BERICHT
aus dem
PSYCHOLOGISCHEN INSTITUT
DER UNIVERSITÄT HEIDELBERG

DER SIGNIFIKANZTEST IN DER PSYCHOLOGISCHEN FORSCHUNG - EIN FALSIFIKATIONSINSTRUMENT IM SINNE DES KRITISCHEN RATIONALISMUS?

von Jörg Sommer

Diskussionspapier Nr. 55 März 1987

Zusammenfassung: Untersucht wird der Signifikanztest (Hypothesentest) in der Form, wie er in der in der gegenwärtigen psychologischen Forschungspraxis üblich ist. Ist er eher als induktives oder als deduktives Verfahren zu rekonstruieren? Zur Klärung dieser Frage wurde ein logisches Instrumentarium zusammengestellt, das die Brücke herstellt zwischen dem mathematischen Kalkül des Signifikanztests und der Aussagenlogik, in deren Rahmen die Kategorien "deduktiv" und "induktiv" definiert sind. Wird die klassische zweiwertige Logik ergänzt durch eine unendlichwertige Logik mit der Wahrscheinlichkeit als Wahrheitswert einer Aussage, kann der Signifikanztest im wesentlichen als deduktives Verfahren rekonstruiert werden. Damit gelingt auch seine Einordnung in die Wissenschaftstheorie des kritischen Rationalismus, auf den sich viele Anhänger einer naturwissenschaftlich-experimentellen Psychologie berufen. Dem widerspricht lediglich die übliche Praxis, die aus der (psychologischen) Objekttheorie abgeleiteten Hypothesen mit der statistischen Alternativhypothese zu verbinden; adäquat im Sinne des kritischen Rationalismus wäre die Verbindung mit der statistischen Nullhypothese, weil nur diese "falsifiziert" (im probabilistisch abgeschwächten Sinn) werden kann.

6900 HEIDELBERG 1, HAUPTSTR. 47-51, TEL. 06221/547346

1. Einleitung

Der Hypothesen- oder Signifikanztest ist seit Jahrzehnten zentraler Bestandteil der Methodologie der akademischen psychologischen Forschung. Das Grundstudium der Psychologie ist nach Meinung der maßgeblichen Fachvertreter an den Universitäten aller fortgeschrittener Industrienationen in Ost und West ohne die Ausbildung in derartigen statistischen Verfahren nicht denkbar. Sie gehören auch zum festen Bestandteil empirischer Forschungen. Berichte darüber können in psychologischen Fachzeitschriften in der Regel nur publiziert werden, wenn sie eine statistische Auswertung aufweisen; eine Kompensation ist allenfalls durch andere mathematische Verfahren (z.B. formalisierte Modelle) möglich.

Obwohl sich in der theoretischen Entwicklung des Hypothesentests durchaus heterogene Strömungen nachweisen lassen - erinnert sei nur an die Konzepte von NEYMAN & PEARSON, FISHER und BAYES - hat sich in der akademischen Psychologie eine sehr einheitliche Handhabung des Verfahrens herausgebildet; Kontroversen gibt es gelegentlich darüber, welcher von den zahlreichen Hypothesentests für eine gegebene empirische Forschung angemessen sei. Aber das Grundschema, nach dem die verschiedenen Verfahren abgewickelt und in die wissenschaftliche Argumentation eingebaut werden, ist sehr einheitlich. Wie SPIELMAN (1974) zeigte, handelt es sich um ein Konglomerat der Konzepte von FISHER sowie NEYMAN & PEARSON. Selbst die gleichen Fehler werden bei der Anwendung des Hypothesentests immer wieder gemacht (BREDENKAMP 1972; ROSENTHAL & GAITO 1963; STELZL 1982; WOLINS 1976. Diese Einheitlichkeit des Gebrauchs zeigt sich sowohl bei der Durchsicht von Publikationen von empirischen psychologischen Untersuchungen und in der Argumentationsweise auf wissenschaftlichen Kongressen wie auch beim Studium der in der Psychologie "gängigen" statistischen Lehrbüchern wie z.B. BORTZ 1977, CLAUSS & EBNER 1972, HOFSTÄTTER & WENDT 1974, KRIZ 1978, LIENERT 1973 und 1978 und SIEGEL 1976 oder auch bei hierzulande gern benutzten amerikanischen Lehrbüchern wie z.B. HAYS 1977. Die EDV-Auswertung hat in Form der "statistical packages" (z.B. BEUTEL et al. 1980) für eine zusätzlichen Normierung in der Handhabung des Hypothesentests durch Psychologen geführt.

2. Gegenstand und Zielsetzung

Gegenstand der vorliegenden Abhandlung ist die derzeitige Praxis in der Handhabung des Hypothesentests durch Psychologen. Diese Praxis soll zunächst beschrieben werden. Unser Hauptanliegen ist dann die Klärung der Frage, ob und inwieweit diese vorfindbare Praxis als logisch stringent ausgewiesen werden kann. Insoweit wäre die Fragestellung nicht neu; erinnert sei z. B. an die Arbeiten von BAKAN (1966), BREDENKAMP (1972), GLASER (1979) und MEEHL (1967). Wo wir jedoch einen Schritt weiterzukommen hoffen, ist das Herstellen einer Verbindung zwischen dem mathematischen Kalkül des Hypothesentests auf der einen Seite und einer allgemeinen Erkenntnis- und Wissenschaftstheorie für die Psychologie auf der anderen. Auch über die Psychologie hinaus besteht "... bis zum heutigen Tage .. eine ungeheure Kluft zwischen logischen und wissenschaftstheoretischen Analysen von Begriffen der Prüfung, der Bestätigung und der Bewährung von Hypothesen auf der einen Seite und von Fachleuten auf dem Gebiet der mathematischen Statistik angestellten Untersuchungen auf der anderen Seite". STEGMÜLLER, der dies 1973 (a, 1) feststellte, konnte schließlich diese Kluft auch nicht schließen, denn seine sehr gründliche Analyse des Hypothesentests bestätigte nur eine Vermutung, die bereits andere Autoren (VETTER 1967, KLEITER 1969) aufgrund nicht so eingehender Untersuchungen geäußert hatten: Es handele sich hier um ein eigenständiges Erkenntnisinstrument, das sich nicht in die üblichen Kategorien der Wissenschaftstheorie "Verifikation und Falsifikation", "Deduktion und Induktion" einordnen lasse. Auch der Versuch eines Brückenschlags in dem ansonsten gern verwendeten und vielzitieren Lehrbuch von OPP (1970) ist fehlgeschlagen - in diesem Fall allerdings deshalb, weil diesem Autor der Hypothesentest, wie er in den Sozialwissenschaften verwendet wird, offensichtlich nicht bekannt ist.

Schon in den dreißiger Jahren ds. Jhds. hat sich POPPER (1966) mit der empirischen Überprüfung von "Wahrscheinlichkeitshypothesen", wie er sie nannte, herumgeschlagen. Er tat dies mit Blick auf die Physik und insofern auch ohne Kenntnis des Hypothesentests (der ja in der heute üblichen Form damals auch

gerade erst in der Entstehung begriffen war). Trotzdem werden wir uns mit seinen Überlegungen beschäftigen. Erstens, weil sich viele Psychologen, die den Hypothsentest verwenden, in ihrer allgemeinen wissenschaftstheoretischen Orientierung auf den von POPPER begründeten "kritischen Rationalismus" berufen (WELLENREUTHER 1982, 33f; WOTTAWA 1977, 37; ZECHA & LUKESCH 1982). Zweitens, weil POPPER dem Ziel, die Prüfung von Wahrscheinlichkeitshypothesen in sein wissenschaftstheoretisches Konzept einzuordnen, schon sehr nahe gekommen war. Wir werden die zwischen ihm und den logischen Positivisten (insbesondere CARNAP und REICHENBACH) geführte Kontroverse wieder aufgreifen, indem wir ein damals verwendetes zentrales Konzept, die mehrwertige Logik oder - wie sie damals genannt wurde - die "Wahrscheinlichkeitslogik", die Fronten wechseln lassen. CARNAP und (in etwas anderer Form) REICHENBACH versuchten, mit dieser Logik eine induktive Erkenntnistheorie zu begründen, was m. E. nicht möglich ist.

Insbesondere mußte dieser Versuch den Widerspruch POPPERs herausfordern, dessen Ablehnung der Induktion ja gerade ein Grundpfeiler seiner Wissenschaftstheorie darstellt. Die mehrwertige Logik ist POPPER vermutlich von Anfang an in Verbindung mit jenen induktivistischen Wissenschaftskonzeptionen begegnet, sodaß er mit der Zurückweisung der Induktion auch gleich die Wahrscheinlichkeitslogik über Bord geworfen hat. Und von diesem Schlag hat sich diese Form der Logik bis heute nicht erholt.

Nun könnte man sich damit zufriedengeben, den Hypothesentest, wie oben zitiert, als eigenständige Erkenntnismethode zu betrachten, wenn nicht in der Psychologie und auch in anderen Sozialwissenschaften immer wieder mehr oder weniger vage Vermutungen geäußert oder zumindest nahegelegt würden, der Hypothesentest könnte doch eine Grundlage für eine induktive Erkenntnis abgeben. So kann man in Lehrbüchern der Statistik etwa lesen, es ginge beim Hypothsentest um einen

- "Schluß von der Stichprobe auf die Grundgesamtheit" (KREYSZIG 1974, 18), oder darum,
- "making inferences from the observed sample to the unobserved population" (BAKAN 1966);
- "in der Statistik" solle "vom Besonderen des empirischen Befundes (der Beobachtung) auf induktivem Wege zu Allgemeinerem vorgestoßen werden (MENGES 1972, 25),

- "die statistische Inferenz beschäftige sich mit der Möglichkeit von Rückschlüssen aus Stichprobendaten auf die Verhältnisse in der Population" (WOTTAWA 1977, 151), oder, etwas genauer:
- es ginge darum, "zu entscheiden, ob beobachtete Unterschiede zwischen zwei Stichproben anzeigen, daß auch die dazugehörigen Populationen diese Unterschiede aufweisen" (SIEGEL, deutsche Ausgabe 1976, 2), oder, ganz allgemein
- sei "die statistische Urteilsbildung integrierender Bestandteil induktiver Erkenntnis" (CLAUSS & EBNER 1972, 19).

Die meisten Autoren führen nicht genauer aus, was sie eigentlich unter "induktiv" verstehen. Für einige (z.B. PRIM & TILMANN 1973, 106) ist eine Argumentation schon dann induktiv, wenn darin Wahrscheinlichkeitsaussagen vorkommen. Zu den gehaltvollsten Definitionen kommen die Autoren, welche die Sache mit dem "Schluß von der Stichprobe auf die Grundgesamtheit" genauer ausführen. Ein "induktiver Schluß" liegt z.B. für LEISER (1978, 114) dann vor, wenn "aus gegebenen Stichprobendaten .. Statistiken berechnet und von ... diesen auf die unbekannten Parameter eines zugrundeliegenden allgemeinen Modells geschlossen" werde. Für MENGES (1972) bilden die "empirischen Beobachtungen die 'Evidenz', d.h. die Basis der Induktion. Von dieser Basis wird - unter Beachtung bestimmter theoretischer Bedingungen - auf ein Allgemeines, in der Regel auf ein Verteilungsgesetz, geschlossen. Dieses Vorgehen dient der Überwindung der Ungewißheit" (266).

Diese Position findet der Leser von Statistik-Lehrbüchern zunächst darin bestätigt, daß er fast immer auf die folgenden beiden Kapitel stößt: Das eine behandelt das Begriffspaar "Stichprobe und Grundgesamtheit" und führt insbesondere aus, daß eine Stichprobe "repräsentativ" für die Grundgesamtheit ist, wenn sie "zufällig" ist. Das andere Kapitel betrifft die Schätzverfahren, die von dem Kennwert einer (beobachteten) Stichprobe (z.B. einem arithmetischen Mittel) ausgehen und am Ende zu einer (Punkt- oder Intervall-) Schätzung des entsprechenden Parameters in der Grundgesamtheit führen.

Verwirrung tritt spätestens dann auf, wenn sich der gleiche Leser publizierten empirischen Forschungsarbeiten zuwendet und feststellen muß, daß fast nie in der Psychologie zufällige oder sonstwie repräsentative Stichproben verwendet werden - und daß daran offensichtlich niemand Anstoß nimmt.

Wie immer man aber zu der Möglichkeit eines "Schlusses von der Stichprobe auf die Grundgesamtheit" stehen mag, eine notwendige Voraussetzung für ein solches Unternehmen ist doch wohl in jedem Fall die Repräsentativität der Stichprobe.

Diese Fragen und Probleme werden wir nicht in der gleichen Reihenfolge behandeln, in der wir sie aufgeworfen und auseinander abgeleitet haben. Um Platz zu sparen, wollen wir die Beschreibung der gegenwärtigen Praxis in der Handhabung des Hypothesentests gleich verbinden mit seiner logischen Analyse. Das setzt aber voraus, daß wir zuerst das logische Instrumentarium entwickeln, dessen wir uns dabei bedienen wollen. Dabei bietet es sich wiederum an, die erwähnte Kontroverse zwischen POPPER und REICHENBACH gleich mit einzuflechten. Zum Schluß wollen wir dann untersuchen, wie sich die mit dem Hypothesentest erzielten Resultate in die übrige wissenschaftliche Argumentation bei der Beurteilung empirischer Untersuchungen aus dem Bereich der Psychologie einordnen lassen.

3. Logisches Instrumentarium

3.1 Wahrscheinlichkeitshypothesen und Hypothesenwahrscheinlichkeiten

Aussagen über die Wahrscheinlichkeit eines Ereignisses bezeichnen wir als Wahrscheinlichkeitshypothesen oder Wahrscheinlichkeitsaussagen, z.B.: "Eine 'Eins' zu würfeln, hat die Wahrscheinlichkeit 1/6 ". Solche Aussagen spielen nicht nur in der mathematischen Statistik, sondern auch in den Objekttheorien der Sozialwissenschaften eine wichtige Rolle. In Lerntheorien kommen z.B. Wahrscheinlichkeitsaussagen vor von der Art "Wenn eine Reaktion verstärkt wird, erhöht sich die Wahrscheinlichkeit ihres Auftretens".

Bezüglich der empirischen Überprüfbarkeit solcher Aussagen besteht unter Erkenntnistheoretikern ein seit langem anhaltender Streit, den wir hier nicht in allen Einzelheiten darstellen können (vgl. z.B. MENGES 1972, 31 f, STEGMÜLLER 1971, 39 f und THOLEY, 1982). Das Grundproblem ist folgendes:

Die Wahrscheinlichkeit ist eine als solche nicht beobachtbare Eigenschaft von Gegenständen, Ereignissen oder Aussagen. Mit Hilfe der Interpretation der Wahrscheinlichkeit als relative Häufigkeit von Ereignissen hat man eine Operationalisierung versucht, die CARNAP (1959, 25) z.B. für die obige Aussage über das Würfeln so formuliert: "(Es) wird eine hinreichend lange Anzahl n von Würfen mit dem fraglichen Würfel vorgenommen und die Zahl m jener Würfe, die eine 'Eins' ergeben, gezählt. Falls die relative Häufigkeit $^{m}/_{n}$ genügend nahe bei 1/6 liegt, so wird der obige Satz als bestätigt angesehen".

Nun kann man sich darüber streiten, was unter "hinreichend lange Anzahl von Würfen" zu verstehen ist und unter "genügend nahe bei 1/6 ". Man kann hier natürlich weitere Präzisierungen vornehmen - und CARNAP hat das in seiner "induktiven Logik" getan - aber ohne willkürlich festgesetzte Konventionen wird man letztes Endes nicht auskommen.

Auch in einem anderen Ansatz zur Operationalisierung der Wahrscheinlichkeit kommt man nicht ganz ohne willkürliche Konventionen aus; sie treten nur an einer anderen Stelle auf, an der sie möglicherweise - das werden wir später sehen - einfacher strukturiert und besser handhabbar sind. In diesem Ansatz wird die Wahrscheinlichkeitshypothese (WH) in eine deterministische Hypothese (dH) transformiert, der man eine gewisse Wahrscheinlichkeit zuschreibt. REICHENBACH (1930) hat das an dem Würfelbeispiel so demonstriert:

"Ob man die Wahrscheinlichkeit den Aussagen oder den Ereignissen zuschreibt, ist nur eine Angelegenheit der Terminologie. Wir haben es bisher als eine Ereigniswahrscheinlichkeit angesehen, wenn man dem Eintreffen der Würfelseite die Wahrscheinlichkeit 1/6 zuschreibt; wir könnten ebenso sagen, daß der Aussage 'die Würfelseite 1 trifft ein' die Wahrscheinlichkeit 1/6 zukommt".

