Statt das Problem der rationalen Wahl von Dummyalternativen zu betrachten, kann man das Problem der rationalen Wahl eines Ablehnungsbereichs für isolierte Hypothesen auch direkt angehen.

Die Heranziehung substantieller Alternativen bietet natürlich keinen Ausweg für das Erklärungs- und Prognoseproblem, scheint aber wenigstens unter Testgesichtspunkten als statistische Variante eines Entscheidungsexperiments gerechtfertigt werden zu können. In diesem Zusammemhang wird manchmal auf Poppers Ansicht verwiesen, Alternativhypothesen seien ein entscheidender Bestandteil eines Tests (Stegmüller 1973: 15f). Dieser Zusammenhang zwischen NP-Tests und Entscheidungsexperimenten ist jedoch sehr oberflächlich. Ein Entscheidungsexperiment im deterministischen Fall setzt voraus, daß die beiden Theorien für ein und dieselbe Situation unterschiedliche, jeweils für sich prüfbare Voraussagen machen. Auch in einem Entscheidungsexperiment wird eine isolierte Hypothese falsifiziert und eine andere isolierte Hypothese bestätigt. Es ist richtig, daß Popper häufig die Wichtigkeit von Entscheidungsexperimenten betont und daß man an einigen Stellen den Eindruck bekommt, Entscheidungsexperimente seien vielleicht sogar der Regelfall (Popper 1973: 25ff). Dies heißt aber nur, daß

die Tests ihre Bedeutung dadurch gewinnen, daß sie – möglicherweise – eine Entscheidung zwischen konkurrierenden Theorien herbeiführen können; es bedeutet *nicht*, daß die Logik des Tests in irgendeiner Weise berührt ist.

Die Entscheidung zwischen Alternativhypothesen ist außerdem nicht die einzige Funktion von Tests. Substantielle Alternativhypothesen können im allgemeinen nur unter hohen Kosten entwickelt werden. Ein wesentlicher Anstoß für die Entwicklung solcher Hypothesen ist die Feststellung, daß alle existierenden Hypothesen empirisch unbefriedigend sind. Ist es unmöglich, eine solche Feststellung zu treffen – und dies wäre der Fall, wenn man nur substantielle Hypothesen gegeneinander testen könnte –, muß man vollkommen blind neue Hypothesen entwickeln, ohne jeden Hinweis darauf, an welcher Stelle neue Hypothesen dringend benötigt werden.

4.3. *Die Neyman-Pearson-Theorie als Entscheidungstheorie*

Die NPT geht vor allem in ihren späteren Formulierungen davon aus, daß die Entscheidung zwischen Hypothesen nichts anderes ist als die Entscheidung zwischen verschiedenen Handlungen: mit einer Hypothese akzeptiert oder verwirft man gewisse weitere Handlungen. ‚Akzeptanz' einer Hypothese bedeutet hier also mehr als „die Hypothese vorläufig als nicht falsifiziert betrachten". Die NPT ist nach Neyman (1957) keine Theorie des statistischen Schließens, sondern eine Theorie des „induktiven Verhaltens". Nach dieser Sichtweise sind mit den in Betracht gezogenen Handlungen Kosten und Nutzen verbunden, und diese Kosten und Nutzen hängen davon ab, welche der in Betracht gezogenen Hypothesen tatsächlich wahr ist. Schon die frühen Formulierungen der NPT gehen davon aus, daß solche Kosten- und Nutzenüberlegungen das Signifikanzniveau bestimmen (Neyman & Pearson 1933: 146). Diese Idee wurde von Wald (1950) zu einer Theorie der Entscheidung unter Unsicherheit ausgebaut. Er nimmt an, daß die Konsequenzen der möglichen Entscheidungen nach einem einheitlichen Maßstab bewertet werden können; üblicherweise wird die Existenz einer Nutzenfunktion unterstellt. Wenn die Hypothese H_1 wahr und die Wahrscheinlichkeit eines Fehlers 1. Art gleich α ist, dann ist der erwartete Nutzen aus der Entscheidung, die Hypothese genau dann zu akzeptieren, wenn das beobachtete Ereignis im Akzeptanzbereich liegt, gleich $EU_1(\alpha) = \alpha U_{21} + (1\text{-}\alpha)U_{11}$ mit U_{ij} als dem Nutzen aus der Entscheidung, Hypothese i zu akzeptieren, wenn Hypothese j wahr ist. Es wird unterstellt, daß $U_{kk} > U_{ji}$ für $i \neq j$, daß also ein Fehler immer schlechter ist als eine richtige Entscheidung. Wenn H_2 wahr ist, ergibt sich ganz analog $EU_2(\beta) = \beta U_{12} + (1\text{-}\beta)U_{22}$, wobei β die Wahrscheinlichkeit eines Fehlers 2. Art ist.

Da innerhalb der NPT die Verwendung subjektiver Wahrscheinlichkeiten für die Gültigkeit der Hypothesen strikt abgelehnt wird, ist der erwartete Gesamtnutzen für das Entscheidungsproblem normalerweise nicht definiert. Für diesen Fall werden in der Entscheidungstheorie verschiedene Entschei-

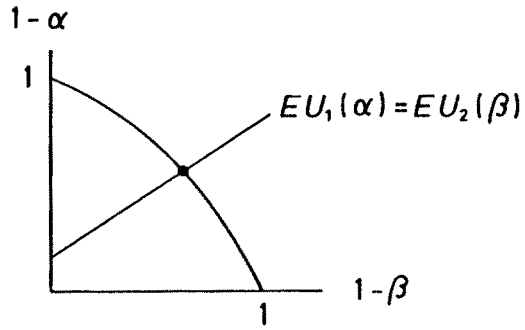

Abb. 3. Wahl des Signifikanzniveaus in der NPT.

dungskriterien diskutiert. In der NPT hat das Maximin-Kriterium – Maximierung des minimalen, also unter den schlechtestmöglichen Umständen erreichbaren Nutzens – eine gewisse Prominenz erlangt.[10] Bei Verwendung des Maximin-Kriteriums ergibt sich als notwendige Optimalitätsbedingung für die Wahl der Fehlerwahrscheinlichkeiten $EU_1(\alpha)=EU_2(\beta)$ oder

$$\beta = [(U_{21}-U_{11})/(U_{12}-U_{22})]\alpha + [(U_{11}-U_{22})/(U_{12}-U_{22})].$$

α und β müssen also auf der so beschriebenen Geraden liegen und sollen so klein wie möglich sein. Der Zusammenhang zwischen $1-\alpha$ und $1-\beta$, der die Wahlmöglichkeiten beschränkt, ergibt sich durch Maximierung der Macht $1-\beta$ für jedes Signifikanzniveau α. Jede Kombination von $1-\alpha$ und $1-\beta$, die nicht auf der so bestimmten Kurve liegt, ist entweder nicht erreichbar oder aber ineffizient, weil die beiden Fehlerwahrscheinlichkeiten α, β größer als notwendig sind. Der Schnittpunkt der Geraden mit der Kurve bestimmt eindeutig die optimale Kombination von Fehlerwahrscheinlichkeiten (Abb. 3 oben).

Es läßt sich leicht zeigen, daß sich nach der oben aufgemachten Rechnung ein (kostenloser) Hypothesentest auf jeden Fall lohnt. Bei Anwendung des Maximin-Kriteriums ohne vorherigen Test wird zugunsten derjenigen Hypothese entschieden, für die der minimale Nutzen, also der Nutzen bei einer Fehlentscheidung, höher ist. Das läuft darauf hinaus, eine der Fehlerwahrscheinlichkeiten gleich 0 und die andere gleich 1 zu setzen. Ein Test bringt damit auf jeden Fall eine Verbesserung, da man dadurch zum optimalen Punkt gelangt.

Die entscheidungstheoretische Perspektive steht im scharfen Gegensatz zur Perspektive Fishers. Nach Fisher liefern Beobachtungen *Evidenz* für oder gegen Hypothesen; die Entscheidung, eine Hypothese vorläufig zu verwerfen, erfolgt durch eine Bewertung dieser Evidenz. Neyman & Pearson dagegen gehen davon aus, daß statistische Tests keine Evidenz liefern:

> We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can in itself provide any valuable evidence of the truth or falsehood of that hypothesis. (Neyman & Pearson 1933: 141f)

Der Gegensatz zwischen Entscheidungstheorie und Evidenztheorie zeigt sich besonders deutlich in der Diskussion um die vorexperimentelle und nachexperimentelle Deutung der Versuchsergebnisse (s. Hacking 1965: 95ff, Mayo 1982, Casella 1988). Dabei geht es u.a. um die Frage, ob die *vorexperimentellen* Optimalitätseigenschaften eines Tests *nach* der Durchführung des Experiments noch irgendeine Bedeutung haben. In der Auseinandersetzung zwischen Fisher und Neyman finden die unterschiedlichen Standpunkte in den unterschiedlichen Interpretationen der Fehlerwahrscheinlichkeiten ihren Ausdruck. Nach Fisher (1959: 44) ist der beobachtete P-Wert vor allem ein vernünftiges Maß für die Abneigung („reluctance"), eine Hypothese für wahr zu halten. Dem P-Wert wird damit eine nachexperimentelle Bedeutung gegeben. Für Neyman und Pearson dagegen sind die Fehlerwahrscheinlichkeiten die Grenzwerte relativer Häufigkeiten von Fehlentscheidungen in einer langen Reihe von Anwendungen des Tests. Das ist eine rein vorexperimentelle Perspektive, denn nach dem Experiment ist für jede Entscheidung die Wahrscheinlichkeit, daß es sich dabei um eine Fehlentscheidung handelt, entweder Null oder Eins, je nachdem, welche Hypothese wahr ist.

Die Frage, die die Vertreter der Evidenztheorien stellen, ist: Warum sollten vorexperimentelle Fehlerwahrscheinlichkeiten nach dem Experiment noch irgendeine Bedeutung haben? Die entscheidungstheoretisch motivierte NPT dagegen versucht zu zeigen, wie Experimentatoren optimale Entscheidungen treffen können. Sie scheint einen Ersatz für methodologische Regeln zu bieten: optimale Entscheidungen sollen an die Stelle einer methodologischen Bewertung treten. Aus dieser Perspektive ist die Existenz einer *nachexperimentellen* Problemstellung ein Rätsel: Welche nachexperimentelle Überlegung kann denn nicht schon vor dem Experiment berücksichtigt werden?

Um diese Positionen beurteilen zu können, muß man sich zunächst klarmachen, daß die NPT keine Lösung für individuell optimales Verhalten bietet. Betrachten wir ein Experiment im Zeitablauf. Der Experimentator muß eine Entscheidung fällen, für die es wichtig ist zu wissen, welche der beiden Hypothesen, H_1 oder H_2, wahr ist; über diesen Punkt herrscht subjektiv völlige Unsicherheit. Er bestimmt zunächst den Ablehnungsbereich für H_1 nach dem Maximin-Kriterium und führt dann das Experiment durch. Nach dem Experiment muß er die endgültige Entscheidung zwischen H_1 und H_2 fällen, die für ihn zu bestimmten Gewinnen oder Verlusten führt, je nachdem, ob nun H_1 oder H_2 wahr ist. Die NPT fordert, daß der Experimentator sich gemäß dem vor dem Experiment gewählten Ablehnungsbereich entscheidet. Grundsätzlich könnte er seine Entscheidung über den Ablehnungsbereich aber auch revidieren. Was ist die optimale Strategie, wenn man diese Revisionsmöglichkeit berücksichtigt? Nach dem Experiment führt, wie erwähnt, jede Entscheidung mit Sicherheit zu einem bestimmten Ergebnis, das nur davon abhängt, welche Hypothese wahr ist. Über letzteren Punkt herrscht dieselbe subjektive Unsicherheit wie vor Durchführung des Experiments. Eine Anwendung des Maximin-Kriteriums

nach dem Experiment führt zu einer Entscheidung ohne Berücksichtigung des Tests, da die Testergebnisse die subjektive Unsicherheit darüber, welche Hypothese wahr ist, weder in Sicherheit noch in Wahrscheinlichkeiten verwandeln können.

Die NPT verpflichtet den Experimentator, sich vor dem Experiment mit der Entscheidung für einen bestimmten Ablehnungsbereich zu binden. Falls der Experimentator sich vor dem Experiment effektiv binden *kann*, besteht die optimale Strategie darin, dies auch zu tun. In der Wissenschaft ist eine solche individuelle Selbstbindung jedoch nicht möglich, da nichts den Experimentator (oder jemand anderen) daran hindern kann, nach dem Experiment die Entscheidung noch einmal zu überdenken. Wenn eine Bindung aber unmöglich ist, dann liefert die NPT keine optimalen Entscheidungen. Das liegt daran, daß die NPT-Lösung dann nicht Bellman-optimal ist. Das Bellman-Kriterium besagt, daß die anfangs gewählte Strategie an jedem Entscheidungspunkt weiter die beste sein muß; andernfalls war sie von Anfang an nicht optimal (White 1976: 299). Berücksichtigt man dieses Kriterium, dann wird man von vorneherein *den* Ablehnungsbereich wählen, den man auch ohne Möglichkeit eines Experiments gewählt hätte.