Im Gegensatz dazu sieht CARNAP (1959) fundamentale Unterschiede in den beiden Varianten. Die Wahrscheinlichkeit der Aussage "Die Würfelseite 'eins' tritt ein" ist für ihn die "induktive Wahrscheinlichkeit", der "Bestätigungsgrad" oder kurz "Wahrscheinlichkeit$_1$". Aussagen mit dieser Art von Wahrscheinlichkeit betrachtet CARNAP "auf alle Fälle als logisch determiniert und nicht synthetisch" (25). Wir schließen uns dieser Auffassung an.

Die Wahrscheinlichkeit von Ereignissen (also z.B. einer Würfelseite) bezeichnet CARNAP (1959, 25) als "statistische Wahrscheinlichkeit" oder "Wahrscheinlichkeit$_2$". Aussagen über diese Wahrscheinlichkeit hält er merkwürdigerweise für synthetisch. Sie müsse empirisch überprüft werden und zwar mit Hilfe jener "hinreichend langen Beobachtungsreihe", die wir vorn bereits zitiert haben.

Dieser Auffassung vermag ich mich nicht anzuschließen. In der Idee des Würfels (genauer: des ungefälschten Würfels) ist die Wahrscheinlichkeit von 1/6 für eine bestimmte Würfelseite schon enthalten. Auch die "Limesdefinition der Wahrscheinlichkeit" ist inbezug auf die Idee des Würfels tautologisch: "Wenn man mit einem ungefälschten Würfel unendlich viele Würfe ausführen würde, zeigte genau 1/6 dieser Würfe die Seite 'Eins'". Da gibt es nichts, was empirisch zu überprüfen wäre.

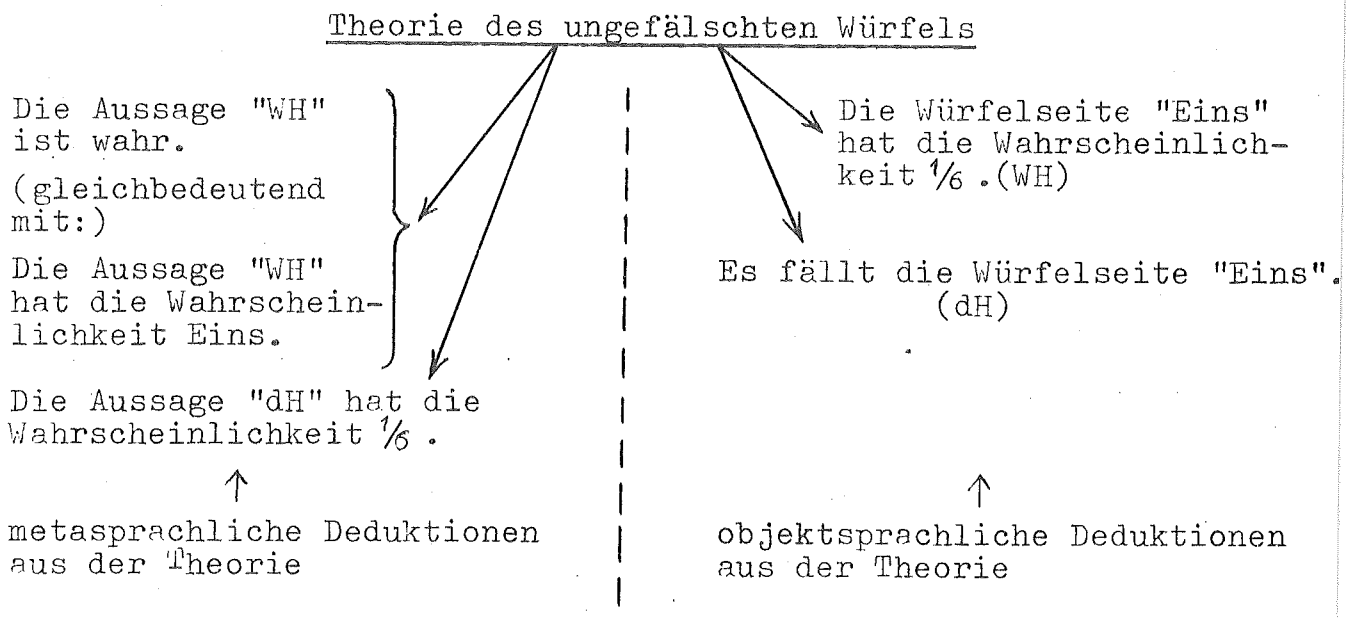
Merkwürdig an der CARNAPschen Position ist auch der ausschliessende Gegensatz, den er zwischen empirisch prüfbaren und logisch abgeleiteten Aussagen behauptet: "Im Gegensatz zu den Wahrscheinlichkeits$_1$-Aussagen drückt eine Wahrscheinlichkeits$_2$-Aussage keine logische Relation aus, sondern besitzt einen Tatsachengehalt" (25). Für eine Falsifikation von Theorien im POPPERschen Sinn braucht man Aussagen, die b e i d e s vereinen: Sie müssen logisch aus der zu prüfenden Theorie abgeleitet sein (also tautologisch inbezug auf die Theorie sein), gleichzeitig aber einen beobachtbaren Sachverhalt behaupten, also einen "Tatsachengehalt" haben. Die Aussage: "Wenn dieser Kupferstab um soundsoviel Grad erwärmt wird, verlängert er sich um soundsoviel Millimeter" kann logisch zwingend aus einer thermodynamischen Theorie abgeleitet werden und ist gleichzeitig empirisch prüfbar, also brauchbar für eine Falsifikation dieser Theorie.

Betrachten wir unter dieser Voraussetzung die beiden Varianten von Aussagen über das Würfeln, kommen wir zu einem wichtigen Ergebnis: Nur die eine von beiden, nämlich die deterministische Hypothese dH, erfüllt beide Forderungen: Sie ist einerseits aus der "Theorie des ungefälschten Würfels" logisch deduziert, behauptet aber gleichzeitig einen beobachtbaren Sachverhalt, nämlich eine "Eins" zu würfeln. Die andere Hypothese (WH) ist zwar auch logisch deduziert, aber empirisch nicht prüfbar - jedenfalls nicht als solche, sondern nur über die angreifbare Hilfskonstruktion einer endlichen Beobachtungsreihe.

Daß nicht alle Aussagen, die aus einer Theorie ableitbar sind, auch empirisch prüfbar sind, überrascht vielleicht nicht sonderlich. Verwirrend ist im vorliegenden Fall aber, daß es in beiden Ableitungen um verschiedene Arten von Wahrscheinlichkeit geht, die in der "Theorie des ungefälschten Würfels" noch eine Einheit bilden: Dort tritt diese Wahrscheinlichkeit ja nur als die "Eigenschaft eines physikalischen Systems" auf (CARNAP 1959, 21). In WH ist daraus die Wahrscheinlichkeit eines Ereignisses geworden, während im andern Fall die Wahrscheinlichkeit einer Aussage (dH) behauptet wird. Im letzteren Fall handelt es sich offensichtlich um eine Aussage über eine Aussage und damit um eine Konstruktion auf mehreren Sprachebenen (Objektsprache und Metasprache).

Das führt uns zu der Frage, ob dieses Beispiel nicht zu verallgemeinern sei. In der Tat wollen wir im folgenden die These belegen, daß jede Deduktion aus einer Theorie auf zwei Sprachebenen erfolgt: Eine deduzierte Aussage ist objektsprachlich und behauptet etwas über den Gegenstandbereich der Theorie. Die andere ist metasprachlich und behauptet etwas über die Wahrscheinlichkeit der objektsprachlichen Aussage. Bevor wir auf das Würfelbeispiel zurückkommen, dazu noch ein allgemeineres Beispiel: Wenn die Aussage: "Morgen wird es regnen" aus einer meteorologischen Theorie abgeleitet ist, verlangt man mit Recht, daß die gleiche Theorie auch eine Angabe über die Sicherheit dieser Prognose macht, etwa so: "Die obige Aussage wird sich mit der und der Wahrscheinlichkeit als richtig erweisen" oder mindestens dies: "Die obige Aussage ist sicher"bzw. "... nicht sicher".

Für das Würfelbeispiel ergeben sich die folgenden Möglichkeiten:



Vergleichen wir die metasprachlichen Aussagen, können wir feststellen, daß die ersten beiden von einer zweiwertigen Logik Gebrauch machen mit den Geltungswerten "wahr" und "falsch" (1 und 0), während in der letzten eine unendlichwertige Logik mit den Geltungswerten von 0 bis 1 (in unserem Fall 1/6 = 0, 166..) verwendet wird.

Damit wird eine Verallgemeinerung der zweiwertigen Logik nutzbar gemacht, die ŁUKASIEWICZ (1920) und POST (1921) - bezugnehmend auf antike Vorbilder - zunächst als mehrwertige Logik einführten . ŁUKASIEWICZ (1930) verallgemeinerte weiter zu einer unendlichwertigen Logik und wies auch bereits auf die Beziehungen zu dem Wahrscheinlichkeitskalkül hin (vgl. auch SINOWJEW 1968, 32).

So weit ich die Literatur überblicke, wurden jedoch in der folgenden Zeit die Möglichkeiten der unendlichwertigen Logik nicht mehr genutzt. Unter der Bezeichnung "Wahrscheinlichkeitslogik" wurde sie stets - wohl durch den Einfluß CARNAPs - als "induktive Logik" betrachtet, die eigenen Regeln folgt (so z.B. HEMPEL 1977, 101; Orig. 1965).

Auch den Begriff des "logischen Spielraums einer Aussage" von v. KRIES (vgl. VETTER 1967, 31) können wir in dieses Konzept einer unendlichwertigen Logik integrieren und gleichzeitig genauer darlegen, wie wir uns die Ableitung der metasprachlichen Aussagen aus einer Theorie vorstellen.

Den logischen Spielraum einer Aussage definieren wir als den Anteil der Verifikatoren im Ereignisraum. Der Ereignisraum (z.B. MENGES 1972, 84) bestehe aus allen denkbaren Überprüfungen der fraglichen Aussage. Wir wollen das am Beispiel der beiden Aussagen WH (Wahrscheinlichkeitshypothese) und dH (deterministische Hypothese mit einer Hypothesenwahrscheinlichkeit zwischen Eins und Null) verdeutlichen. Um die Aussage WH zu prüfen, stellt man sich vor ("Gedankenexperiment"), immer wieder einen ungefälschten Würfel zu betrachten und zu überlegen, ob die Wahrscheinlichkeit für eine "Eins" ⅙ beträgt. Natürlich wird jede Nachprüfung die Aussage WH bestätigen, weil in dem Begriff "ungefälschter Würfel" bereits die "Wahrscheinlichkeit ⅙ für eine 'Eins'" enthalten ist. Andernfalls wäre es eben kein ungefälschter Würfel! Der Ereignisraum enthält also in diesem Fall ausschließlich Verifikatoren; ihr Anteil und damit die Wahrscheinlichkeit der Aussage WH ist Eins.

Betrachtet man den auf analoge Weise gebildeten Ereignisraum zur Aussage dH, wird man feststellen, daß genau ⅙ seiner Elemente Verifikatoren sind, seine Wahrscheinlichkeit also ⅙ beträgt. Aber auch das folgt logisch zwingend aus dem Begriff

(der "Theorie") des ungefälschten Würfels. Insofern hat CARNAP Recht, wenn er solche Aussagen als "analytisch" betrachtet. Aber auch VETTER (1967, 31) hat Recht, wenn er hier generell von "faktischen" Aussagen spricht; denn der objektsprachliche Teil ("Es fällt eine 'Eins'") ist a u c h empirisch prüfbar.

Bisher sahen wir uns nicht genötigt, irgendeine Konvention oder willkürliche Setzung einzuführen. Man k a n n das tun, wenn man aus dem Bereich der unendlichwertigen Logik wieder zu der (uns vertrauten) zweiwertigen Logik zurückkehren will. Dann muß ein Intervall der Geltungswerte "ausgezeichnet" werden (BOCHEŃSKI & MENNE 1973, 120), für das eine Aussage der unendlichwertigen Logik als "wahr" bzw. "falsch" im Sinne der zweiwertigen Logik gelten soll. Indem wir die in der Statistik übliche Notation vorwegnehmen, legen wir ein "Signifikanzniveau α" im Intervall (0 ... 1) willkürlich fest. Der Wert von α soll nur die Bedingung erfüllen, nahe bei Null zu liegen. Üblich sind in den Sozialwissenschaften Werte von 0,01 und 0,05, gelegentlich auch noch 0,1 oder o,001. Als "wahr" im Sinne der zweiwertigen Logik wird dann eine Aussage bezeichnet, deren Geltungswert im Intervall $\{(1-\alpha) \ldots 1\}$ liegt und als "falsch", wenn der Geltungswert im Restintervall liegt, also in $\{0 \ldots (1-\alpha)\}$.

Während aber für die Operationalisierung von Wahrscheinlichkeitshypothesen willkürliche Setzungen unvermeidlich sind, braucht man sie bei deterministischen Hypothesen (mit einer bestimmten Wahrscheinlichkeit) nicht notgedrungen. Insbesondere sind auch die obigen Konventionen für den Hypothesentest nicht zwingend erforderlich, sondern nur eine Frage der Arbeitserleichterung. Im Prinip kann nämlich jeder Test als sog. "exakter Test" ausgeführt werden, bei dem die "Irrtumswahrscheinlichkeit" ihrem Wert nach berechnet wird, anstatt nur zu prüfen, ob ein Signifikanzniveau über- oder unterschritten wird. Deshalb ist einer Operationalisierung probabilistischer Theorien über deterministische Hypothesen mit einer bestimmten Hypothesenwahrscheinlichkeit (unendlichwertige Logik) der Vorzug zu geben gegenüber einer Operationalisierung über Wahrscheinlichkeitshypothesen.

3.2 Logische Schlüsse

Um die Aussage "eine 'Eins' wird gewürfelt" mit einem Geltungswert von ⅙ behaupten zu können, genügt als hinreichende Voraussetzung die Unverfälschtheit des Würfels. Wir kommen damit zu der Aussagenverbindung (vgl. auch ZAWIRSKI 1935):

P_1 Wenn ein Würfel ungefälscht ist, dann (kann mit dem Wahrheitswert ⅙ behauptet werden:) würfelt man eine "Eins".

P_2 Der Würfel ist ungefälscht.

(Daraus folgt mit einem Wahrheitswert von ⅙):

K Man würfelt eine 'Eins'."

In diesem Syllogismus ist P_1 die 1. Prämisse - Gesetzesaussage. P_2 ist die 2. Prämisse - Randbedingung. K ist die Konklusion. Die metasprachlichen Teile des Ausdrucks sind in Klammern gesetzt.

Notwendig ist die Unverfälschtheit des Würfels nicht, um die Aussage "man würfelt eine 'Eins'" mit einem Geltungswert von ⅙ zu behaupten, denn Fälschungen eines Würfels mögen andere Augenzahlen beeinflussen, die Wahrscheinlichkeit der "Eins" aber unberührt lassen. Andererseits folgt die Konklusion aus den Prämissen mit dem angegebenen Geltungswert zwingend, weshalb wir hier eine "strenge Implikation" vor uns haben. In der zweiwertigen Logik verwenden wir dafür die Schreibweise $p \rightarrow q$, in der mehrwertigen Logik $p \overset{w}{\Rightarrow} q$. Darin sind p und q Aussagenvariablen, w steht für den Geltungswert*. Der nach rechts gerichtete Pfeil bedeutet "... folgt zwingend...". Auf der Grundlage dieser strengen Implikation sind die folgenden Schlüsse zulässig (ZAWIRSKI 1935):

2-wertige Logik	Bezeichnung der Syllogismen	mehrwertige Logik*
$\dfrac{p \rightarrow q \quad p}{q}$	modus ponendo ponens	$\overset{w}{\dfrac{p \overset{w}{\Rightarrow} q \quad p}{q}}$
$\dfrac{p \rightarrow q \quad \neg q}{\neg p}$	zweiter modus tollendo tollens	$\overset{w}{\dfrac{p \overset{w}{\Rightarrow} q \quad \neg q}{\neg p}}$

Das obige Würfelbeispiel entspricht dem modus ponendo ponens. Dabei steht "der Würfel ist ungefälscht" für p und "man würfelt eine 'Eins'" für q. $\neg$ p steht für "nicht - p".

Berücksichtigen wir nun noch die oben erwähnte Möglichkeit, Geltungswerte der mehrwertigen Logik auszuzeichnen, können wir das

*) Mit der Schreibweise lehne ich mich an OPP (1970), 44 an.

Signifikanzniveau mit $\alpha = 0{,}05$ (willkürlich) festlegen. Dann kann die Aussage "wenn ein Würfel ungefälscht ist, dann würfelt man eine 'Eins'" nicht mehr behauptet werden, da ihr Geltungswert mit $w = 1/6 = 0{,}17$ nicht in das ausgezeichnete Intervall von $\{(1-\alpha) \ldots 1\} = \{0{,}95 \ldots 1\}$ fällt. Das ist im gegebenen Fall auch intuitiv einsichtig: Die Tatsache, daß ich (selbst auf Anhieb, was in allen diesen Beispielen vorausgesetzt wird) eine "Eins" gewürfelt habe, besagt noch gar nichts. Aus mehrmaligem Würfeln wären schon eher Rückschlüsse auf die Beschaffenheit des Würfels zu ziehen. Ein entsprechender Syllogismus könnte lauten:

Wenn ein Würfel ungefälscht ist,
$(w = 0{,}972)$*
erscheinen bei dreimaligem Würfeln mindestens zwei verschiedene Zahlen.

Der Würfel ist ungefälscht.
(daraus folgt mit $w = 0{,}972$)
Bei dreimaligem Würfeln erscheinen mindestens zwei verschiedene Zahlen.

Jetzt liegt der Geltungswert in dem ausgezeichneten Intervall $(0{,}972 > 0{,}95)$, der Ausdruck kann also behauptet werden.

Der modus tollendo tollens lautet in der unendlichwertigen Version:

Wenn ein Würfel ungefälscht ist, $(w = 0{,}972)$ erscheinen beim dreimaligen Würfeln mindestens zwei verschiedene Zahlen.	$p \underset{0.972}{\Rightarrow} q$
Es erscheinen n i c h t mindestens zwei verschiedene Zahlen. $(w = 0{,}972)$	$\neg q$
Der Würfel ist gefälscht.	$\underset{0.972}{\neg p}$

Entsprechend den im vorangehenden Abschnitt angestellten Überlegungen kann dieser Ausdruck behauptet werden. Auch das ist intuitiv einsichtig. "$\neg q$" besagt ja, daß ich mit einem Würfel a u f A n h i e b dreimal die gleiche Zahl werfe, und das gibt hinsichtlich der Ungefälschtheit des Würfels in der Tat zu denken. Andererseits bin ich natürlich nie absolut sicher, auch einmal mit einem ungefälschten Würfel auf Anhieb dreimal die gleiche Zahl zu werfen, und das kommt darin zum Ausdruck, daß der Geltungswert des obigen Ausdrucks eben nicht 1 ist,

)* Die Zahl 0,972 ergibt sich aus $1-6(1/6)^3$; jede Serie aus drei gleichen Würfelzahlen hat die Wahrscheinlichkeit $(1/6)^3$. Möglich sind aber sechs solcher Serien.

sondern etwas darunter liegt. Die Differenz zu 1 - also 0.028 - können wir (ebenfalls unter Vorwegnahme der in der Statistik üblichen Terminologie) als "Irrtumswahrscheinlichkeit" bezeichnen.

Als weitere Relation zwischen Aussagen benötigen wir noch die Äquivalenz. Sie liegt bei notwendigen u n d hinreichenden Bedingungen vor. Ein Beispiel für die unendlichwertige Logik wäre der Münzwurf: Für die Aussage "Ich werfe eine Zahl" mit einem Geltungswert von ½ ist jetzt die Unverfälschtheit nicht mehr nur hinreichend, sondern auch notwendig (wobei natürlich vorausgesetzt wird, daß sich die "Unverfälschtheit" auf das Verhalten der Münze beim Werfen bzw. beim Fallen bezieht). Die Aussagen "Ich werfe eine Zahl" (mit w = ½) und "die Münze ist unverfälscht" (mit dem gleichen Geltungswert) sind äquivalent, in einem Text können sie gegeneinander ausgetauscht werden. Man schreibt formalisiert $p \leftrightarrow q$. - Alle Syllogismen, die mit einer implikativen Aussagenverbindung möglich sind, sind auch erlaubt bei einer äquivalenten Aussagenverbindung (SEGETH 1973, 56 - 63), darüberhinaus aber auch noch der Schluß vom Nachsatz auf den Vordersatz (replikative Abschwächung der Äquivalenz - BOCHEŃSKI & MENNE 1973, 45; Umkehr der Replikation in eine Implikation - BOCHEŃSKI & MENNE 1973, 31 und Anwendung des modus ponendo ponens). Die unendlichwertige Version wird analog der vorausgehenden Fälle gebildet.

$$\frac{p \leftrightarrow q \quad q}{p}$$

Andere Relationen zwischen Aussagen (z.B. die Replikation), sowie andere Syllogismen werden wir für die Rekonstruktion des Signifikanztests nicht benötigen, sodaß ich hier nicht auf sie einzugehen brauche.

3.3 Deduktive und induktive Logik

Um einen Schluß als "induktiv" charakterisieren zu können, ist es zweckmäßig, von formalen (syntaktischen) Kriterien der Schlußfigur auszugehen, da bei einer inhaltlichen Analyse zu leicht Mißverständnisse durch eine unterschiedliche Interpretation entstehen können. Das auf ŁUKASIEWICZ (1913) zurückgehende System von BOCHEŃSKI (1965) unterscheidet von den deduktiven Schlußformen zunächst die reduktiven. Während bei den deduktiven Schlüssen vom

Vordersatz auf den Nachsatz geschlossen wird (z.B. Vordersatz: "Wenn es regnet...", Nachsatz: "...dann ist die Straße naß" - es regnet - also ist die Straße naß), geht man beim reduktiven Schließen in umgekehrter Richtung vor, schließt also vom Nachsatz auf den Vordersatz. Diese Art des Schließens ist logisch nicht zwingend. Der induktive Schluß ist ein Sonderfall des reduktiven, in dem nur anstatt des Einzelfalls ("die Straße ist naß") mehrere Fälle stehen, z.B.:

1. Prämisse	Vordersatz:	Wenn die Straße immer naß ist, wenn es regnet,
	Nachsatz:	dann ist sie auch naß, wenn es regnet, in den Fällen A, B, C usw.
2. Prämisse:		Die Straße ist naß, wenn es regnet, in den Fällen A, B, C usw.
Schluß (Konklusion):		Also ist die Straße immer naß, wenn es regnet.

Dieser Formalismus soll selbstverständlich keine Lösung des "Induktionsproblems" darstellen. Entweder, man akzeptiert, daß induktive Schlüsse nicht zwingend sind - und dann kann man weiter von einer "induktiven Logik"reden. Oder, man legt fest, daß Schlüsse entweder logisch zwingend sind oder es keine Schlüsse sind - dann gibt es eben keine "induktive Logik". Der obige Formalismus zeigt auch sehr deutlich, daß sich das "Induktionsproblem" nicht durch die Einführung des Wahrscheinlichkeitsbegriffs lösen läßt. Ich mag die obigen Fälle A, B, C erweitern auf hundert, tausend oder zehntausend - niemals kann ich eine Wahrscheinlichkeit daraus ableiten für den Satz "Die Straße ist immer naß, wenn es regnet".

Von dieser BOCHEŃSKIschen Position aus gesehen ist Vieles nicht-induktiv, was von einigen Sozialwissenschaftlern als induktiv betrachtet wird (OPP 1970, 36-44; PRIM & TILMANN 1973, 106-108; GROEBEN & WESTMEYER 1975, 84-86). Dort wird immer noch die CARNAPsche bzw. HEMPELsche Position aufrechterhalten oder zumindest referiert, ohne daß inzwischen eine Lösung der dort aufgetretenen Probleme gefunden werden konnte (STEGMÜLLER 1973 a)

Die Beispiele aus der mehrwertigen Logik, die ich bisher angeführt habe, sind alle deduktiv. Sofern sie dem modus ponendo ponens entsprechen, wird auch hier vom Vorder- auf den Nachsatz geschlossen (vgl. das Bsp. am Anfang des Abschnittes 3.2). Damit genügen sie dem BOCHEŃSKIschen Kriterium für deduktive Schlüsse. Der deduktive Charakter meiner Beispiele ist aber auch erkennbar, wenn man verbale Beschreibungen der fraglichen Schlußweisen heranzieht, z.B. die von STEGMÜLLER (1973 b , 76 - 77): Induktiv sind "wahrheitskonservierende Erweiterungsschlüsse". "Wahrheitskonservierend heißt für ihn, daß, "falls die Prämissen wahr sind, ... sich diese Wahrheit auf die Conclusio überträgt"; Erweiterungsschlüsse liegen vor, wenn die "Conclusio dem logischen Gehalt nach stärker ist als die Klasse der Prämissen". "Wahrheitskonservierend" in diesem Sinne sind unsere Beispiele, aber es sind keine "Erweiterungsschlüsse". In der Konklusion wird nichts ausgesagt, was in den Prämissen nicht schon enthalten wäre: Es handelt sich durchwegs um tautologische Aussagenverbindungen, und das ist ein sicheres Merkmal deduktiver Schlüsse.

4. Die Analyse des Hypothesentests

Es gibt zahlreiche Darstellungen der logischen Struktur des Hypothesentests, z.B. von BREDENKAMP (1972), GLASER (1969), HACKING (1965), STEGMÜLLER (1973 a), ZAHLEN (1966). Unter "logisch" verstanden aber alle diese Autoren die der Mathematik immanente Folgerichtigkeit. Für sie war demnach die Aufgabe gelöst, wenn sie die mathematische Struktur des Hypothesentests transparent gemacht hatten. Das muß auch unsere Analyse leisten, und insoweit können wir auf den genannten Arbeiten aufbauen. Zusätzlich wollen wir aber die Brücke zur Aussagenlogik schlagen und danach müssen wir die elementaren Einheiten unserer Analyse wählen. Diese Einheiten oder Schritte, in die wir den Hypothesentest zerlegen, sollen also einerseits den allseits anerkannten Strukturierungen der o.g. Autoren nicht widersprechen, müssen aber andererseits in möglichst elementare Strukturen der Aussagenlogik abbildbar sein. Wer die folgende Analyse mit den Augen des mathematischen Statistikers betrachtet, wird daher nichts wesentlich Neues finden. Unsere Analyse wird erst dann - wie wir hoffen - interessant, wenn man sie vom Standpunkt des Brückenbauers von der Mathematik zur Logik beurteilt.

Eingangsgrößen:

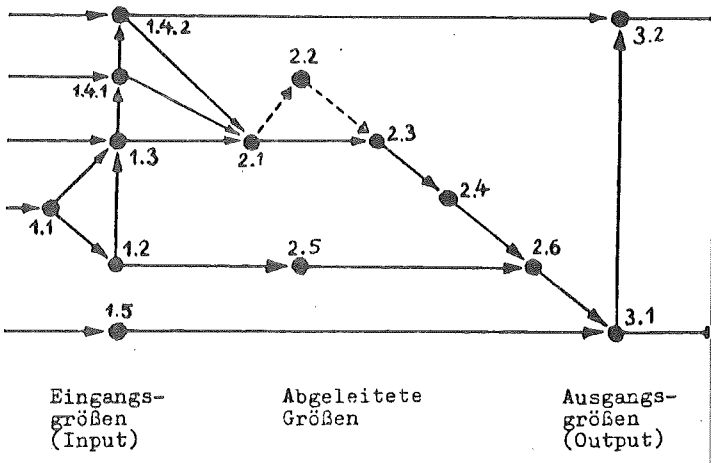
1.1 Quantifizierte empirische Daten
1.2 Typ der prüfgröße
1.3 Umfang, empirische Verankerung und meßtheoretische Charakteristika der Grundgesamtheit
1.4 Statistische Hypothesen
1.4.1 Nullhypothese
1.4.2 Alternativhypothese
1.5 Signifikanzniveau

Abgeleitete Konstrukte:

2.1 Verteilung der abhängigen Variablen in der Grundgesamtheit
2.2 Parameter der Grundgesamtheit
2.3 Stichprobenraum
2.4 Prüfverteilung
2.5 beobachteter Wert der Prüfgröße
2.6 Wahrscheinlichkeit der Prüfgröße

Ausgangsgrößen:

3.1 Entscheidung bezüglich der Nullhypothese
3.2 Entscheidung bezüglich der Alternativhypothese



Hauptbestandteile des Hypothesentest

Die Pfeile sind zu interpretieren als "... führt zu ...". Gestrichelte Pfeile führen zu Konstrukten, die nicht bei allen Tests benötigt werden.

Wir können den Hypothesentest als offenes System bezeichnen. Als Eingangsgrößen (input) sind empirische Daten und Festlegungen des Signifikanzniveaus, der Prüfgröße(n) und der statistischen Hypothesen erforderlich; Ausgangsgrößen (output) ist die Auszeichnung einer der beiden statistischen Hypothesen mit einem Geltungswert.

4.1 Eingangsgrößen

4.1.1 Quantifizierte empirische Daten (abhängige Variable)

Indem man Beobachtungsresultate ein-eindeutig oder mindestens eindeutig auf ein Zahlensystem abbildet, erhält man eine endliche Menge quantitativer Daten, die wir als Stichprobe bezeichnen. Eingangsgrößen für das System des Hypothesentests sind sie in zweierlei Hinsicht: Einmal in ihrer Eigenschaft als Zahlen, mit denen gerechnet werden kann. Zweitens aber auch als Daten, die unter bestimmten, definierten Bedingungen erhoben wurden und die insofern gewissen Voraussetzungen genügen oder nicht. STEGMÜLLER (1973a, 128 - 131) bezeichnet sie als "statistische Oberhypothesen", ZAHLEN (1966) als "a-priori-Hypothese".

4.1.1.1 Zufällige Veränderliche

Durch die Versuchs- bzw. Beobachtungsbedingungen muß sichergestellt sein, daß die Stichprobendaten als Realisation einer zufälligen Veränderlichen betrachtet werden können. Eine notwendige Voraussetzung hierfür ist es, daß die Daten überhaupt variieren, also eine Varianz größer als Null haben (andernfalls wären es keine Variablen, sondern Konstante). Diese Bedingung ist erfüllt, wenn sich wenigstens ein Datum in der Stichprobe seinem Wert nach von den anderen unterscheidet.

Zufällig ist die Veränderliche, wenn die Wahrscheinlichkeit $p(x_i)$, daß die von der Versuchsperson Vp_i stammende Realisation der Veränderlichen X, nämlich x_i, die Realisationswahrscheinlichkeit $p(x_j)$ eines anderen Wertes der Veränderlichen X, nämlich x_j, die von der Vp_j stammt, nicht beeinflußt: $p(x_j) = p(x_j / x_i)$ (MENGES 1972, 96).

Diese Forderung kann auch bei nicht-zufälligen Stichproben erfüllt sein, wie sie in der Psychologie häufig verwendet werden (zufällig wäre eine Stichprobe, wenn jeder Merkmalsträger der Grundgesamtheit die gleiche Chance hätte, in die Stichprobe zu kommen; siehe z.B. STEGMÜLLER 1973 a, 135). Nicht erfüllt wäre dagegen das Kriterium von MENGES, wenn sich die Vpn gegenseitig beeinflussen würden (z.B. voneinander abschreiben, Reaktionen nachahmen, sich anspornen oder behindern würden usw.). Dem versucht man in der Psychologie durch Vereinzelung der Vpn., Unterbinden von Gerüchten über die Untersuchung usw. vorzubeugen. Das obige Kriterium wäre ebenfalls verletzt durch Clusterbildung beim Herstellen von Teilstichproben, sodaß (zwar nicht die A u s w a h l , wohl aber) die A u f t e i l u n g der Vpn streng nach Zufall erfolgen muß. Das schließt z.B. ein Mitspracherecht der Vpn bei ihrer Zuordnung zu Kontroll- oder Experimentalgruppen aus.