Die Alternative zur Interpretation der NPT als eines individuellen Entscheidungskalküls ist die Interpretation als Methodologie. Die Grundidee dieser Methodologie besteht darin, methodologische Regeln so zu wählen, daß ihre Verwendung möglichst wenige Fehlentscheidungen erwarten läßt. Diese Idee ist mit dem Falsifikationismus grundsätzlich gut vereinbar. Nach Popper (1984: 22ff) sind methodologische Regeln Konventionen. Das Problem der statistischen Testtheorie ist offensichtlich das Problem der vernünftigen Wahl solcher Konventionen. Dieses Problem diskutiert Popper nirgends explizit, aber es ist klar, daß diese Konventionen nicht beliebig sind. Die einzige Möglichkeit, eine Reduzierung methodologischer Regeln auf bloße Spielregeln zu vermeiden, ist, die Methodologie als eine Technologie aufzufassen, also als eine auf wissenschaftlichen Erkenntnissen beruhende Lehre darüber, wie man bestimmte Erkenntnisziele am besten erreicht.[11] Methodologische Regeln müssen auf Grund ihrer Zweckmäßigkeit beurteilt werden. Entscheidend sind die zu erwartenden Folgen bei der Anwendung dieser Regeln, also die zu erwartende Qualität der auf Grund dieser Regeln getroffenen Entscheidungen. Hacking (1980) hat bereits darauf hingewiesen, daß man die Fehlerwahrscheinlichkeiten in der NPT als Eigenschaften von – wie er es nennt – Schlußregeln auffassen sollte. Eine Entscheidung für oder gegen eine Hypothese – nach Hacking ein induktiver Schluß – wird dadurch gerechtfertigt, daß man auf die Qualität der Regel verweist, auf Grund derer die Entscheidung getroffen wurde. Eine Kritik der Entscheidung muß also auf der Regelebene erfolgen. Das schließt nachexperimentelle Evidenzbetrachtungen aus; der vorexperimentelle Standpunkt ist der richtige, weil nur dieser Standpunkt es ermöglicht, die Qualität der methodologischen Regeln zu diskutieren.

Evidenztheorien sind mit dem hier vertretenen Standpunkt vereinbar, wenn man sich klar macht, daß es eben eine Frage vernünftiger Konventionen ist, was man als Evidenz betrachtet. Bei Fisher fehlt eine explizite Formulierung dieser Konventionen und damit notwendigerweise auch eine Diskussion ihrer Eigenschaften. Vertreter anderer Evidenztheorien wie der Likelihoodtheorie oder der bayesianischen Theorie formulieren ihre Konventionen zwar explizit, weigern sich aber meist, die erwarteten Folgen ihrer Anwendung zu diskutieren. Dabei ziehen sie sich auf den Standpunkt zurück, solche Eigenschaften wie Fehlerwahrscheinlichkeiten seien nur aus einer vorexperimentellen Perspektive sinnvoll, während Evidenzüberlegungen auf der nachexperimentellen Perspektive beruhen. Sie übersehen aber, daß sie damit den einzigen Standpunkt aufgeben, von dem aus sich solche Konventionen in Bezug auf ihre Zweckmäßigkeit beurteilen lassen.

Soweit läßt sich also die Verwendung von Fehlerwahrscheinlichkeiten gegen Angriffe etwa auf der Grundlage der Likelihood-Theorie verteidigen. Das bedeutet allerdings nicht, daß deswegen die NPT insgesamt akzeptabel ist. Im Gegenteil: die entscheidungstheoretisch motivierten Bestandteile der NPT sind mit dem grundsätzlich methodologischen Charakter der Theorie nicht vereinbar. Wenn man die Auswahl der Alternativhypothesen und des Ablehnungsbereichs von den Interessen oder der Nutzenfunktion des Experimentators abhängig macht, behandelt man den Test als ein Problem individuell rationaler Entscheidung. Aus dieser Sicht macht jedoch ein NP-Test keinen Sinn, wie wir gesehen haben. Die Alternative, eine konsequente Behandlung der NPT als Methodologie, erfordert aus falsifikationistischer Sicht die Ausschaltung pragmatischer Elemente.

5. EINE FALSIFIKATIONISTISCHE TESTTHEORIE

Rekapitulieren wir kurz die Problemstellung. Gesucht wird eine methodologische Regel, die den Ablehnungsbereich einer statistischen Hypothese festlegt. Die Regel sollte folgenden Ansprüchen genügen:

- Eine eindeutige Beurteilung der Hypothese allein auf Grundlage der bekannten Daten muß möglich sein.
- Fehlentscheidungen sollten möglichst selten sein.

Aus der ersten Bedingung folgt, daß Alternativhypothesen nicht zur Bestimmung des Ablehnungsbereichs herangezogen werden dürfen, während es die zweite Bedingung wünschenswert erscheinen läßt, die Wahrscheinlichkeit eines Fehlers 2. Art bzw. die Macht eines Tests zu berücksichtigen. Wenn die beiden Anforderungen miteinander vereinbar sein sollen, muß also irgendein Substitut für die Macht eines Tests gefunden werden, das nicht auf der Verwendung von Alternativhypothesen beruht. Dieses Substitut ist der empirische Gehalt. Nach der falsifikationistischen Auffassung ist empirischer Gehalt wünschenswert. Es liegt also nahe, eine von Popper (1984: 85) vorgeschlagene allgemeine methodologische Regel zu verwenden, die besagt, man solle den empirischen Gehalt von Hypothesen so groß

wie möglich machen. Der empirische Gehalt übernimmt bei der vorgeschlagenen Regel die Funktion der Macht eines Tests; die Regel zur Bestimmung des Ablehnungsbereichs heißt nicht mehr „Maximiere die Macht des Tests bei gegebener Irrtumswahrscheinlichkeit", sondern „Maximiere den empirischen Gehalt bei gegebener Irrtumswahrscheinlichkeit". In der NPT ist die Macht eines Tests ein Maß der Verwundbarkeit der zu testenden Hypothese, allerdings ein Maß, das ganz spezifische Alternativen berücksichtigt. Eine Vergrößerung der Macht eines Tests bedeutet, daß die Hypothese gegenüber dieser Alternative verwundbarer wird; eine Vergrößerung des empirischen Gehalts dagegen bedeutet eine Erhöhung der Verwundbarkeit ohne Bezug auf spezifische Alternativhypothesen. In diesem Sinne greift die hier vorgeschlagene Testtheorie einen Grundgedanken der NPT auf.

Ich zeige zunächst anhand einer diskreten Verteilung, wie man die Regel der Maximierung des empirischen Gehalts auf statistische Hypothesen anwenden kann, und gehe dann auf spezielle Probleme ein.

5.1 *Der empirische Gehalt statistischer Hypothesen*

Die zu testende Hypothese besagt im einfachsten Fall, daß die Größe X in der Situation S einer diskreten Verteilung mit einer endlichen Zahl von Elementarereignissen gehorcht, wie sie z.B. durch die Wahrscheinlichkeitsmassefunktion $f(X)$ der Abb. 4 (unten) dargestellt wird. Diese Hypothese soll zunächst anhand einer einzigen Beobachtung getestet werden; die Irrtumswahrscheinlichkeit sei vorgegeben.