Die Voraussetzung einer zufälligen Veränderlichen ist auch bei sog. "abhängigen Stichproben" zu realisieren, die z.B. dann entstehen, wenn dieselben Vpn mehrmals unter verschiedenen Bedingungen beobachtet werden. Die Daten sind dann zwar abhängig zwischan den Bedingungen, innerhalb einer Bedingung aber unabhängig.

4.1.1.2 Diskrete und stetige Veränderliche

Eine weitere Eigenschaft der Daten bestimmt sich aus der Meßvorschrift, die festlegt, ob den Beobachtungen beliebige reele Zahlen zugeordnet werden können (stetige Veränderliche), oder ob bestimmte Intervalle, die nicht zuordenbar sind, die Veränderliche von einem Wert zum anderen "Sprünge" machen läßt (diskrete Veränderliche).

4.1.1.3 Skalenniveau

Inwieweit auch das Skalenniveau der zufälligen Veränderlichen für einen Signifikanztest definiert sein muß, ist umstritten. BREDENKAMP (1972, 123-124) stellt fest: "So führt ANDERSON (1961) etwa aus, daß eine parametrische Varianzanalyse mit einer ordinal skalierten Variablen durchgeführt werden könne, wenn die statistischen Annahmen der Normalverteilung und der Varianzhomogenität erfüllt seien. Niemand hat diesem Argument unseres Wissens bisher widersprochen. Die Statistik ist - darüber dürften sich alle Autoren, STEVENS eingeschlossen (vgl. STEVENS 1968), einig sein - gegenüber dem Skalenniveau der Zahlen neutral".

Andererseits werden in allen Lehrbüchern der Statistik für jeden Test bestimmte Skalenniveaus als Voraussetzung genannt. Behauptet werden kann in den Sozialwissenschaften meist aber nur die Nominal- und die Ordinalskala. Die Annahme der Intervallskala bzw. höherer Skalentypen als (angebliche) Voraussetzung der am häufigsten verwendeten Tests, nämlich der parametrischen, muß in der Regel als Hypothese stehen bleiben.

Die logische Rekonstruktion des Signifikanztests wird zeigen, daß an keiner Stelle Voraussetzungen bezüglich des Skalenniveaus gemacht werden müssen. Seine Bedeutung liegt in den Bereichen, die im Prozeß der Gewinnung und Verarbeitung empirischer Daten vor und nach dem Signifikanztest angesiedelt sind, also in der Operationalisierung theoretischer Konstrukte und meßtheoretischer Überlegungen einerseits und bei dem Rückbezug empirischer Daten auf die Theorie andererseits.

4.1.1.4 Verteilungsform

Dieses Merkmal der Stichprobendaten wird in Statistikbüchern meist nicht explizit genannt, spielt aber gleichwohl in der Praxis der Testanwendung eine große Rolle. Insbesondere muß anhand der

Verteilungsform der Daten entschieden werden, ob die Normalverteilungs-Hypothese für die Variable in der Grundgesamtheit beibehalten werden kann, was für die so weit verbreiteten parametrischen Tests von Belang ist (sofern man bei größeren Stichproben nicht mit dem zentralen Grenzwertsatz operieren kann). Das Vorgehen besteht darin, daß man die Häufigkeitsverteilung der Stichprobendaten auf ihre Symmetrie und annähernd glockenförmige Gestalt hin untersucht. Größere Abweichungen von diesen Merkmalen führt zur Verwerfung der Normalverteilungshypothese und zur Wahl eines nicht-parametrischen Tests.

4.1.2 Typ der Prüfgröße

Aus der inhaltlichen Fragestellung, die der Datenerhebung zugrunde liegt, läßt es sich ableiten, in welcher Stichprobenstatistik die gewünschte Information repräsentiert ist. In vielen Fällen handelt es sich dabei um ein Maß für die zentrale Tendenz einer Verteilung, z.B. das arithmetische Mittel bzw. die Differenz zwischen zwei derartigen Mittelwerten; in der Varianzanalyse kommt es auf das Verhältnis von je zwei Varianzen an usw. Diese als Prüfgröße bezeichnete Stichprobenstatistik ist jetzt nur als Typ zu charakterisieren, d.h. es muß eine Rechenoperation angegeben werden, die zu der Berechnung der Prüfgröße verwendet wird.

4.1.3 Umfang, empirische Verankerung und meßtheoretische Charakteristika der Grundgesamtheit

Lehrbücher der psychologischen Methodenlehre bzw. Statistik vermitteln oft den Eindruck, es seien nur solche Stichproben für eine statistische Auswertung brauchbar, die, um repräsentativ für eine definierte Grundgesamtheit oder Population zu sein, zufällig aus dieser gezogen sein müßten. In der Forschungspraxis der Psychologie kommen dagegen solche zufälligen Stichproben nur sehr selten vor. In den meisten Fällen handelt es sich um "anfallende" Stichproben aus Personen, die sich freiwillig zur Verfügung stellen oder gegen Bezahlung an der Untersuchung teilnehmen. Da diese Praxis offenbar als unproblematisch betrachtet wird, scheint ihr eine andere Logik als die der realen Grundgesamtheiten und zufälligen Stichproben zugrundezuliegen.

Das erhellt auch daraus, daß faktische Zufallsstichproben nur aus endlichen Grundgesamtheiten gezogen werden können (da jedes Element der Grundgesamtheit die gleiche Chance haben müßte gezogen zu werden). Für solche endlichen Grundgesamtheiten wären aber Korrekturen der statistischen Schätzverfahren erforderlich ("finite population factor"). Diese werden in den Statistikbüchern zwar gelegentlich erwähnt (z.B. HAYS 1973, 287-288), aber in der Forschungspraxis so gut wie nie angewandt.

Dieser Praxis werden wir daher eher gerecht, wenn wir die Grundgesamtheiten als rein fiktive Gebilde betrachten, die dadurch zustandekommen, daß man sich die mit der verwendeten Stichprobe praktizierte Datenerhebung unendlich oft unter denselben Bedingungen wiederholt vorstellt.

Dieses Konzept unterscheidet sich dann aber in nichts mehr von dem an dem Würfelbeispiel demonstrierten "Ereignisraum" (3.1), sodaß die dort entwickelten Überlegungen zum Umgang mit probabilistischen Theorien voll auf das Konzept der Grundgesamtheit im Hypothesentest anwendbar sind. Auseinandersetzen müssen wir uns lediglich noch mit der Auffassung BREDENKAMPs (1972, 28f), daß eine Teilklasse der Hypothesentests (die sog. Randomisierungstests) die Annahme einer (auch nur fiktiven) Grundgesamtheit ganz überflüssig machen würden. Diese Auffassung werden wir zurückweisen (4.4.3.2), sodaß wir unsere weiteren Rekonstruktionsversuche auf der These aufbauen können: Jedem Hypothesentest liegt die Vorstellung einer fiktiven Grundgesamtheit zugrunde, welche identisch ist mit dem Konzept des "Ereignisraumes" in der Wahrscheinlichkeitstheorie.

Grundsätzlich werden alle hypothetischen und empirischen Feststellungen über den Typ der abhängigen Variablen, die inbezug auf die Stichprobendaten getroffen wurden, auch zur Definition der Veränderlichen in der Grundgesamtheit verwendet. Man geht also davon aus, daß es sich um eine zufällige Veränderliche in der Grundgesamtheit handelt, die das gleiche Skalenniveau hat wie die Daten der Stichprobe und ebenso wie diese entweder diskret oder stetig verteilt ist.

Darüberhinaus muß vorausgesetzt werden, daß für diese Veränderliche ein arithmetisches Mittel existiert und die Varianz größer als Null ist. Hinreichende Bedingung hierfür ist es, daß die Veränderliche stochastisch konvergiert und keine Einpunktverteilung hat (FISZ 1966, 59 und 178 sowie KREYSZIG 1974, 86). Die erste

dieser beiden Bedingungen ist in den Sozialwissenschaften praktisch immer erfüllt, da durch die Meßvorschriften aus Gründen der Praktikabilität die Variable nach oben und unten begrenzt wird, d.h. ihr absoluter Betrag kann nicht beliebig groß werden; auch theoretisch gilt darüberhinaus in vielen Fällen, daß die Wahrscheinlichkeit einer Variablen X gegen Null geht, wenn sie gegen $+\infty$ oder $-\infty$ strebt: $\lim_{|X| \to \infty} p(|X|) = 0$.

Eine Einpunktverteilung ist auszuschließen, wenn sich in der Stichprobe mindestens ein Wert der betreffenden Veränderlichen befindet, der sich von den anderen unterscheidet. Auch dieser empirische Sachverhalt ist in den Sozialwissenschaften praktisch immer gegeben.

4.1.4 Statistische Hypothesen

Die Fragestellung einer Untersuchung besteht in den Sozialwissenschaften meist nur aus einer verbalen Formulierung von Hypothesen, da die zugrundeliegenden Theorien nur in den seltensten Fällen formalisiert bzw. quantifiziert sind. Ganz allgemein behaupten diese Hypothesen den Einfluß einer oder mehrerer sog. "unabhängiger Variabler" auf eine oder mehrere "abhängige Variablen", also die Beobachtungsdaten, welche mit Hilfe einer Meßvorschrift in die zufälligen Veränderlichen (4.1.1.1) transformiert werden.

In der empirischen Psychologie werden hauptsächlich zwei Arten von Hypothesen formuliert, die Differenz- und Zusammenhangshypothesen. Bei der Differenzhypothese geht man so vor, daß man die Merkmalsträger (in der Regel Versuchspersonen) in Teilstichproben aufteilt, die sich durch den Ausprägungsgrad der unabhängigen Variablen unterscheiden. Die verschiedenen Ausprägungsgrade werden dabei entweder hergestellt - Experiment - oder vorgefunden (aufgesucht) - Erhebung (Befragung). In beiden Fällen, Experiment und Erhebung, wird dann untersucht, ob sich die Teilstichproben inbezug auf die abhängige(n) Variable(n) unterscheiden. Daraus gelangt man zu einer statistischen Differenzhypothese, indem man zu jeder Teilstichprobe eine eigene Grundgesamtheit postuliert und unterstellt, entsprechende Differenzen bestünden auch in diesen Grundgesamtheiten.

Bei der Zusammenhangshypothese erfaßt man auch die unabhängige(n) Variable(n) als zufällige Veränderliche und untersucht dann deren Zusammenhang mit der (den) abhängigen Variablen (oft wird hier nicht mehr zwischen abhängigen und unabhängigen Variablen unterschieden; man untersucht dann nur noch Zusammenhänge zwischen verschiednen Variablen). Die statistische Zusammenhangshypothese behauptet, daß die Zusammenhänge auch in der Grundgesamtheit bestehen würden, aus welcher die Merkmalsträger (Versuchspersonen) entnommen seien.

4.1.4.1 Nullhypothese

Für die Differenzhypothese kann die statistische Nullhypothese sehr allgemein formuliert werden. Sie behauptet dann, alle den Teilstichproben zugeordneten Grundgesamtheiten seien identisch:

$$H_0'' \triangleq GG_1 \Leftrightarrow GG_2 \Leftrightarrow GG_3 \ldots$$

Dies impliziert eine spezifischere Variante, die besagt, daß auch die Verteilungsfunktion einer beliebigen Veränderlichen X in den verschiednen Grundgesamtheiten identisch sei:

$$H_0' \triangleq F_1(X) = F_2(X) = F_3(X) \ldots$$

Wenn aber die Verteilungsfunktionen gleich sind, müssen auch die Verteilungsparameter gleich sein, sodaß eine noch spezifischere Variante der Nullhypothese formuliert werden kann:

$$H_0 \triangleq \Theta_1 = \Theta_2 = \Theta_3 \ldots)^*$$

Für die Zusammenhangshypothese gibt es nur eine Form der Nullhypothese:

$$H_0 \triangleq \rho(X, Y, Z \ldots) = 0,$$

wobei die Prüfgröße ρ irgendein Maß für den Zusammenhang der Variablen X, Y, Z ... ist, das die Bedingung erfüllt, bei fehlendem Zusammenhang den Wert Null anzunehmen (z.B. ein Korrelationskoeffizient).

Wir haben bisher absichtlich nicht den ein- und zweiseitigen Fall bei der Formulierung der H_0 unterschieden. Manche Autoren tun das und setzen im einseitigen Fall $H_0 \triangleq \Theta_1 \leq \Theta_2$ oder $\rho(X, Y, Z..)$

)* Mathematisch äquivalent sind: $H_0 = \Theta_1 - \Theta_2 = 0$ (z.B. t-Test für Differenzen von Mittelwerten) und $H_0 = \Theta_1/\Theta_2 = 1$ (Varianzanalyse)

≤ 0 (HAYS 1977, 369; RENN 1975, 15; SCHAICH 1977, 191). THOLEY (1982) hat aber gezeigt, daß man dadurch in einen schwerwiegenden Widerspruch gerät. Die so formulierten Hypothesen sind nämlich logisch schwächer als die zweiseitigen, gleichwohl (unter sonst gleichen Bedingungen) leichter widerlegbar. Um die gegenwärtige Praxis des Hypothesentests in der psychologischen Forschung zu verteidigen, sollte die H_0 deshalb im ein- und zweiseitigen Fall identisch formuliert werden, wie das z.B. auch CARLSON (1973, 208), sowie CLAUSS & EBNER (1972, 182) tun. BORTZ (1977, 149-150), der ebenfalls so vorgeht, bringt dafür noch andere Argumente als THOLEY (1982) bei. - Der Seitigkeit der Fragestellung kann man bei der Formulierung der Alternativhypothese Rechnung tragen (siehe nächster Abschnitt).

4.1.4.2 Alternativhypothese

Die Alternativhypothese H_1 wird so formuliert, daß sie inbezug auf die Nullhypothese H_0 disjunkt und erschöpfend ist, d.h. es gilt H_0 oder H_1 - nicht beide - und beide Hypothesen umfassen alle möglichen Aussagen über das jeweils ins Auge gefaßte Merkmal der Grundgesamtheit(en).

Für die allgemeinste Variante der H_0'' existiert keine entsprechend allgemeine Variante der Alternativhypothese. Für die spezifischere Variante H_0' lautet die Alternativhypothese:

$$H_1' \cong F_1(X) \neq F_2(X) \neq F_3(X) \ldots$$

Das Testen dieser Hypothese erfolgt mit sog. "tests of fit" (KENDALL & STUART 1967, 419-465) oder "Omnibustets" (LIENERT 1973, 121), da Verteilungsunterschiede Unterschiede zwischen allen möglichen Verteilungsparametern einschließen können. Ist eine der Verteilungsfunktionen F bekannt, oder wird sie als bekannt vorausgesetzt ("theoretische Verteilung"), spricht LIENERT (1973, 64) auch von "Anpassungstests". Ist das nicht der Fall, sind verteilungsfreie Verfahren im engeren Sinn gegeben (distribution free tests, KENDALL & STUART 1967, 450-461). Die Bezeichnungen in der Literatur insgesamt sind jedoch nicht einheitlich.

4.1.4.3 Die Abstimmung der beiden statistischen Hypothesen aufeinander

Die grundlegende Forderung, daß die beiden statistischen Hypothesen gemeinsam disjunkt und erschöpfend sein sollten, hatten wir bereits erwähnt. Im vorangehend dargestellten Fall der einseitigen Hypothesen scheint diese Forderung verletzt zu sein. Denn wenn die statistischen Hypothesen z.B. lauten: $H_0 \triangleq \Theta_1 = \Theta_2$ und $H_1 \triangleq \Theta_1 > \Theta_2$, fehlt ja offensichtlich der Fall $H_2 \triangleq \Theta_1 < \Theta_2$. Ein Vorschlag KAISERs (1960), zweiseitige und trotzdem gerichtete Hypothesetests zu verwenden (wobei dann drei statistische Hypothesen formuliert und getestet werden müssen), konnte sich bisher nicht durchsetzen. Der gegenwärtigen Forschungspraxis wird man daher eher gerecht, wenn man den bei einseitigen Hypothesen nicht inbetracht gezogenen Fall (im obigen Beisp. also H_2) zu den"statistischen Oberhypothesen" oder Eingangsgrößen (siehe Abschn. 4.1.1) schlägt und nicht zu den statistischen Hypothesen. Das deckt sich auch insofern mit der gegenwärtigen Praxis psychologischer Forschung und dem in Lehrbüchern der Statistik vermittelten Wissen, als dort in der Regel erwartet wird, daß man bereits v o r der Formulierung der statistischen Hypothesen und in einem gesonderten Argumentationszusammenhang begründet, warum man bei einseitiger Fragestellung einen bestimmten Bereich von Untersuchungsergebnissen von vornherein außer Betracht läßt. Wenn es dann in einem späteren Schritt an die Formulierung der statistischen Hypothesen geht, beziehen sich diese im einseitigen Fall auf einen schon eingeschränkten Bereich möglicher Resultate der Untersuchung, sind aber für d i e s e n Bereich dann wieder erschöpfend.