Die einzelnen Elementarereignisse reichen von $x=1$ bis $x=7$; die Wahrscheinlichkeiten reichen von $W(x=1) = W(x=7) = 1/16$ bis zu $W(x=4) = 4/16$. Angenommen, die vorgegebene Irrtumswahrscheinlichkeit sei 2/16. Diese Irrtumswahrscheinlichkeit kann zum Beispiel realisiert werden, indem man entweder das Ereignis $x=2$ oder das Ereignis $x=6$ als Ablehnungsbereich festlegt. In beiden Fällen gäbe es genau eine zusätzliche Beobachtung – über den deterministischen Gehalt hinaus –, die zur Falsifikation der Hypothese führt. Diese eine zusätzliche Beobachtung ist der empirische Gehalt, den man bei einer solchen Wahl des Ablehnungsbereiches für die vorgegebene Irrtumswahrscheinlichkeit erhält. Allerdings kann man bei

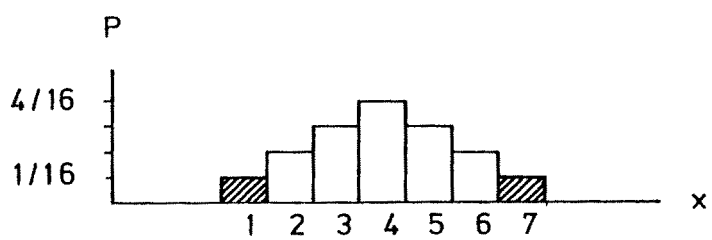

Abb. 4. Maximierung des empirischen Gehalts.

gleicher Irrtumswahrscheinlichkeit mehr an empirischem Gehalt bekommen, indem man zuerst die Ereignisse mit den kleinsten Wahrscheinlichkeiten in den Ablehnungsbereich nimmt, dann die Ereignisse mit den nächstgrößeren Wahrscheinlichkeiten usf., bis die aufsummierten Wahrscheinlichkeiten gerade gleich der vorgegebenen Irrtumswahrscheinlichkeit sind. Im vorliegenden Fall wird man also die beiden Ereignisse $x=1$ und $x=7$ als Ablehnungsbereich festsetzen. Das ergibt eine Irrtumswahrscheinlichkeit von 2/16 wie verlangt.Der empirische Gehalt ist jetzt aber doppelt so hoch wie bei einem Ablehnungsbereich von $x=2$, denn es sind nun zwei Ereignisse ($x=1$, $x=7$), die zur Falsifikation führen. Die Maximierung des empirischen Gehalts erfordert also, daß man die ‚Schwänze' einer unimodalen Verteilung als Ablehnungsbereich wählt – so, wie es dem üblichen Signifikanztest entspricht. Ein anderes Ergebnis ist, daß eine Gleichverteilungshypothese anhand einer Beobachtung nicht testbar ist. Das erscheint auch vernünftig: auf Grund der Symmetrie der Situation kann es keine Regel geben, die bei einer Gleichverteilung bestimmte Elementarereignisse als Ablehnungsbereich auswählt.

Eine Ergänzung der vorgeschlagenen Regel wird dadurch notwendig, daß z.B. für eine Irrtumswahrscheinlichkeit von 1/16 die Wahl des Ablehnungsbereichs uneindeutig ist. Die Ergänzung fordert, daß das höchste Niveau der Irrtumswahrscheinlichkeit gewählt werden soll, das nicht über dem vorgegebenen Niveau liegt und bei dem die Wahl des Ablehnungsbereichs eindeutig ist.[12] Für eine vorgegebene Irrtumswahrscheinlichkeit von 1/16 ist der Ablehnungsbereich nach dieser Regel leer: eindeutig bestimmte Ablehnungsbereiche existieren nur für die Irrtumswahrscheinlichkeiten 0, 2/16, 6/16, 12/16 und 16/16. Je feiner die Verteilung, desto unwichtiger wird diese Ergänzung.

Ein weiteres technisches Problem entsteht bei Verteilungen mit unendlich vielen Elementarereignissen. In diesem Fall wird der Ablehnungsbereich unendlich groß; es ist dann möglich, bei gegebener Irrtumswahrscheinlichkeit verschiedene unendlich große Ablehnungsbereiche zu konstruieren. In diesem Fall sollte man die Größe des Akzeptanzbereiches minimieren und den empirischen Gehalt z.B. als Kehrwert der Größe der Akzeptanzbereichs definieren. Diese Modifikation ist nicht ad hoc. Ein höherer empirischer Gehalt äußert sich darin, daß die aus der Hypothese ableitbare Prognose schärfer ist. Da eine statistische Hypothese nach falsifikationistischer Auffassung prognostiziert, daß das zu beobachtende Ereignis im Akzeptanzbereich liegt, bedeutet ein kleinerer Akzeptanzbereich eine Verschärfung dieser Prognose und somit größeren empirischen Gehalt.

Ich habe bis jetzt stillschweigend vorausgesetzt, daß man den empirischen Gehalt messen kann, indem man die Elementarereignisse im Ablehnungsbereich zählt. Bei dieser Annahme handelt es sich um einen Bestandteil der methodologischen Regel, denn bis auf den trivialen Fall, in dem eine Satzmenge eine andere enthält, lassen sich Gehaltsvergleiche nicht ohne zusätliche Voraussetzungen durchführen. [13] Es ist also keineswegs selbst-

verständlich, daß der empirische Gehalt durch die Zahl der Elementarereignisse gemessen wird; allerdings sehe ich weder schlagende Gegenargumente noch sinnvolle Alternativen. Daher erscheint mir vorläufig keine weitere Diskussion dieses Punktes erforderlich.

Die Tatsache, daß man sich bei der Messung des empirischen Gehalts auf bestimmte Konventionen einigen muß, wird noch deutlicher, wenn man zum Problem des Tests anhand mehrerer Beobachtungen übergeht. Hier bieten sich zwei grundsätzlich verschiedene Möglichkeiten an, um die falsifizierenden Ereignisse abzuzählen: man kann die Ereignisse im Ablehnungsbereich nach der Anordnung der Beobachtungen unterscheiden oder auch nicht. Für welche der beiden Zählweisen soll man sich entscheiden?

Die Formulierung der Gesetzeshypothese impliziert die Unabhängigkeit aller Beobachtungen, weil das Vorliegen der Situation S nach dieser Hypothese hinreichend dafür ist, daß X gemäß $f(X)$ verteilt ist. Damit kann es keinen Einfluß einer Ausführung des Experiments auf eine andere – korrekte – Ausführung geben, wenn die Hypothese wahr ist. Daher könnte man n Beobachtungen der Größe X einfach als *eine* Beobachtung eines Vektors auffassen, dessen Verteilung durch das n-fache Produkt der Massefunktion $f(X)$ bestimmt ist. Auf diese Verteilung könnte man nun die oben formulierte Regel anwenden. Man würde dann die Zahl der verschiedenen Beobachtungen des n-dimensionalen Vektors *unter Berücksichtigung der Anordnung* als Maß des empirischen Gehalts verwenden. Die sich daraus ergebende Testtheorie würde jedoch auch bei einer großen Stichprobe vom Umfang n die n-fache Beobachtung des häufigsten Wertes als bestmögliche Bestätigung der Hypothese auffassen. Das ist mit der üblichen Vorstellung unvereinbar, nach der man für eine diskrete Hypothese eine gute Anpassung der relativen Häufigkeiten der Elementarereignisse an die jeweiligen Wahrscheinlichkeiten erwartet, wenn die Stichprobe groß genug ist. Außerdem ist das n-fache Produkt einer Gleichverteilung wieder eine Gleichverteilung. Wie wir oben gesehen haben, kann man bei einer Gleichverteilung keine Regel für die Festlegung des Ablehnungsbereichs finden. Eine Berücksichtigung der Anordnung würde also unnötigerweise dazu führen, daß Gleichverteilungshypothesen auch anhand mehrerer Beobachtungen nicht testbar sind.