Es gilt ferner die Regel, daß nur solche statistische Hypothesen miteinander zu kombinieren sind, welche den gleichen Allgemeinheitsgrad haben. Für die allgemeinste Variante der Nullhypothese (H_0'') fehlt die entsprechende Alternativhypothese, da kein entsprechender Hypothesentest existiert. H_0'' hatten wir trotzdem aufgeführt, da sie auch in vielen Lehrbüchern erwähnt wird und ausserdem mit-falsifiziert werden kann, wenn eine spezifischere Nullhypothese zurückgewiesen wird. Die Umkehrung des letzten Satzteils gilt nicht: Wenn Grundgesamtheiten verschieden sind, können doch einzelne Variablen, die an ihnen beobachtet werden, gleiche Verteilungsfunktionen haben; sind diese verschieden, können doch bestimmte Parameter gleich sein; unterscheiden sich

diese, können immer noch andere Parameter gleich sein - ja, sie m ü s s e n es in bestimmten Hypothesentests sogar. Dieses sog. Homomeritätspostulat besagt, daß nur der in der Alternativhypothese explizit formulierte Unterschied zwischen den Grundgesamtheiten besteht, während alle sonstigen Merkmale der Grundgesamtheiten gleich sind (RENN 1975, 50). In der Varianzanalyse z.B. behauptet die Alternativhypothese Unterschiede in der Varianz der Mittelwerte von Grundgesamtheiten ("Varianz zwischen"); das Homomeritätspostulat fordert in diesem Fall gleiche Varianz der Einzelwerte innerhalb der Grundgesamtheiten ("Varianz innerhalb"), gleiche Verteilungstypen usw.

Das Homomeritätspostulat ist eine sehr starke Forderung, die in der Forschungspraxis aber einfach vorausgesetzt werden muß, mehr oder weniger in jedem Einzelfall plausibel gemacht wird, oder in bestimmten Anwendungsbereichen auch durch eigene Hypothesentets belegt werden kann (z.B. BARTLETT-Test für Varianzhomogenität bei varianzanalytischen Versuchsplänen, vgl. u.a. BORTZ 1977, 345-346). Ein wirkungsvolles Mittel zum Erfüllen dieser Forderung ist auch die Randomisierung, also die Zufallsaufteilung der Merkmalsträger (i. d. R. Versuchspersonen) auf die verschiedenen Versuchsbedingungen, was freilich nur in experimentellen Anordnungen möglich ist (also nicht bei Erhebungen).

4.1.5 Signifikanzniveau

Das Signifikanzniveau α , auch "Umfang" eines Tests genannt, ist die Wahrscheinlichkeit, die Nullhypothese zu verwerfen, obwohl sie zutrifft. Der Wert von α wird in den Sozialwissenschaften aufgrund einer willkürlichen Konvention meist mit 0,05, 0,01 oder 0,001 vorgegeben.

4.2 Abgeleitete Konstrukte

4.2.1 Verteilung der abhängigen Variablen in der Grundgesamtheit

Dieses Systemelement ist das komplexeste im statistischen Test, sowohl als Konstrukt, als auch hinsichtlich der Ableitungsregeln und deren logischer Struktur.

4.2.1.1 Konstrukt

Gemeinsam ist allen Verteilungsannahmen bei den verschiedenen Signifikanztests, daß man von der Nullhypothese ausgeht. Alle weiteren Annahmen sind mehr oder weniger testspezifisch.

Parametrische Tests aufgrund großer Stichproben, welche die Anwendung des zentralen Grenzwertsatzes ermöglichen (siehe 4.2.4), erfordern die schwächste Verteilungsannahme: Es kann irgendein Verteilungstyp sein, er muß nur ein arithmetisches Mittel und eine Varianz größer als Null haben.

Die nächststärkere Verteilungshypothese erfordern parametrische Tests für mittlere und kleine Stichproben, sowie eine Reihe von nicht-parametrischen Tests. So setzt z.B. der t- und der F-Test (Varianzanalyse) normalverteilte Daten in der Grundgesamtheit voraus; sog. nichtparametrische Tests, welche die Binomialverteilung als Prüfverteilung verwenden, beruhen auf einer Zweipunktverteilung des untersuchten Merkmals in der Population.

Ganz ohne Verteilungshypothesen kommen angeblich Randomisierungstests für Rang- und Meßdaten aus (SIEGEL 1976, 31; RENN 1975, 38). Bei der Herleitung des Stichprobenraumes und der Prüfverteilung werden wir jedoch sehen, daß auch hier eine spezielle Verteilungshypothese gegeben ist, nämlich die Annahme einer Gleichverteilung (siehe auch LIENERT 1973, 107).

Bei den Tests für Rangdaten (z.B. U-Test) besagt diese Gleichverteilungshypothese, daß die Grundgesamtheit nur aus Rangzahlen besteht, die auch in der Stichprobe vorkommen und daß jeder dieser Ränge die gleiche Wahrscheinlichkeit hat.

Bei Randomisierungstets für metrische Daten (z.B. FISHER-PITMAN-Test, vgl. LIENERT 1973, 420-428) ist die Gleichverteilungshypothese spezifischer: Sie bezieht sich nur auf die in der Stichprobe realisierten Meßwerte, unterstellt also, daß auch in der Grundgesamtheit nur diese Meßwerte gegeben sind und zwar alle mit der gleichen Wahrscheinlichkeit: "Die Stichprobe ist ein kompositionsgetreues Abbild der Grundgesamtheit" (LIENERT 1973, 417).

4.2.1.2 Ableitungsregeln

Die Annahme einer beliebigen Verteilung, die überhaupt ein arithmetisches Mittel und eine Varianz größer als Null hat, hatten wir bereits als Eingangsgröße behandelt.

Wird ein bestimmter Verteilungstyp verlangt, geht man von der Häufigkeitsverteilung der Stichprobendaten aus und vergleicht sie mit der Dichtefunktion bzw. Wahrscheinlichkeitsverteilung des verlangten Verteilungstyps. Hinreichende Ähnlichkeit begründet dann die Annahme der Verteilungshypothese. "Hinreichende Ähnlichkeit" ist über einen Anpassungstest operationalisierbar (z.B. χ^2-Test für den Vergleich einer empirischen mit einer theoretischen Verteilung). In der Praxis verzichtet man jedoch hierauf häufig, zumal bei kleineren Stichproben ($n < 30$) manche Anpassungstests nicht mehr anwendbar sind. Man weicht dann auf ein mehr intuitives Verfahren aus, indem man die Graphen der Stichprobendaten und der theoretischen Verteilung morphologisch vergleicht. Geht es z.B. um die Normalverteilungshypothese, achtet man darauf, ob sich die Stichprobendaten annähernd symmetrisch und glockenförmig verteilen (vgl. 5.1.1.4).

Eine andere Gruppe von Ableitungsregeln, die insbesondere bei bestimmten diskreten Veränderlichen zugrundegelegt wird, ergeben sich bereits aus den Meßvorschriften der Datenerhebung. Wird dort vereinbart, daß ein Merkmal nur in zwei disjunkten Zuständen erfaßt wird, heißt das gleichzeitig, daß die daraus entstehende Variable in der Grundgesamtheit eine Zweipunktverteilung hat; werden drei Zustände erfaßt, resultiert eine Dreipunktverteilung usw.

Zur Ableitung der für die Randomisierungstests erforderlichen Gleichverteilung der abhängigen Variablen in der Grundgesamtheit kann das "Prinzip vom unzureichenden Grund" herangezogen werden, das auf LAPLACE zurückgeht. MENGES (1972, 32) formuliert es so: "Wenn kein zureichender Grund für die Annahme besteht, eine Modalität werde sich leichter verwirklichen als eine andere, so betrachtet man alle Modalitäten als gleich leicht verwirklichbar - oder als gleichmöglich".

4.2.1.3 Logik der Ableitungsregeln

Die Herleitung eines beliebigen Verteilungstyps, bei dem ein arithmetisches Mittel existiert und die Varianz größer als Null ist, läßt sich deduktiv begründen, sofern in der Stichprobe sich mindestens ein Wert von den anderen unterscheidet und der absolute Betrag der abhängigen Variablen (z.B. aufgrund der als Eingangsgröße festgelegten Meßvorschriften) nicht unendlich groß werden kann. Die eben genannten Bedingungen sind hinreichend, sodaß der Schluß in einer zweiwertigen Logik auf einer Implikation aufbaut.

Die Ableitung eines bestimmten Verteilungstyps mit Hilfe eines Anpassungstests ist logisch komplex, weil dieser alle logischen Schritte noch einmal wiederholt, die in einem Signifikanztest enthalten sind. Das muß nicht auf einen unendlichen Regreß hinauslaufen, da die meisten Anpassungstests schwächere Verteilungshypothesen verwenden als die Tests, deren Voraussetzungen mit den Anpassungstests geprüft werden (RENN 1975, 51).

Die Verwendung morphologischer Merkmale der Häufigkeitsverteilung der Stichprobendaten als Anhaltspunkt für die Verteilung in der Grundgesamtheit ist ein (logisch nicht zwingender) reduktiver Schluß. Die erste Prämisse besagt, daß aus einem bestimmten Verteilungstyp der abhängigen Variablen eine spezielle Form der Häufigkeitsverteilung in der Stichprobe folgt, allerdings nur mit einer gewissen (maximalen) Wahrscheinlichkeit. D.h., es können auch andere Formen der Stichprobenverteilung resultieren, aber mit geringerer Wahrscheinlichkeit. Ist die abhängige Variable in der Grundgesamtheit z.B. normalverteilt, hat eine symmetrische und glockenförmige Häufigkeitsverteilung der Stichprobendaten die größte Wahrscheinlichkeit unter allen möglichen Verteilungsformen. In der zweiten Prämisse kann man nur die Form der Häufigkeitsverteilung in der Stichprobe beschreiben, weil nur dies ein empirischer Sachverhalt ist. Man schließt damit vom Nachsatz auf den Vordersatz, der aber nur eine hinreichende Bedingung beinhaltet. Damit ist der Schluß reduktiv.

Die wenigsten logischen Probleme ergeben sich, wenn (z.B. für Binomialtests) eine Zweipunktverteilung des beobachteten Merkmals in der Grundgesamtheit konstatiert werden muß. Die Logik ist deduktiv und basiert auf den als Eingangsgröße festgelegten Meßvorschriften. Die Festlegung, ein Merkmal sei nur in zwei Zuständen zu erfassen und die Aussage, es handele sich um ein zweipunktverteiltes Merkmal, sind äquivalent.

Das Prinzip vom unzureichenden Grund ist logisch selbst nicht ableitbar, muß also axiomatisch gesetzt werden, um die Gleichverteilung herzuleiten. Inwieweit es wenigstens plausibel ist, ist umstritten (MENGES 1972, 32). Einleuchtend ist es bei den klassischen Beispielen der Wahrscheinlichkeitstheorie (Würfel, Urnen mit Kugeln usw.) - ob auch in sozialwissenschaftlichen Bereichen, ist schon fraglicher. Wird der Satz vom unzureichenden Grund aber einmal als erste Prämisse akzeptiert, folgt der Rest deduktiv, logisch zwingend. In der zweiten Prämisse muß dann behauptet werden, daß im vorliegenden Fall in der Tat kein Grund für die Annahme vorliegt, eine Modalität (ein Meßwert) würde sich unter der Nullhypothese leichter verwirklichen als eine andere. Empirisch wäre diese Feststellung zu begründen z.B. aufgrund früherer Untersuchungen des gleichen Merkmals, aus denen sich möglicherweise keine Gründe gegen die Gleichwahrscheinlichkeitsannahme ergeben haben. Inwieweit das in der Forschungspraxis tatsächlich überprüft wird, muß dahingestellt bleiben; in den Lehrbüchern findet sich jedenfalls keine entsprechende Einschränkung hinsichtlich der Anwendungsvoraussetzungen für Randomisierungstests. Dieser Punkt ist problematisch und muß im Abschnitt "Diskussion" noch einmal eingehender behandelt werden. Zusammenfassend ist vorläufig festzustellen, daß die Ableitungsregeln zwar deduktiv in einer zweiwertigen Logik auf einer Implikation aufgebaut werden können, die Prämissen selbst aber unsicher sind.

4.2.2 Parameter der abhängigen Variabeln in der Grundgesamhtheit

Dieser Schritt ist nicht für alle Signifikanztests erforderlich; er entfällt jedoch nicht - wie man erwarten könnte - generell für die sog. nichtparametrischen Tests. Für den "Ein-Stichproben-Fall" ist nämlich auch in dieser Gruppe von Verfahren die Kenntnis eines Parameters der Grundgesamtheit Voraussetzung, z.B. der Wert des Parameters einer Zweipunktverteilung (Wahrscheinlichkeit der einen Alternative) beim Binomialtest. Nach einer mathematisch präzisen Definition (MENGES 1972, 296) würde dieser Signifikanztest und eine Reihe weiterer Tests, die z.B. von LIENERT (1973) oder SIEGEL (1976) unter den nicht-parametrischen Verfahren geführt werden, als parametrisch bezeichnet werden müssen. RENN (1975, 38) hält es angesichts dieser Probleme in Anlehnung an einige amerikanische Autoren für angemessener, anstelle des Prädikats "nicht-parametrisch" den Begriff "voraussetzungsärmer" zu verwenden, aber selbst

diese Abschwächung reicht nicht aus, was wir bei den Verteilungsannahmen gesehen haben.

4.2.2.1 Konstrukt

"...Parameter... sind Werte, die den Verteilungen nach einer bestimmten Vorschrift als charakteristische Größen zugeordnet werden" (MENGES 1972, 215). Die Parameter erfüllen weiter die Bedingung, daß die Verteilungsfunktion einer zufälligen Veränderlichen von ihnen abhängt, d.h. der Parameterraum umfaßt die Menge der "festen Zahlen", die zu einer Definition der Verteilungsfunktion notwendig sind (MENGES 1972, 268). Für die in der Sozialforschung meist verwendeten Verteilungstypen ist die Zahl der Parameter bekannt, z.B. für die Normalverteilung und die Binomialverteilung zwei, für die POISSON- und die Zweipunktverteilung einer.

4.2.2.2 Ableitungsregeln

Zur Herleitung der Parameter dienen sogenannte "Punktschätzverfahren", die bestimmten Gütekriterien, wie z.B. der Erwartungstreue, Konsistenz usw. genügen müssen.

Ein weitverbreitetes, auf R.A. FISHER zurückgehendes Verfahren ist die Maximum-Likelihood-Methode. Das Prinzip soll in Anlehnung an MENGES (1972, 279f) erläutert werden.

Vor Beginn der Untersuchung kann man die abhängige Variable als zufällige Veränderliche betrachten, deren Wahrscheinlichkeitsverteilung bzw. Dichtefunktion von einem oder mehreren festen, aber unbekannten Parametern der Grundgesamtheit abhängt. Nach der Datenerhebung haben sich bestimmte Werte der zufälligen Veränderlichen realisiert, womit nun diese Stichprobendaten als konstante Größen betrachtet werden. Umgekehrt kann man die unbekannten Parameter der Grundgesamtheit nun als Variablen sehen und sich fragen, welche Werte diese Variablen haben müßten, damit die erhaltenen Stichprobenwerte am wahrscheinlichsten sind. Diese Wahrscheinlichkeit wird als Likelihood bezeichnet. Bezeichnet man mit Θ_1, Θ_2... Θ_m die Parameter in der Grundgesamtheit und mit x_1, x_2... x_n die

in der Stichprobe realisierten Werte der abhängigen Variablen, dann ist die Likelihoodfunktion L definiert als
$L(\Theta_1, \Theta_2 \ldots, \Theta_m \mid x_1, x_2 \ldots x_n) = p(x_1, x_2 \ldots, x_n \mid \Theta_1, \Theta_2 \ldots \Theta_m)$, wenn es sich um eine diskrete Variable handelt. Der Ausdruck rechts vom = - Zeichen ist die Wahrscheinlichkeitsverteilung dieser Variablen. Bei stetigen Veränderlichen erscheint hier eine Dichtefunktion. Soweit die Wahrscheinlichkeitsverteilung bzw. Dichtefunktion ihrem Typ nach durch eine Gleichung (Funktionalform) gegeben ist mit den Parametern als Unbekannten und gegebenen Stichprobenwerten, ist die Likelihoodfunktion eindeutig bestimmt.

Die Likelihoodmethode ist eine bei den Signifikanztesten sehr häufig verwendete Ableitungsregel. Für einige Parameter liefert sie jedoch nicht erwartungstreue, sondern nur asymptotisch erwartungstreue Schätzungen. Deshalb kann dieses Verfahren für einen sehr wichtigen Fall nicht verwendet werden, nämlich für die Schätzung der Populationsvarianz aufgrund der Varianz kleinerer Stichproben. Hier wird ein anderes Schätzverfahren verwendet, das auf dem Konzept des Erwartungswertes E beruht, der wiederum auf dem Stichprobenraum)* definiert ist. Es kann gezeigt werden (z.B. HAYS 1977 , 273-274 und 283-284), daß $E\left(\frac{n}{n-1}\ s^2\right) = \sigma^2$, wobei n = Stichprobenumfang, s^2 = Stichprobenvarianz und σ^2 = Populationsvarianz. D.h., die Stichprobenvarianz muß mit dem Faktor $\frac{n}{n-1}$ multipliziert werden, um als erwartungstreue Schätzung der Populationsvarianz gelten zu können.

4.2.2.3 Logik der Ableitungsregeln

Das Prinzip der Likelihoodfunktion kann logisch als deduktiver Syllogismus auf implikativer Grundlage in einer unendlichwertigen Logik mit der Likelihood als Wahrheitswert rekonstruiert werden. Die maximale Likelihood ist ein ausgezeichneter Wert des Geltungswertbereichs. Formalisiert lautet die Ableitungsregel:

$$\begin{array}{l} p \underset{L_{max}}{\longrightarrow} q \\ p \\ \hline L_{max} \\ q \end{array}$$

p ≙ "In der Stichprobe sind die Werte x_1, x_2, ... x_n der Veränderlichen X realisiert und X hat die Wahrscheinlichkeitsverteilung W bzw. die Dichtefunktion D"

q ≙ "ein bestimmter Parameter Θ_k der Veränderlichen X in der Grundgesamtheit hat den Wert ϑ_k ".

*) Dieses Konstrukt wird im Abschnitt 4.2.3 ausführlich dargestellt, sodaß ich jetzt nicht weiter darauf eingehe.

Um das Schätzverfahren nach dem Prinzip des Erwartungswertes zu begründen (nicht: anzuwenden, s.u.), muß zunächst für den betr. Parametertyp gezeigt werden, daß die Stichprobenstatistik erwartungstreu ist. Dazu setzt man den Parameter in der Grundgesamtheit als Konstante und prüft nach, ob die Stichprobenstatistik, arithmetisch gemittelt über den gesamten Stichprobenraum, den gleichen Wert ergibt. Gegebenenfalls muß man, um das zu erreichen, Korrekturfaktoren zu Hilfe nehmen. Bei der Anwendung des Schätzverfahrens wird dann der Parameter in der Grundgesamtheit als Unbekannte betrachtet und die Stichprobenstatistik als ein realisierter Wert aus dem Stichprobenraum. Die Logik des Verfahrens sei am Beispiel der Varianzschätzung demonstriert. Dabei ist s^2 die Varianz einer Stichprobe vom Umfang n:

P_1 Wenn die Varianz σ^2 einer Grundgesamtheit den Wert a hat, dann ist $E(\frac{n}{n-1} s^2) = a$.

P_2 Wenn $E(\frac{n}{n-1} s^2) = a$, dann ist $\frac{n}{n-1} s^2 = \hat{\sigma}^2$ eine erwartungstreue Schätzung für σ^2.

P_3 Wenn $\hat{\sigma}^2$ eine erwartungstreue Schätzung für σ^2 ist, (dann gilt nach dem Erwartungswertprinzip:) $\hat{\sigma}^2$ kann als Wert für σ^2 gesetzt werden.

P_4 $\hat{\sigma}^2$ ist eine erwartungstreue Schätzung für σ^2. (es folgt nach dem Erwartungswertprinzip:)

$\hat{\sigma}^2$ kann als Wert für σ^2 gesetzt werden.

Die beiden ersten Prämissen sind unproblematisch (zweiwertige Logik). Schwierigkeiten bereitet die Prämisse P_3, weil hier der Nachsatz nicht zwingend aus dem Vordersatz folgt. Also liegt keine zweiwertige Logik vor. Der Nachsatz folgt aber auch nicht mit einer gewissen Wahrscheinlichkeit, sondern eben nur "nach dem Erwartungswertprinzip". Deshalb ist der Syllogismus auch nicht einer mehrwertigen Logik zuzuordnen. Das Erwartungswertprinzip ist logisch allenfalls als "Aussagenfunktor" zu charakterisieren (BOCHEŃSKI & MENNE 1973, 22), aber er ordnet seinen Argumenten (dem Vorder- und Nachsatz) keinen Wert zu, der als Geltungswert interpretiert werden könnte.

Man könnte die Position einnehmen, daß es sich bei den Schätzungen nach dem Erwartungswertprinzip gar nicht um logische Schlußfolgerungen handelt, sondern nur um eine Arbeitsregel im Sinne einer materialen Implikation $a \Rightarrow b$, was heißt: Immer, wenn in einem Kalkül a auftritt, kann es durch b ersetzt werden. Für den

Sonderfall eines mathematischen Kalküls ist a = b. Dabei ist a ⇒ b bzw. a = b völlig willkürlich gesetzt, von jeder inhaltlichen Folgerichtigkeit wird abgesehen. So kann für a z.B. gesetzt werden: "2 mal 2 ist 5" und für b: "Der Schnee ist weiß". Aus a ⇒ b folgt dann, daß der Schnee weiß ist, da die Konklusion aus einer Implikation (auch aus einer materialen) auch dann wahr ist, wenn die erste Prämisse falsch ist (KONDAKOW 1978, 218).

Dem ist entgegenzuhalten, daß das Erwartungswertprinzip durchaus eine inhaltliche Folgerichtigkeit erkennen läßt und insofern von einer strengen Implikation Gebrauch macht, auch wenn deren logische Rekonstruktion nicht ohne weiteres möglich ist. Das ist nicht allzu gravierend, da diese Schätzungen nur in einem Teil der statistischen Tests verwendet werden, also nicht als notwendiger Bestandteil dieses Verfahrens betrachtet werden müssen.

4.2.3 Stichprobenraum

Der Stichprobenraum ist ein generelles (d.h. für alle Signifikanztests geltendes) Verbindungsglied zwischen Grundgesamtheit und Prüfverteilung. Wenn der Stichprobenraum in den Lehrbüchern der Statistik trotzdem bei den meisten Signifikanztests gar nicht erwähnt wird, liegt das daran, daß die Prüfverteilung im praktischen Anwendungsfall nicht immer erneut hergeleitet wird (Ausnahme: einige Randomisierungstests), sondern in standardisierter, tabellierter Form vorliegt. Die mathematische Konstruktion dieser tabellierten Prüfverteilungen kann aber immer über die Zwischenkonstruktion des Stichprobenraumes geführt werden, sodaß er behandelt werden soll, wenn es um die Grundstruktur des Signifikanztests geht.

4.2.3.1 Konstrukt

Der Stichprobenraum ist die Menge aller möglichen Stichproben eines gegebenen Umfangs n aus der Grundgesamtheit. (SCHAICH 1977, 186, FISZ 1966, 276; MENGES 1972, 325; BORTZ 1977, 113; HAYS 1977, 263). Da bereits die Grundgesamtheit unter der Voraussetzung der Nullhypothese stand, gilt diese auch für den Stichprobenraum. Dessen fiktive Stichproben denkt man sich nach dem gleichen Auswahlverfahren der Elemente zustandekommen wie die faktisch vorliegende Stichprobe. Handelte es sich dort z.B. um eine Auswahl "ohne Zurücklegen", d.h. wurden die Elemente der Stichprobe, die zufällig ausgewählt wurden, nicht mehr zur Grundgesamtheit gerechnet, also bei weiteren Ziehungen ausgeschlossen, dann gilt dieses Verfahren auch für die fiktiven Stichproben des Stichprobenraumes. Wurden faktisch mehrere Stichproben herangezogen (z.B. bei den meisten

experimentellen Anordnungen), dann besteht der Stichprobenraum aus ebensolchen Gruppen von Stichproben des entsprechenden Umfangs, die jedoch, da die Nullhypothese vorausgesetzt wird, alle der gleichen Grundgesamtheit angehören.

4.2.3.2 Ableitungsregeln

Gewöhnlich wird als Ableitungsregel angegeben, alle möglichen zufälligen Stichproben eines gegebenen Umfangs n aus einer gegebenen Grundgesamtheit zu ziehen. Das führt bei infiniten Populationen zu infiniten Stichprobenräumen und auch bei finiten Populationen sind die Stichprobenräume so groß, daß sie meist nicht faktisch dargestellt werden können, sondern eine rein fiktive Konstruktion bleiben. Gravierender aber ist noch die mangelnde Generalisierbarkeit der obigen Ableitungsregel. In Randomisierungstests wird ein anderes Verfahren verwendet. Da aber gerade diese Tests den Schlüssel zu einem neuen Verständnis des Signifikanztestes liefern (BREDENKAMP 1972, 30f), ist zu prüfen, ob nicht die hier verwendeten Ableitungsregeln umgekehrt auch für alle anderen Tests gültig sind und zum gleichen Resultat führen.

Um die für die Randomisierungstests gültigen Ableitungsregeln zu generalisieren, faßt man die Verteilung der abhängigen Variablen in der Grundgesamtheit als (endliche) Häufigkeitsverteilung auf, wobei die Häufigkeit der einzelnen Werte ihrer Wahrscheinlichkeit bzw. Dichte in der Grundgesamtheit entspricht. (Stetige Variable müssen für diesen Zweck in Intervalle unterteilt werden). Diese "Modellpopulation" (model population, KURTZ 1965, 106f) ist natürlich nur eine Annäherung an die Wahrscheinlichkeitsverteilung bzw. Dichtefunktion, da Wahrscheinlichkeiten und Dichten meist eine stetige Veränderliche darstellen, Häufigkeiten aber eine diskrete. Läßt man aber die Gesamtzahl der Fälle in der Häufigkeitsverteilung größer werden, ist eine beliebig genaue Angleichung an die Wahrscheinlichkeits- bzw. Dichtefunktion zu erreichen.

Die Herleitung des Stichprobenraumes kann dann mit einem exakten Verfahren vorgenommen werden, indem man systematisch alle Kombinationen von N Elementen zur n ten Klasse ohne Wiederholung bildet (N ist die Zahl der Fälle in der Modellpopulation, n der Stichprobenumfang). Die Wahrscheinlichkeit bzw. Dichte der abhängigen Variablen spiegelt sich insofern angemessen in den Häufigkeiten der einzelnen Stichproben wieder, als Werte mit größerer Wahrscheinlichkeit/Dichte in der Modellpopulation häufiger vertreten sind und deshalb entsprechend häufiger in die Kombinationen (Stichpro-

ben) eingehen. Nach den Regeln der Kombinatorik ist der Umfang des Stichprobenraumes nach diesem Verfahren $\binom{N}{n}$.

4.2.3.3 Logik der Ableitungsregeln

Den verallgemeinerten Ableitungsregeln der Randomisierungstests liegt die Feststellung zugrunde, daß die durch systematische Kombination von N Elementen einer Modellpopulation zur n ten Klasse gewonnenen Stichproben in ihrer Gesamtheit die Häufigkeitsverhältnisse in der Modellpopulation widerspiegeln. Ein Element, das in der Modellpopulation selten auftritt, wird auch in den Kombinationen dieser Elemente entsprechend selten auftreten usw. Dieser Schluß ist logisch zwingend und kann deduktiv in einer zweiwertigen Logik abgebildet werden.

Die Ableitung der Modellpopulation als Häufigkeitsverteilung aus einer Wahrscheinlichkeitsverteilung bzw. Dichtefunktion beruht auf der Häufigkeitsinterpretation der Wahrscheinlichkeit als axiomatischem Bestandteil der Ableitungsregeln. Auf dieser Grundlage ist die Feststellung, eine bestimmte Häufigkeitsverteilung sei eine maximale Annäherung an eine bestimmte Dichtefunktion bzw. Wahrscheinlichkeitsverteilung, ebenfalls logisch zwingend und deduktiv rekonstruierbar.

4.2.4 Prüfverteilung

4.2.4.1 Konstrukt

Berechnet man für jede Stichprobe des Stichprobenraumes die (unter 4.1.2) definierte Prüfgröße, erhält man eine Menge von Werten dieser Größe, die damit ebenfalls als zufällige Veränderliche betrachtet werden kann. Diese Werte können zu einer Häufigkeits- bzw. Wahrscheinlichkeitsverteilung oder Dichtefunktion zusammengestellt werden, der Prüfverteilung (LIENERT 1973, 10, RENN 1975, 21) oder Testverteilung (KREYSZIG 1974 , 156), Stichprobenkennwerteverteilung (BORTZ 1977, 113), Verteilung der Stichprobenfunktion (FISZ 1966 , 277). Die Bezeichnung "Stichprobenverteilung" (BREDENKAMP 1972, 21, aber auch sonst in Lehrbüchern weit verbreitet) - wohl eine Übersetzung des englischen "sampling distribution" (HAYS 1977 , 263f) halte ich für ungünstig, da sie zu leicht mit der "Verteilung der abhängigen Variablen in der Stichprobe" verwechselt werden kann.

4.2.4.2 Ableitungsregeln

Nur bei kleinen, endlichen Grundgesamtheiten, wie sie z.B. bei den Randomisierungstests zugrundegelegt werden, kann die Prüfverteilung entsprechend der obigen operationalen Definition hergeleitet werden: Für jede Stichprobe des Stichprobenraumes (zumindest für die kritischen Regionen) berechnet man die Prüfgröße. Die ermittelten Werte stellt man zu einer Häufigkeitsverteilung zusammen, welche die Prüfverteilung darstellt. Bei grösseren endlichen bzw. bei unendlichen Grundgesamtheiten dient die obige operationale Definition nur noch zur Begriffsklärung. Zur Herleitung der Prüfverteilung versucht man bestimmte Verteilungstypen, die mathematisch definiert und in der Regel auch tabelliert sind, aus den gegebenen Voraussetzungen deduktiv abzuleiten. Handelt es sich z.B. bei der abhängigen Variablen um eine zweipunktverteilte zufällige Veränderliche mit dem Parameter p in der Grundgesamtheit, und es wurde eine Stichprobe mit dem Umfang n ohne Zurücklegen gebildet, resultiert als Prüfverteilung eine Binomialverteilung mit den Parametern n und p. Wurde unter sonst gleichen Bedingungen eine Stichprobe ohne Zurücklegen gebildet, ist die Prüfverteilung eine hypergeometrische Verteilung mit den Parametern n, N und p, wobei N = Umfang der Grundgesamtheit. Eine bestimmte Prüfverteilung ergibt sich mathematisch zwingend aus einer Reihe von Voraussetzungen, die in den vorangehenden Schritten überprüft wurden:

- Art der abhängigen Variablen,
- deren Verteilung in der Grundgesamtheit, ggf. mit entsprechenden Parametern,
- Art der Stichprobenbildung,
- Umfang der Stichprobe,
- Art der Prüfgröße.

Die jeweils hinreichenden Voraussetzungen können noch in zwei wichtige Teilmengen untergliedert werden:

Die eine Menge von Voraussetzungen führt auch für kleine Stichproben zu exakten Prüfverteilungen - Beispiele: Binomial-, hypergeometrische und t-Verteilung.