Ich schlage daher vor, den empirischen Gehalt ohne Berücksichtigung der Anordnung zu messen. Die Verteilung von n Beobachtungen ohne Berücksichtigung der Anordnung ist für eine diskrete Hypothese mit m Elementarereignissen eine Multinomialverteilung mit $(n+m-1)!/n!(m-1)!$ Zellen. Der Ablehnungsbereich wird maximiert, indem man die k Zellen mit den kleinsten Wahrscheinlichkeiten in den Ablehnungsbereich aufnimmt, so daß (1) die aufsummierten Wahrscheinlichkeiten das vorgegebene Signifikanzniveau gerade nicht überschreiten und (2) der Ablehnungsbereich eindeutig ist. Dieser Test wird in der Liteatur als der multinomiale Anpassungstest bezeichnet; er kann unter bestimmten Bedingungen durch den Chi-Quadrat-Anpassungstest approximiert werden.[14]

5.2. *Tests anhand echter Folgerungen*

Die Gesetzeshypothese, daß n beliebige Beobachtungen bei jeweils korrekter Durchführung des Experiments einer bestimmten Multinomialverteilung gehorchen, ist äquivalent zu der ursprünglichen Gesetzeshypothese, da man eine Hypothese aus der anderen herleiten kann: Wegen der durch die Gesetzeshypothese implizierten Unabhängigkeit der Beobachtungen folgt die Multinomialverteilungshypothese aus der ursprünglichen Hypothese; aus der Multinomialverteilungshypothese folgt die ursprüngliche Hypothese, weil man, wenn die ursprüngliche Hypothese nicht wahr wäre, prinzipiell die Experimente so gestalten könnte, daß die Multinomialverteilungshypothese für die ausgewählten Experimente nicht gelten würde.

Die Multinomialverteilungshypothese ist also eine Folgerung, aber keine echte, d.h. logisch schwächere Folgerung aus der ursprünglichen Hypothese. Allerdings könnte man aus der ursprünglichen Hypothese eine echte Folgerung ableiten und diese Folgerung mit Hilfe des multinomialen Anpassungstests testen. Da jede statistische Testtheorie notwendigerweise das Konsequenzprinzip verletzt, kann man nicht beliebige echte Folgerungen heranziehen und eine etwaige Falsifikation auf die ursprüngliche Hypothese übertragen.[15] Die Auswahl einer echten Folgerung, anhand derer die Hypothese getestet werden könnte, müßte durch eine eigene methodologische Regel erfolgen. Die Frage ist aber, ob Tests anhand echter Folgerungen überhaupt sinnvoll sind.

Man kann zwei Arten von echten Folgerungen unterscheiden: (1) Folgerungen, die sich daraus ergeben, daß die Verteilung durch Zusammenfassung einiger Elementarereignisse zu einem neuen Elementarereignis vergröbert wird, und (2) Folgerungen, die wie im Fall von Stoppregelexperimenten durch Einführung neuer Randbedingungen in der Wenn-Komponente der Gesetzeshypothese entstehen.

Die Betrachtung der ersten Art von Folgerungen führt nur dazu, daß die Wahl des Ablehnungsbereichs eingeschränkt wird, weil gewisse Ereignisse immer zusammen in den Ablehnungs- oder Akzeptanzbereich aufgenommen werden müssen. Die Einführung einer zusätzlichen Beschränkung bei der Wahl des Ablehnungsbereichs kann jedoch keine Verbesserung bringen. Damit wäre eine methodologische Regel, die eine echte Folgerung der ersten Art für den Test auswählt, sinnlos.

Eine Folgerung der zweiten Art ist nicht durch alle Beobachtungen von Ergebnissen korrekt durchgeführter Experimente prüfbar. Das liegt an den zusätzlichen Randbedingungen, die in der Wenn-Komponente der Folgerung auftauchen. Damit ist die Folgerung jedoch schlechter prüfbar als die ursprüngliche Hypothese. Zwar kann man möglicherweise zeigen, daß es bei gegebener Irrtumswahrscheinlichkeit Beobachtungen gibt, die die Folgerung, nicht aber die ursprüngliche Hypothese widerlegen. Ein solches Ergebnis würde den direkten Vergleich des empirischen Gehalts von ursprünglicher Hypothese und Folgerung mit Hilfe der Teilklassenbeziehung

unmöglich machen. Aber da es um die Wahl einer methodologischen Regel und nicht um Einzelfallentscheidungen geht, wiegt es schwerer, daß für echte Folgerungen der zweiten Art gewisse Beobachtungen von vornherein irrelevant sind. Die beste methodologische Regel ist diejenige, nach der möglichst viele Beobachtungen für die Hypothese relevant sind. Damit scheiden auch echte Folgerungen der zweiten Art aus. Der multinomiale Anpassungstest sollte also auf die ursprüngliche Hypothese angewandt werden; die Berücksichtigung von Stoppregeln ist nicht sinnvoll.

5.3. *Unabhängigkeit*

Für Tests anhand mehrer Beobachtungen wurde vorgeschlagen, die Anordnung der Beobachtungen zu vernachlässigen. Dagegen könnte man einwenden, aus einer auffälligen Regelmäßigkeit in der Anordnung lasse sich auf eine Verletzung der Unabhängigkeitsannahme und damit auf die Falschheit der Hypothese schließen.[16] Der Einwand muß zunächst präzisiert werden. Unter einer Regelmäßigkeit wollen wir eine deterministische Regelmäßigkeit verstehen; die zusätzliche Einbeziehung stochastischer Prozesse ist, wie wir sehen werden, nicht notwendig. Eine solche deterministische Regelmäßigkeit können wir durch eine Rekursionsformel, die die Folge der Beobachtungen liefert, beschreiben. Im Falle einer Reihe von Münzwürfen könnte man sich als extrem verdächtiges Ergebnis vorstellen, daß die Münze immer abwechselnd Kopf und Zahl zeigt. Mit e_i für das Ergebnis des i-ten Wurfs lautet die Rekursionsformel dann z.B. e_i = Kopf, $e_{i+1} \neq e_i$, $e_i \in \{\text{Kopf, Zahl}\}$.