Für die andere Gruppe von Voraussetzungen sind die Prüfverteilungen Grenzverteilungen, die nur erreicht werden, wenn der Stichprobenumfang gegen unendlich geht. Beispiele sind die Normalverteilung des arithmetischen Mittels als Grenzverteilung beliebiger Verteilungen der abhängigen Variablen (zentraler Grenzwertsatz von

LJAPUNOFF) und die χ^2-Verteilung von normierten Differenzen zwischen beobachteten und theoretischen Häufigkeiten der Werte einer zufälligen Veränderlichen. Bei endlichen Stichproben, wie sie ja in der empirischen Forschungspraxis gegeben sind, ist nur eine mehr oder weniger gute Annäherung an diese Grenzverteilung gegeben, eine Konvergenz. Um eine Konvergenz als hinreichend zu behaupten, bedarf es bestimmter Konventionen über den mindestens erforderlichen Stichprobenumfang, der natürlich für jede Grenzverteilung unterschiedlich ist. In den Sozialwissenschaften wurden diese Mindest-Stichprobenumfänge meist in dem Bereich zwischen 30 und 80 festgelegt.

4.2.4.3 Logik der Ableitungsregeln

Bezeichnet man die Gesamtheit der genannten Voraussetzungen mit "p" und die Aussage: "Eine bestimmte Prüfverteilung f (U) ist gegeben (U sei die jeweilige Prüfgröße)" mit q,
dann ergibt sich der nebenstehende Syllogismus. $\frac{p \rightarrow q \quad p}{q}$
Da p als die Menge der hinreichenden Bedingungen vorausgesetzt wird, ist der Syllogismus tautologisch (logisch zwingend), also deduktiv und hat zwei Geltungswerte (wahr/falsch).

4.2.5 Beobachteter Wert der Prüfgröße

4.2.5.1 Konstrukt

Die unter 4.1.2 als Typ bereits mathematisch definierte Prüfgröße (z.B. als Differenz zwischen arithmetischen Mitteln) nimmt in der bzw. den erhobenen Stichprobe(n) einen bestimmten Wert an. Das ist der beobachtete Wert der Prüfgröße.

4.2.5.2 Ableitungsregeln

Den beobachteten Wert der Prüfgröße erhält man, indem man die Stichprobendaten entsprechend der Definition der Prüfgröße verrechnet.

4.2.5.3 Logik der Ableitungsregeln

Die Logik der Ableitungsregeln ist zweiwertig und deduktiv. Das bedarf keiner weiteren Begründung, da es sich nur um das Ausführen einer Rechnung nach vorgegebenen Rechenanweisungen der elementaren Algebra handelt.

4.2.6 Wahrscheinlichkeit der Prüfgröße, Rejektionsbereich

4.2.6.1 Konstrukt

Gesucht sind die Werte der Prüfgröße, welche (1.) die Alternativhypothese stützen und (2.) zusammengenommen in der Prüfverteilung eine Wahrscheinlichkeit von α haben. Je nachdem, ob es sich um einen ein- oder zweiseitigen Test handelt, liegen diese Werte an einem oder beiden Enden der Prüfverteilung. Der oder die Bereich(e) der Prüfverteilung, für welche die Prüfgröße die eben genannten beiden Bedingungen erfüllt, heißt Rejektionsbereich oder kritische Region, der Rest Annahmehereich.

Bei den sog. "exakten Tests" bildet man den Wertebereich aus dem beobachteten Wert der Prüfgröße und der sich unmittelbar daran anschließenden Werte, welche die Alternativhypothese stützen. Dann berechnet man die Wahrscheinlichkeit dieser Werte in der Prüfverteilung. Bei zweiseitigen Tests muß diese Wahrscheinlichkeit verdoppelt werden.

4.2.6.2 Ableitungsregeln

Bei einseitiger Fragestellung muß zunächst festgelegt werden, ob aufgrund der Alternativhypothese eine Prüfgröße rechts oder links vom Median der Prüfverteilung zu erwarten ist. Liegt er rechts, wird ein Wert von u_α so festgelegt, daß rechts von ihm ein Flächenanteil der Prüfverteilung abgeschnitten wird, das genau α entspricht. u_α wird also so bestimmt, daß für stetige Prüfgrössen die Gleichung
$\int_{u_\alpha}^{\infty} f(U)\, dU = \alpha$ erfüllt ist, für diskrete die Gleichung $\sum_{U \geq u_\alpha} p(U) = \alpha$.
Ist aufgrund der Alternativhypothese eine links vom Median der Prüfverteilung liegende Prüfgröße zu erwarten, wird entsprechend von $-\infty$ bis u_α integriert bzw. die Summe für $U \leq u_\alpha$ gebildet.

Bei zweiseitigen Hypothesen werden zwei Werte von U, nämlich u_{α_1} rechts und u_{α_2} links vom Median der Prüfverteilung so festgelegt, daß an beiden extremen Enden der Prüfverteilungen je ein Flächenanteil von $\alpha/2$ abgeschnitten wird.

Die auf diese Weise gebildeten extremen Teile der Prüfverteilung heißen kritische Region oder Rejektionsbereich, der restliche Teil Annahmebereich.

In der Forschungspraxis bedient man sich meist Tabellen, die für häufig vorkommende Prüfverteilungen in Lehrbüchern der Statistik oder eigenen Tabellenwerken zur Verfügung stehen.

Bei den sogenannten "exakten Tests" verzichtet man auf die Hilfsgröße u_α , sondern berechnet die Wahrscheinlichkeit von u_b (im obigen Sinn) selbst. Für die Herleitung gilt die am Anfang dieses Abschnitts dargestellte Integrations- bzw. Summierungsanweisung, in welcher u_α durch u_b zu ersetzen ist. Dannach kann ebenfalls wieder festgestellt werden, ob die ermittelte Wahrscheinlichkeit $\leq \alpha$ oder $> \alpha$ ist.

4.2.6.3 Logik der Ableitungsregeln

Es handelt sich um eine zweiwertige deduktive Logik, da die Bedingungen, welche der Annahmebereich erfüllt, sich zwingend aus einer gegebenen Prüfverteilung ergeben.

4.3 Ausgangsgrößen

4.3.1 Entscheidung bezüglich der Nullhypothese

4.3.1.1 Konstrukt

Die Prüfverteilung wurde unter der Voraussetzung konstruiert, daß die Nullhypothese gilt. Damit ist auch die im vorangehenden Schritt ermittelte Wahrscheinlichkeit der Prüfgröße eine Wahrscheinlichkeit unter der gewählten Prüfverteilung und der Nullhypothese. Der Grundgedanken des Hypothesentests ist es, die Nullhypothese zu verwerfen, wenn die Wahrscheinlichkeit des beobachteten Wertes der Prüfgröße nur noch sehr klein ist. Was unter "sehr klein" zu verstehen ist, wurde durch das Signifikanzniveau festgelegt. Allgemeiner kann man das so formulieren: In dem Experiment oder der Erhebung hat sich etwas ereignet, das, wäre die Nullhypothese richtig, sehr unwahrscheinlich ist; deshalb lehnt man die Nullhypothese ab.

Im andern Fall, wenn die Wahrscheinlichkeit der Prüfgröße grösser als α ist, wird die Nullhypothese beibehalten - was aber nicht heißt, daß sie verifiziert oder auch nur im Sinne einer Wahrscheinlichkeitslogik gestützt wäre. Es müssen dann beide statistischen Hypothesen als möglich betrachtet werden.

4.3.1.2 Ableitungsregeln

Die Ableitungsregeln sind in den wiederum operationalen Definitionen der vorangehenden Abschnitte bereits enthalten: Ist die Wahrscheinlichkeit des beobachteten Wertes der Prüfgröße $\leq \alpha$, lehnt man die Nullhypothese ab; andernfalls wird sie zusammen mit der Alternativhypothese beibehalten.

4.3.1.3 Logik der Ableitungsregeln

Die Ableitungsregeln können als unendlichwertige Version des modus tollendo tollens betrachtet werden. Hierfür formulieren wir die

1. Prämisse: Wenn die Nullhypothese zutrifft (p),
(gilt mit $(1-\alpha)$:)
der beobachtete Wert der Prüfgröße
fällt in den Annahmebereich (q).

2. Prämisse: Der beobachtete Wert der Prüfgröße
fällt n i c h t in den Annahmebereich ($\neg$q).
(daraus folgt mit $(1-\alpha)$:)

Konklusion: Die Nullhypothese trifft n i c h t zu ($\neg$p).

Stärker formalisiert lautet dieser Syllogismus:

$$\frac{p \overset{1-\alpha}{\Longrightarrow} q \quad \neg q}{\neg p}\ 1-\alpha$$

An dieser Stelle sei auf einen Sachverhalt hingewiesen, der bisher als selbstverständlich vorausgesetzt wurde: Jeder Schritt unseres logischen Kalküls baut auf den vorangehenden Schritten auf. Letztere konstituieren eine zunehmende Menge von Rahmenbedingungen, unter der allein der jeweils folgende Schritt gültig ist. Wollte man dies explizit darstellen, müßten die Prämissen in unseren Syllogismen eine fortschreitende Kette aus Konjunktionen enthalten, für den gegenwärtigen Schritt also lauten: "Wenn die Nullhypothese zutrifft u n d die Grundgesamtheit richtig definiert wurde u n d die Prüfgröße die gewünschten Informationen der Daten repräsentiert u n d die Prüfverteilung F(U) als adäquat betrachtet wird u n d ... ".

Dementsprechend ist auch die einfache Konklusion in obigen Syllogismus folgerichtig nur unter der Bedingung, daß die vorangehenden Schritte als gültig erachtet werden. Andernfalls müßten auch diese Konklusionen eine fortschreitende Kette darstellen, im obigen Fall eine aus nicht-ausschließenden Disjunktionen ("und/oder"). Die Darstellung der logischen Struktur würde damit sehr umständlich werden, weshalb wir den hier beschriebenen Weg gewählt haben.

Wenn der beobachtete Wert der Prüfgröße in den Annahmebereich fällt bzw. (bei "exakten Tests") eine Wahrscheinlichkeit $> \alpha$ haben, also ein "nicht signifikantes Ergebnis" vorliegt, lautet der Syllogismus:

$$\begin{array}{l} p \xrightarrow[1-\alpha]{} q \\ \underline{q} \\ p \end{array}$$

Er ist offensichtlich reduktiv - logisch nicht zwingend - da auf der Grundlage einer Implikation vom Nachsatz auf den Vordersatz geschlossen wird. Eine deduktive Rekonstruktion des Schlusses ist - selbst in einer mehrwertigen Logik - deshalb nicht möglich, weil die Nullhypothese nicht auch (oder nur) als notwendige Bedingung dafür behauptet werden kann, daß der beobachtete Wert der Prüfgröße in den Annahmebereich fällt (das wäre dann eine Äquivalenz oder Replikation). Denn auch, wenn die Nullhypothese nicht richtig ist, kann der beobachtete Wert der Prüfgröße in den Annahmebereich fallen und zwar mit einer Wahrscheinlichkeit von β , die ohne weiteres wesentlich größer sein kann kann als α, in der Regel aber unbekannt ist.

Will man den Hypothesentest in einer deduktiven Logik rekonstruieren, folgt daraus, daß für den Fall, daß der beobachtete Wert der Prüfgröße in den Annahmebereich fällt, ü b e r h a u p t k e i n e l o g i s c h z w i n g e n d e S c h l u ß f o l g e r u n g möglich ist. Ein "nicht signifikantes" Ergebnis ist also - kurz gesagt - gar kein Ergebnis oder jedenfalls ein Ergebnis, das die Entscheidung zwischen Nullhypothese und Alternativhypothese offen läßt. Damit entspricht die in der Psychologie übliche Form des Hypothesentests dem von STEGMÜLLER (1973,) charakterisierten Typ (3). Dieser teilt die Menge möglicher Ausgänge in zwei disjunkte Bereiche, nämlich (a) den Annahmebereich der Alternativhypothese und (b) Urteilsenthaltung.

4.3.2 Entscheidung bezüglich der Alternativhypothese

4.3.2.1 Konstrukt

Wenn die Nullhypothese verworfen wurde, bleibt nur die Alternativhypothese, da beide als disjunkt und erschöpfend vorausgesetzt wurden.

4.3.2.3 Logik der Ableitungsregel

Das Merkmal der statistischen Hypothesen, in der Regel nur in zwei Versionen aufzutreten, die disjunkt und erschöpfend sind, ermöglicht es, sie in eine zweiwertige Logik einzuordnen. Hier gilt der "Satz vom ausgeschlossenen Dritten". Das heißt, daß die Alternativhypothese gelten muß, wenn die Nullhypothese verworfen wurde und umgekehrt.

Die Verwendung einer zweiwertigen Logik für die Beurteilung der Alternativhypothese mag etwas überraschen, weil für die Beurteilung der Nullhypothese eine mehrwertige Logik erforderlich war, andererseits die beiden Hypothesen aber formal eng verwandt erscheinen. In der Tat ist aber der gesamte Signifikanztest, wie wir gesehen haben, auf der Nullhypothese aufgebaut, die Alternativhypothese spielt als konstruktives Element kaum eine Rolle. Die Argumentation zielt daher zunächst ausschließlich auf eine Beurteilung der Nullhypothese ab. Das erklärt ihre Sonderstellung auch im Hinblick auf die logische Struktur ihrer Ableitung.

4.4 Systemelemente mit unterschiedlicher Lokalisation

Es handelt sich hier um verschiedene Transformationsvorschriften, welche die Anwendung von standardisierten und in Tabellenform vorliegender Prüfverteilungen ermöglichen sollen. Eine bekannte Form ist z.B. die Überführung einer normalverteilten Variablen in eine standardnormalverteilte mit dem arithmetischen Mittel Null und der Varianz Eins.

Diese Transformationen sind z.T. bereits in den Meßvorschriften einer Untersuchung festgelegt, also Eingangsgröße; sie können aber auch in einem späteren Stadium eingeführt werden, betreffen aber mindestens den beobachteten Wert der Prüfgröße und die Prüfverteilung.

Auf ihre logische Struktur brauchen wir hier nicht einzugehen, weil sie keine konstituierenden Elemente des Systems sind. D.h., daß sich grundsätzlich jeder Signifikanztest auch ohne diese Transformationen durchführen läßt, nur werden die Berechnungen langwieriger und umfangreicher.

5. Zusammenfassende Schlußfolgerungen

5.1 Der Hypothesentest, für sich allein betrachtet

Wenn man den Begriff "Falsifikation" auch im Rahmen einer mehrwertigen Logik zuläßt, fügt sich der Hypothesentest - für sich allein betrachtet - weitgehend in das Wissenschaftskonzept des kritischen Rationalismus ein. Es handelt sich um ein System von Aussagen, die zum einen Teil deduktiv aus einer Theorie abgeleitet sind, zum andern Teil empirische Sachverhalte beschreiben. Die "Theorie", um die es dabei geht, ist ein stochastisches Modell (Verteilungsannahmen, statistische Oberhypothesen). "Herzstück" ist die Lokalisation der Prüfgröße in der Prüfverteilung. Darüber wird eine Aussage gemacht, die sowohl aus dem stochastischen Modell deduziert ist - inbezug auf dieses also tautologisch ist - als auch empirisch prüfbar ist. Sie lautet: "Der beobachtete Wert der Prüfgröße fällt in den Annahmebereich der Prüfverteilung". Die Negation davon: " ... fällt in den Ablehnungsbereich ..." wäre dann der Basissatz, den die Theorie "verbietet" (POPPER 1966). "Verboten" wird jetzt (im Rahmen einer mehrwertigen Logik) allerdings nicht mehr strikt, sondern mit einem Geltungsanspruch von 95, 99 oder 99,9 Prozent, je nachdem, wie man das Signifikanzniveau ansetzen will.

Damit ist die Voraussetzung für eine Falsifikation im Sinne des kritischen Rationalismus gegeben. Der Hypothesentest ist weder ein induktives Erkenntnisinstrument, noch eine eigenständige Erkenntnismethode "jenseits von POPPER und CARNAP" (STEGMÜLLER 1973, 15-59), deren Logik lediglich auf mathematischer Ebene beschreibbar wäre.