Nun läßt sich aber jede endliche Stichprobe durch eine genügend komplexe Rekursionsformel beschreiben. Diese Überlegung stammt aus einer Weiterentwicklung der von Misesschen Theorie der Zufallsfolgen, die der Häufigkeitsinterpretation der Wahrscheinlichkeit zugrundeliegt.[17] In der modernen Variante dieser Theorie wird die Komplexität der die Stichprobe beschreibenden Rekursionsformel als ein Maß der Zufälligkeit der Stichprobe gedeutet. Der Begriff der Zufälligkeit spielt für uns keine Rolle; wir ziehen nur die rein formalen Ergebnisse aus der Häufigkeitstheorie der Wahrscheinlichkeit als Hilfsmittel heran. Im vorliegenden Zusammenhang dient die Komplexität der Rekursionsformel – wobei der Begriff der Komplexität mathematisch scharf gefaßt werden kann – als ein Mittel, den obigen Einwand gegen eine Vernachlässigung der Anordnung der Stichprobenelemente zu präzisieren. Eine solche präzise Fassung könnte lauten:

> Wenn wir eine Stichprobe ziehen, die durch eine relativ einfache Rekursionsformel beschrieben werden kann, ist dies ein Hinweis darauf, daß die statistische Unabhängigkeit der Beobachtungen verletzt sein könnte. Eine Testtheorie für isolierte Hypothesen muß dem Rechnung tragen, indem sie in geeigneter Weise die Anordnung der Stichprobenelemente berücksichtigt.

Ich will diesen Einwand nicht weiter ausführen; auch in dieser immer noch reichlich ungenauen Fassung läßt sich zeigen, woran ein Argument

dieser Art scheitern muß. Die Theorie der Zufallsfolgen entstammt einer Tradition, in der ein wesentliches Problem darin bestand, eine Zufallsfolge zu *definieren*; die moderne Fassung der Theorie behandelt das Problem, eine Zufallsfolge zu *erkennen*. In jedem Fall handelt es sich darum, mit einer Folge im mathematischen Sinn umzugehen. Wie aber schon weiter oben ausgeführt wurde, liefern nach der propensity-Theory die durchgeführten Experimente keine solchen Folgen, da für eine gegebene Stichprobe die *relevante* Anordnung der Elemente nicht gegeben ist. Wenn die Hypothese wahr ist, sollte jede Anordnung für Testzwecke gleich gut sein; wenn die Hypothese falsch ist, dann gibt es keine Grenzen für die statistischen Abhängigkeiten, die zwischen den einzelnen Experimenten bestehen können. Statistische Abhängigkeiten können durch beliebige intervenierende Variable zustandekommen; es ist daher keineswegs sicher, daß die Anordnung der Stichprobenelemente z.B. nach der zeitlichen Folge der Experimente die richtige Anordnung ist, um statistische Abhängigkeiten ans Licht zu bringen. Wenn aber alle Anordnungen als gleichwertig angesehen werden müssen, bricht jedes Argument, das auf eine Überprüfung der statistischen Unabhängigkeit hinausläuft, zusammen: bei genügend vielen Beobachtungen existiert immer eine Anordnung, nach der ein vorgegebener Unabhängigkeitstest zu einer Ablehnung führt. Das gilt erst recht, wenn Regelmäßigkeiten durch irgendetwas Allgemeineres als eine Rekursionsformel beschrieben werden.

Das Argument gegen die Existenz einer kanonischen Anordnung der Stichprobenelemente impliziert, daß die Unabhängigkeit der Beobachtungen kein für sich prüfbarer Bestandteil einer statistischen Hypothese ist. Ein Test auf Unabhängigkeit der Beobachtungen kann also nur ein Test auf eine bestimmte Form der Abhängigkeit sein. Die Vernachlässigung der Anordnung der Beobachtungen führt nicht dazu, daß Falsifikationsmöglichkeiten verlorengehen.

5.4. *Kontinuierliche Verteilungen*

Die methodologischen Regeln zur Festlegung des Ablehnungsbereichs werden auf Hypothesen angewendet, die sich direkt auf beobachtbare Elementarereignisse beziehen. Die Beobachtbarkeit der Elementarereignisse ist nur gegeben, wenn die Hypothesen diskrete Verteilungen postulieren. Das gilt auch dann, wenn man annimmt, daß die interessierende Variable ein Kontinuum von Werten annehmen kann, da Messungen dieser Variablen nur endlich genau sein können. Die zu prüfende Hypothese ist immer eine Hypothese über die Verteilung der Meßwerte und nicht über die Verteilung der zu messenden Größe. Tatsächlich gehen viele Statistiker (z.B. Basu 1988: 23 Fn) davon aus, daß echte kontinuierliche Verteilungen oder sogar Verteilungen mit unendlich vielen Elementarereignissen irrelevant sind.

Zumindest auf der Ebene von Hypothesen, die direkt mit Beobachtungen konfrontiert werden können, muß man kontinuierliche Verteilungen als

Approximationen sehr feiner diskreter Verteilungen betrachten. Eine solche Approximation ist nur dann gut, wenn die Wahrscheinlichkeit dafür, daß dasselbe Ereignis mehrmals beobachtet wird, vernachlässigbar ist. Unter diesen Umständen ist es gleichgültig, ob man den Ablehnungsbereich mit oder ohne Berücksichtigung der Anordnung festlegt, da sich ein Unterschied zwischen beiden Verfahrensweisen nur für Ereignisse der Wahrscheinlichkeit Null ergibt. Der Ablehnungsbereich besteht dann aus dem Bereich der Elementarereignisse kleinster Dichte, da nur so der Ablehnungsbereich maximal bzw. der Akzeptanzbereich minimal wird.

Die Auffassung, daß kontinuierliche Verteilungen nur Approximationen diskreter Verteilungen sind, begegnet folgendem Argument, das von Redhead (1974) gegen Gillies' (1971) falsifikationistische Testtheorie vorgebracht wurde und das eine Vereinfachung von Neymans Paradox darstellt.

Redheads Argument: Eine kontinuierlich verteilt Größe kann auf verschiedene Arten transformiert werden. Wenn man z.B. auf ein normalverteiltes X die folgende Transformation anwendet, gelangt man zu einer Größe $Y = h(X)$, die ebenfalls wieder mit demselben Mittelwert und derselben Varianz normalverteilt ist.

$$Y = h(X) = \begin{cases} F^{-1}(3/2 - F(X)), & X > 0 \\ F^{-1}(1/2 - F(X)), & X < 0. \end{cases}$$

Dabei ist F die Verteilungsfunktion der Normalverteilung, also die kumulierte Wahrscheinlichkeitsdichte, und F^{-1} die Umkehrfunktion dazu. Die Transformation beruht darauf, daß die Größe F(X) im Intervall [0, 1] gleichverteilt ist. Diese Transformation hat die Eigenschaft, ein x aus den Schwänzen der ursprünglichen Verteilung auf ein y im Zentrum der neuen Verteilung abzubilden. Was bei Betrachtung der Größe X zur Ablehnung führen würde, ist bei der Betrachtung der Größe Y also eine Bestätigung der Hypothese.

Nach der hier vertretenen Auffassung ist es jedoch unzulässig, den Ablehnungsbereich anhand der Verteilung der transformierten Größe Y zu bestimmen. Eine solche Transformation hat kein Gegenstück für Größen, die nur diskrete Werte annehmen können. Damit ist Redheads Argument nicht mehr anwendbar. Allein die Skala, auf der die Größe X tatsächlich gemessen wird, ist maßgeblich für die Messung des empirischen Gehalts.

Wir wenden die Überlegungen zum empirischen Gehalt kontinuierlicher Verteilungshypothesen auf eine einfache Normalverteilungshypothese mit Mittelwert Null an. Da der empirische Gehalt jetzt ohne Berücksichtigung der Anordnung die Bereiche der niedrigsten Dichte umfaßt, ergibt sich im Falle zweier Beobachtungen für eine Irrtumswahrscheinlichkeit von α als Akzeptanzbereich eine Kreisscheibe um den Ursprung mit Wahrscheinlichkeitsmasse $1-\alpha$. Der Durchmesser der Kreisscheibe ist bei gegebener Irrtumswahrscheinlichkeit desto größer, je höher die Varianz der Normalverteilung ist; mit der Varianz geht auch der Durchmesser der Kreisscheibe gegen unendlich.

Aus diesen Überlegungen und dem disjunktiven Testkriterium folgt, daß die dem *t*-Test zugrundeliegende Nullhypothese keinen empirischen Gehalt hat. Für jede Beobachtung und jede Irrtumswahrscheinlichkeit existiert eine Varianz, so daß die Beobachtung im Akzeptanzbereich der entsprechenden Normalverteilungshypothese liegt. Der *t*-Test kann damit nur durch Voraussetzung einer Oberhypothese gerechtfertigt werden, wie dies in der NPT geschieht. Diese Oberhypothese hat selbst keinen empirischen Gehalt und ist damit äußerst problematisch, weil nicht unabhängig prüfbar.

6. SCHLUß

Das Hauptergebnis der vorliegenden Arbeit läßt sich kurz zusammenfassen: Im Falle diskreter Hypothesen mit endlich vielen Elementarereignissen ist der multinomiale Anpassungstest bzw. – wenn genügend Beobachtungen vorliegen – der Chi-Quadrat-Anpassungstest ein geeigneter Test, mit Hilfe dessen eine isolierte statistische Hypothese falsifiziert werden kann. Dieses Ergebnis kann man sowohl als eine Rechtfertigung der statistischen Praxis wie auch als eine systematische Ausführung von Randbemerkungen unorthodoxer Theoretiker und Philosophen ansehen. So schreibt z.B. Hacking:

> Statistical orthodoxy of different schools has become so entrenched that many now maintain that it is impossible to test when rival hypotheses are lacking. Hence many curious things are said about such a valuable device as the χ^2 test. (Hacking 1973: 499)

Ich betrachte es als eine wesentliche Stütze meiner Theorie, daß sie im Gegensatz zur NPT den Chi-Quadrat-Test erklärt. Die größte Stütze der NPT dagegen ist die Tatsache, daß sie den *t*-Test erklärt. Diese Erklärung wird durch die hier vorgeschlagene Testtheorie nicht in Frage gestellt; die einzige andere Erklärung des *t*-Tests, nämlich die Fishersche, ist ad hoc. Die Möglichkeit einer Falsifikation statistischer Hypothesen läßt Tests wie den *t*-Test, die die Wahrheit einer nicht falsifizierbaren Oberhypothese voraussetzen, allerdings in einem etwas anderen Licht erscheinen. Die bekannte Tatsache, daß der *t*-Test auf einer nicht falsifizierbaren Oberhypothese beruht, kann jetzt als ernsthafter Einwand gegen diesen Test angeführt werden, da die Voraussetzung einer Oberhypothese nicht, wie Neyman und Pearson behaupten, unvermeidlich ist.

NOTES

* Ich danke Gunnar Andersson, Volker Gadenne, Alfred Hamerle, Karl-Josef Koch, Martin Kukuk, Joachim Möller, Paul Michels und Hans-Günther Seifert-Vogt für Diskussionen und kritische Hinweise.

[1] S, auch Hacking (1980: 154). In der Neyman-Pearson-Theorie bedeutet Akzeptanz allerdings etwas anderes; vgl. unten S. 17.

[2] Vgl. für diese Ansicht auch Spielman (1974).

[3] Für eine systematische Darstellung s. Cramér (1946: 452ff) und Baird (1983); vgl. dazu Fisher (1970: 118).

[4] Vgl. Baird (1983) für einen allerdings nicht ausgearbeiteten Vorschlag.

[5] Neyman (1929), zitiert nach Neyman & Pearson (1930; 101 Fn), und Neyman (1952; 43ff).
[6] Baird (1983) weist anhand des Chi-Quadrat-Unabhängigkeitstests nach, daß Fishers Testtheorie diese Bedingung verletzt.
[7] Baird (1983: 114) verweist auf Bowley (1948: 494ff).
[8] Qualitativ das gleiche Problem tritt immer dann auf, wenn die Stichprobengröße in irgendeiner Weise von den Beobachtungen abhängt.
[9] Zum Fehler 3. Art vgl. Gigerenzer *et al.* (1989: 104) und den dort angeführten Artikel von Kimball (1957).
[10] Vgl. Lehmann (1959: 12f). Das Kriterium wird dort Minimax- statt Maximin-Kriterium genannt, da Lehmann auf die Minimierung des maximalen Schadens anstelle die Maximierung des minimalen Nutzens abstellt.
[11] Vgl. hierzu Hans Albert (1987: insbes. 70ff., 1989); dort wird vor allem auch gezeigt, daß eine solche Auffassung nicht zu einem Zirkel führt.
[12] S. dazu Lancaster (1969: 32f), von dem ich diese Lösung übernehme.
[13] Vgl. Popper (1984: 78ff). Die hier vorgeschlagene Vergleichsmöglichkeit entspricht Poppers Dimensionsvergleich mit Hilfe von relativ atomaren Sätzen (ebd.: 90f).
[14] Vgl. Horn (1977) für eine Übersicht über verschiedene Anpassungstests bei diskreten Verteilungen.
[15] Diese Folgerungsproblematik ist das Propensity-Gegenstück zum Problem der Wahl der Referenzmenge in der Häufigkeitstheorie.
[16] Hacking (1965: 78) und Stegmüller (1973: 147) werfen die Frage auf, wie man die Vernachlässigung der Anordnung *rechtfertigen* kann. Da es sich hier um Konventionen handelt, scheint mir die Frage falsch gestellt. Die richtige Frage muß lauten: Welche Folgen hat es, wenn man die Anordnung vernachlässigt?
[17] Für einen Überblick siehe Fine (1973: ch. V).

LITERATURVERZEICHNIS

Albert, Hans: 1987, *Kritik der reinen Erkenntnislehre*, Tübingen: Mohr (Siebeck).
Albert, Hans: 1989, ‚Die Möglichkeit der Erkenntnis', in Salamun, Kurt (ed.), *Karl R. Popper und die Philosophie des Kritischen Rationalismus*, Amsterdam: Rodopi 1989, 3–18.
Andersson, Gunnar: 1984, 'How to Accept Fallible Test Statements: Popper's Criticist Solution', in Andersson, Gunnar (ed.), *Rationality in Science and Politics*, Dordrecht: Reidel, 1984, 47–68.
Armstrong, David: 1983, *What is a Law of Nature?*, Cambridge: Cambridge University Press.
Baird, Davis: 1983, 'The Fisher/Pearson, Chi-squared Controversy: A Turning Point for Inductive Inference, *British Journal for the Philosophy of Science* **34**, 105–118.
Basu, Dev: 1988, 'Statistical Information and Likelihood', in: Gosh, J. K. (ed.), *Statistical Information and Likelihood. A Collection of Critical Essays by Dr. D. Basu*, New York: Springer 1988, 20–42.
Berger, James O., Berry, Donald A.: 1988, 'The Relevance of Stopping Rules in Statistical Inference (with discussion)', in Gupta, Shanti S. and Berger, James O. (eds.), *Statistical Decision Theory and Related Topics IV*, Vol. I, New York: Springer 1988, 29–72.
Bowley, Arthur L.: 1948, *Elements of Statistics*, 6th ed., London: Staples Press.
Bowley, Arthur L. and Connor, L. R.: 1923, 'Tests of Correspondence Between Statistical Grouping and Formulae', *Economica* **3**, 1923, 1–9.
Cramér, Harald: 1946, *Mathematical Methods of Statistics*, Princeton: Princeton University Press.
Casella, George: 1988, 'Conditionally Acceptable Frequentist Solutions (with discussion)', in Gupta, Shanti S. and Berger, James O. (eds.), *Statistical Decision Theory and Related Topics IV*, Vol. I, New York: Springer 1988, 73–117.
Fine, Terence: 1973, *Theories of Probability*, New York: Academic Press.
Fisher, Ronald A.: 1929, 'Tests of Significance in Harmonic Analysis', *Proceedings of the Royal Society of London* A **125**, 54–59.

Fisher, Ronald A.: 1959, *Statistical Methods and Scientific Inferences*, 2nd rev. ed., Edinburgh: Oliver and Boyd.

Fisher, Ronald A.: 1970, *Statistical Methods for Research Workers*, 14th rev. and enl. ed., New York: Hafner.

Gadenne, Volker: 1990, ‚Unvollständige Erklärungen', in Sukale, M. (ed.), *Sprache, Theorie und Wirklichkeit*, Frankfurt/M.: Lang 1990, 263–287.

Giere, Ronald N.: 1973, 'Objective Single-case Probabilities and the Foundations of Statistics', in Suppes, Patrick *et al.* (eds.), *Logic, Methodology and Philosophy of Science IV*, Amsterdam/London: North-Holland 1973, 467–483.

Giere, Ronald N.: 1977, 'Testing Versus Information Models of Statistical Inference', in Colodny, Robert G. (ed.), *Logic, Laws & Life*, Pittsburgh: University of Pittsburg Press 1977, 19–70.

Gigerenzer, Gerd; Swijtink, Zenc; Porter, Theodore; Daston, Lorraine; Beatty, John; Krüger, Lorenz: 1989, *The Empire of Chance*, Cambridge: Cambridge University Press.

Gillies, Donald A.: 1971, 'A Falsifying Rule for Probability Statements', *British Journal for the Philosophy of Science* **22**, 231–261.

Gillies, Donald A.: 1973, *An Objective Theory of Probability*, London: Methuen.

Hacking, Ian: 1965, *Logic of Statistical Inference*, Cambridge: Cambridge University Press (Seitenangaben nach: 1st paperback ed. 1976).

Hacking, Ian: 1973, 'Propensities, Statistics and Inductive Logic', in Suppes, Patrick *et al.* (eds.), *Logic, Methodology and Philosophy of Science IV*, Amsterdam/London: North-Holland 1973, 485–500.

Hacking, Ian: 1980, 'The Theory of Probable Inference: Neyman, Peirce and Braithwaite', in Mellor, David H. (ed.), *Science, Belief and Behaviour*, Cambridge: Cambridge University Press 1980, 141–160.

Hempel, Carl G. and Oppenheim, Paul: 1948, 'Studies in the Logic of Explanation', repr. in: Pitt, Joseph C. (ed.), *Theories of Explanation*, Oxford: Oxford University Press 1988, 9–46.

Hempel, Carl G.: 1977, *Aspekte wissenschaftlicher Erklärung*, Berlin: De Gruyter.

Horn, Susan D.: 1977, 'Goodness-of-fit Tests for Discrete Data: A Review and an Application to a Health Impairment Scale, *Biometrics* **33**, 237–248.

Kimball, A. W.: 1957, 'Errors of the Third Kind in Statistical Consulting', *Journal of the American Statistical Association* **52**, 133–142.

Lancaster, H. O.: 1969, *The Chi-squared Distribution*, New York: Wiley.

Lehmann, Erich, L.: 1959, *Testing Statistical Hypotheses*, New York: Wiley.

Mayo, Deborah G.: 1982, 'On After-trial Criticisms of Neyman-Pearson Theory of Statistics', in *PSA 1982*, Vol. I, 145–158.

Neyman, Jerzy: 1929, 'Contributions à la théorie de vraisemblance des hypothèses statistiques', *Revue Trimestrielle Statistique de la République Polonaise* **6**, 1–28.

Neyman, Jerzy: 1952, *Lectures and Conferences on Mathematical Statistics*, Graduate School of the U.S. Department of Agriculture.

Neyman, Jerzy: 1957, '"Inductive Behaviour" as a Basic Concept in the Philosophy of Science', *Revue l'Institute International de Statistique* **25**, 7–22.

Neyman, Jerzy: 1965, 'Behavioristic Points of View on Mathematical Statistics', in *On Political Economy and Econometrics – Essays in Honor of Oskar Lange*, Oxford: Pergamon Press 1965, 99–115.

Neyman, Jerzy und Pearon, Egon S.: 1930, 'On the Problem of Two Samples', repr. in Neyman, Jerzy and Pearson, Egon S.: *Joint Statistical Papers*, Cambridge: Cambridge University Press, 1967, 140–185.

Neyman, Jerzy und Pearson, Egon S.: 1933, 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', repr. in Neyman, Jerzy and Pearson, Egon S.: *Joint Statistical Papers*, Cambridge: Cambridge University Press, 1967, 140–185.

Popper, Karl R.: 1973, *Objektive Erkenntnis*, Hamburg: Hoffmann und Campe.

Popper, Karl R.: 1984, *Logik der Forschung*, 8. weiter verb. und verm. Aufl., Tübingen: Mohr (Siebeck).

Readhead, Michael L. G.: 1974, 'On Neyman's Paradox and the Theory of Statistical Tests', *British Journal for the Philosophy of Science* **25**, 265–271.
Spielman, Stephen: 1974, 'The Logic of Tests of Significance', *Philosophy of Science* **41**, 211–226.
Stegmüller, Wolfgang: 1973, *Jenseits von Popper und Carnap : Die logischen Grundlagen des statistischen Schließens*, broschierte Studienausgabe, Berlin: Springer.
Wald, Abraham: 1950, *Statistical Decision Functions*, New York: Wiley.
White, Douglas J.: 1976, *Fundamentals of Decision Theory*, Amsterdam: North-Holland.

Fakultät für Wirtschaftswissenschaften u. Statistik
Universität Konstanz
Postfach 5560
7750 Konstanz, Germany