POPPER (1966, Orig. 1934, 157-158) mußte zwar einräumen, daß Wahrscheinlichkeitsansätze nicht im strikten Sinn falsifizierbar sind (verifizierbar schon gar nicht), jedenfalls nicht - das müssen wir heute hinzufügen - im Rahmen einer zweiwertigen Logik. Aber er hat eine Lösung angedeutet, die (obwohl er den Hypothesentest offensichtlich nicht kannte) unserer Rekonstruktion sehr nahe kommt. Er schreibt:

"Die anerkannten Basissätze (können) einem Wahrscheinlichkeits-Ansatz besser oder schlechter entsprechen; sie können einen mehr oder weniger typischen Abschnitt einer Wahrscheinlichkeitsfolge 'realisieren'. An diesem Umstand kann nun die methodologische Regel anknüpfen: Diese könnte ja z.B. verlangen, daß die Basissätze dem Wahrscheinlichkeitsansatz so und so gut entsprechen, d.h., sie könnte eine willkürliche Grenze ziehen und gewisse Abschnitte als erlaubt, andere, etwa stark atypische Abschnitte, als verboten erklären."

An der gleichen Stelle macht POPPER auch Vorschläge, wie die Willkürlichkeit der Grenzziehung zu reduzieren sei. Man könnte sich dabei auf die jeweils "erreichbare Meßgenauigkeit" beziehen. Aus unserer Rekonstruktion ergibt sich eine Möglichkeit, der Willkürlichkeit ganz zu entkommen: Man müßte sich nur entschließen, die statistischen Hypothesen im Rahmen einer mehrwertigen Logik zu beurteilen. Die sehr gekünstelte Dichotomisierung in "signifikante" und "nicht-signifikante" Resultate ergibt sich aus dem Hang zur zweiwertigen Logik - eher eine "Denkmode" als eine Denkmethode.

5.2 Der Hypothesentest im Gesamtrahmen sozialwissenschaftlicher Argumentation

Ordnen wir den Hypothsentest in den Gesamtzusammenhang der sozialwissenschaftlichen Argumentation im Anschluß an eine empirische Forschung ein, müssen wir unterscheiden, ob die zu prüfende inhaltliche (psychologische oder sonst sozialwissenschaftliche) Hypothese mit der statistischen Null- oder Alternativhypothse verknüpft wird. Der zweite Fall ist der häufigste in Wissenschaften mit primitivem Entwicklungsstand der Theorien - also auch in der Psychologie.

5.2.1 Verknüpfung der inhaltlichen Hypothese mit der statistischen Alternativhypothse

Hierbei geht es meist um die Frage, ob bestimmte Beobachtungs- oder Experimentalbedingungen überhaupt einen Einfluß haben auf die abhängige Variable oder nicht; oder daß zwischen zwei oder mehr Variablen ein Zusammenhang besteht oder nicht. Dieser Art von Fragestellungen werden ungerichtete statisti-

sche Alternativhypothesen zugeordnet, die dann einem zweiseitigen Hypothesentest unterworfen werden. Ist die theoretische Entwicklung etwas weiter gediehen, kann zusätzlich noch die Richtung des Einflusses oder des Zusammenhangs angegeben werden: Das führt dann zu gerichteten statistischen Hypothesen und einseitigen Tests. In allen diesen Fällen wird die aus der psychologischen Theorie abgeleitete Hypothese als Zusammenhangs- oder Differenzhypothese formuliert und mit der statistischen Alternativhypothese verbunden (vgl. Abschn. 4.1.4.2): Wenn die Theorie bzw. die daraus abgeleitete Hypothese zutrifft, sind Unterschiede in der Prüfgröße aus verschiedenen Experimentalgruppen bzw. Zusammenhänge zwischen verschiedenen Variablen zu erwarten. Die statistische Nullhypothese ist eine "Gegenhypothese", welche einen Widerspruch zwischen Daten und inhaltlicher Theorie behauptet.

Die erhobenen Daten stützen fast immer zunächst - prima facies - die Alternativhypothese, denn irgend einen Unterschied zwischen zwei oder mehr Versuchsgruppen oder einen - mehr oder weniger starken - Zusammenhang zwischen Variablen wird man fast immer finden. Den Fall, daß die Daten wirklich einmal genau der Nullhypothese entsprechen, brauchen wir deshalb wegen praktischer Irrelevanz nicht weiter verfolgen, obwohl er theoretisch natürlich denkbar ist.

Nun lassen aber Daten, welche zumindest in ihrer Tendenz zunächst die Alternativhypothese stützen, noch alternative Interpreationen zu, die nicht der theoretischen Voraussage entsprechen und daher ausgeschieden werden müssen, wenn das Resultat der empirischen Untersuchung eindeutig interpretierbar sein soll.

Die erste alternative Interpretation erklärt die aufgetretenen Unterschiede bzw. die gefundenen Zusammenhänge mit systematisch wirkenden Störvariablen bzw. mangelhafter Operationalisierung der theoretischen Konstrukte. Diese Interpretation müßte durch eine Diskussion der Versuchsanordnung ausgeschlossen werden und hat mit dem Hypothesentest nichts zu tun.

Die zweite alternative Interpretation beruft sich auf unsystematisch wirkende Störvariablen ("Zufallsschwankungen") und insistiert im übrigen auf der Nullhypothese: "In Wirklichkeit" seien die Differenzen bzw. Zusammenhänge gleich Null; die dennoch beobachteten Werte ungleich Null seien durch Zufallseffekte bei der Stichprobenauswahl, Stichprobenaufteilung bzw. Messung zu erklären. Diese Interpretation kann - falls ein signifikantes Resultat erzielt wird - durch den Hypothesentest zurückgewiesen werden. Übrig bleibt dann die Alternativhypothese, welche die Beobachtungsresultate nun tatsächlich auf die von der inhaltlichen Theorie behaupteten Effekte zurückführt. Das bringt allerdings die wissenschaftliche Erkenntnis nicht weiter, wenn man die Position des kritischen Rationalisten einnimmt: Er interessiert sich ausschließlich für falsifizierende Daten. Man kann noch nicht einmal behaupten, daß hier ein Falsifikationsversuch an der überprüften Theorie gescheitert sei, die Theorie sich also (einmal mehr) bewährt habe. Denn das gewählte Verfahren ist ja zur Falsifikation überhaupt ungeeignet. Es falsifiziert die statistische Theorie bzw. Hypothese, aber nicht das, was man eigentlich falsifizieren will, nämlich die inhaltliche Theorie.

Ein nicht-signifikantes Resultat bringt uns erst recht nicht weiter, denn in diesem Fall ist ja schon auf der Ebene des Hypothesentests Urteilsenthaltung geboten und damit natürlich auch beim Rückschluß auf die Theorie.

Eine Rekonstruktion des wissenschaftlichen Erkenntnisprozesses im Sinne des kritischen Rationalismus ist also bei dieser Anwendungsvariante des Hypothsentests nicht möglich.

5.2.2 Verbindung der inhaltlichen Hypothese mit der statistischen Nullhypothse

Bei fortgeschrittener theoretischer Entwicklung - die freilich in der Psychologie bisher nur selten erreicht wurde - sind auch quantitative Voraussagen möglich. Wenn bestimmte, in einer Theorie formulierte Antezedenzbedingungen erfüllt sind, werden z.B. die durchschnittlich zu erwartenden Zahlenwerte einer geeigneten abhängigen Variablen prognostiziert (vgl. z.B. SOMMER 1966). Man spricht hier kurz von "erwarteten Wer-

ten"* (in bestimmten Fällen auch von "erwarteten Häufigkeiten"), die empirisch mit beobachteten Werten konfrontiert werden können. Auch dabei ist in der Psychologie und in den Sozialwissenschaften überhaupt eine gewisse Fehlervarianz aufgrund unsystematischer Störvariablen zu erwarten, die in der Größenordnung der von der Theorie prognostizierten Effekte liegen kann. Deshalb stimmen, auch wenn die Theorie richtig ist, erwartete und beobachtete Werte in der Forschungspraxis kaum jemals vollständig überein. Damit ergibt sich die Frage, ob diese Differenzen "noch durch Zufall zu erklären sind", oder ob es sich bereits um systematische Abweichungen handelt. Auch hier können Hypothesentests eingesetzt werden (besonders beliebt ist hier der χ^2-Test), allerdings sind sie anders in die Argumentationskette eingebaut als im vorangehenden Fall.

Der "Basissatz", der jetzt im POPPERschen Sinne von der Theorie "verboten" wird, lautet: "Die Abweichungen zwischen den erwarteten und beobachteten Werten gehen über das Ausmaß hinaus, das aufgrund unsystematisch wirkender Störvariablen zu erwarten ist". Dieser Sachverhalt wird aber jetzt in der statistischen Alternativhypothese abgebildet, während die von der inhaltlichen Theorie geforderte Übereinstimmung zwischen beobachteten und erwarteten Werten in der statistischen Nullhypothese konstatiert wird. Diesesmal bedeutet ein "signifikantes" Ergebnis, daß der von der Theorie "verbotene" Sachverhalt eingetreten ist und diese Theorie somit als falsifiziert gelten muß.

Auch der andere Ausgang des Hypothesentests - keine Signifikanz, Beibehalten der Nullhypothese neben der Alternativhypothese - fügt sich in die Wissenschaftskonzeption des kritischan Rationalismus ein: Schon der Hypothesentest verlangt ja in diesem Falle "Urteilsenthaltung" und diese ist auch beim weiteren Rückschluß auf die Theorie geboten. Daten, welche die Theorie bestätigen, besagen nichts für deren Richtigkeit. Immerhin kann man wenigstens noch behaupten: "Die Daten widersprechen nicht der Theorie" und/oder: "Ein Falsifikationsversuch ist gescheitert an der Theorie, sie hat sich (einmal

)* Nicht zu verwechseln mit dem Begriff des "Erwartungswertes" aus der mathematischen Statistik, der sich auf (unendlich große) Stichprobenräume bezieht

mehr) bewährt". Denn bei dieser Anwendungsvariante des Hypothesentests wäre eine Falsifikation ja tatsächlich möglich gewesen.

Ein Streit darüber, ob die inhaltliche Hypothese mit der statistischen Null- oder Alternativhypothese verbunden werden sollte, war in den sechziger Jahren schon einmal von den amerikanischen Psychologen GRANT, BINDER, WILSON & MILLER, EDWARDS sowie WILSON, MILLER & LOWER geführt worden; die Kontroverse ist bei BREDENKAMP 1972, 76 - 80 dargestellt. Er weist nach, daß die Fraktion, die für eine Verknüpfung der inhaltlichen Hypothese mit der Nullhypothese plädiert, sich in Widersprüche verwickelt. Auch die gegenteilige Position, die u.a. von EDWARDS gehalten wird, ist angreifbar, aber eher deshalb, weil hier mit unglücklichen und zudem für die psychologische Forschung nicht gerade typischen Beispielen operiert wurde. BREDENKAMP (1972) zieht jedenfalls aus dieser Kontroverse die Konsequenz, daß "es eher der POPPERschen und HOLZKAMPschen Analyse der Prüfung wissenschaftlicher Hypothesen entsprechen (würde), wenn man die Prognose mit der statistischen Nullhypothese gleichsetzen würde" (79).

Auch MEEHL (1967) plädiert - übrigens ohne Bezug auf die o.g. amerikanischen Autoren - für die Verbindung der inhaltlichen Hypothese mit der Nullhypothese. Er begründet das insbesondere damit, daß eine Verbesserung der Beobachtungsverfahren (Erhöhung der Präzision und Zuverlässigkeit) unter sonst gleichen Bedingungen die Wahrscheinlichkeit erhöht, die Nullhypothese zurückweisen zu können. Das führt aber in dem für die Psychologie üblichen Fall, daß die Alternativhypothese mit der inhaltlichen Hypothse verbunden ist, zu dem paradoxen Ergebnis, daß zunehmende Präzision der Messungen "provide an easier hurdle for the theory to surmont", sodaß dann immer mehr psychologische Hypothesen und Theorien als "bestätigt" betrachtet werden oder zumindest als "nicht falsifiziert". Das ist sehr wahrscheinlich ein Grund für die wachsende Menge von "trivialen ad-hoc-Hypothesen, die so gut wie keinen Erkenntnisgewinn bringen, ... die theoretische Zersplitterung, Fragmentierung und Trivialisierung (MERTENS & FUCHS 1978, 115), die Kritiker der Psychologie zehn Jahre später konstatieren müssen.

Die andere Anwendungsvariante des Hypothesentests mit einer Verbindung der inhaltlichen Hypothese mit der Nullhypothese - die MEEHL in der zitierten Arbeit übrigens schon gar nicht mehr in der Psychologie, sondern in den "physical sciences" ansiedelt - führt dagegen zu dem einleuchtenden Resultat, daß mit wachsender experimenteller Präzision Theorien eher widerlegt werden können.

Die gegenwärtig in der Psychologie vorwiegende Praxis, die inhaltlichen Hypothesen mit der Alternativhypothese des Hypothesentests zu verbinden, läßt sich nur aufrechterhalten, wenn man zusätzlich auch den sog. "Fehler zweiter Art" (β-Fehler) in das Kalkül einbezieht. Diesen Weg hat BREDENKAMP (1972) beschritten und einen "modifizierten Signifikanztest" vorgeschlägen, der sich bis jetzt allerdings nicht allgemein durchsetzen konnte. Deshalb gingen wir hier nicht näher darauf ein, da wir ja nur die gegenwärtig vorherrschende Forschungspraxis betrachten wollten.

Literatur

AHRENS, H.J. (1974²): Spezielle Methoden der Psychologie. In ROGGE, K.-E. (Hrsg.): Steckbrief der Psychologie. Heidelberg: Quelle & Meyer

ANDERSON, N.H. (1961): Zitiert nach BREDENKAMP (1972)

ATKINSON, R.C. & ESTES, W.K. (1963): Stimulus sampling theory. In LUCE et al. (Eds.): Handbook of mathematical Psychology II, 121 - 168

AUSTEDA, F. (ohne Jahr): Wörterbuch der Philosophie. München: Lebendiges Wissen

BAKAN, D. (1966): The test of significance in psychological research. Psychol. Bull. 66, 423-437

BEUTEL, P., H. KÜFFNER & W. SCHUBÖ (1980³): SPSS 8 - Statistik-Programm-System für die Sozialwissenschaften. Stuttgart: G. Fischer

BOCHEŃSKI, J. M. (1965³): Die zeitgenössischen Denkmethoden. Bern: Francke

BOCHEŃSKI, J. M. & MENNE, A. (1973⁴): Grundriß der Logistik. Paderborn: Schöningh (UTB)

BORTZ, J. (1977): Lehrbuch der Statistik für Sozialwissenschaftler. Berlin: Springer

BREDENKAMP, J. (1972): Der Signifikanztest in der psychologischen Forschung. Frankfurt: Akad. Verlagsges.

CARNAP, R. (1952): The continuum of induktive methods. Chicago: Univ. Press

CARNAP, R. (1959, Orig. 1950): Induktive Logik und Wahrscheinlichkeit. Bearbeitet von W. STEGMÜLLER. Springer: Wien

CATTELL, R. B. (1965): The scientific analysis of personality. Harmondworth: Penguin

CLAUSS, G. & H. EBNER (1972): Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Frankfurt: Harry Deutsch

FISHER, R.A. (1935): Zitiert nach SPIELMAN (1974)

FISZ, M. (1966⁴): Wahrscheinlichkeitsrechnung und mathematische Statistik. VEB Verlag Dt. Wissensch.: Berlin

GLASER, W. R. (1979): Statistische Entscheidungsprozeduren über Hypothesen in den Sozialwissenschaften. In H. ALBERT & K. H. STAPF (Hrsg.): Theorie und Erfahrung. Stuttgart: Klett, 117- 138

GROEBEN, N. & H. WESTMEYER (1975): Kriterien psychologischer Forschung. München: Juventa

HACKING, I.(1965): The logic of statistical inference. Cambridge: Univ. Press

HAYS, W.L.(1977²): Statistics for the social sciences. London: Holt, Rinehart & Winston

HEMPEL, C.G. (1977, Orig. 1965): Aspekte wissenschaftlicher Erklärung. Berlin: De Gruyter

HOFSTÄTTER, P. R. & D. WENDT (1974): Quantitative Methoden der Psychologie. Frankfurt: Barth

KALBFLEISCH, J.G. (1979): Die Prüfung statistischer Hypothesen. In MENGES, G. (Hrsg.): Handwörterbuch der mathematischen Wirtschaftswissenschaft Bd. II, 141 - 152. Gabler

KAISER, H.F. (1960): Directional statistical decisions. Psychol. Rev. 67, 160-167

KENDALL, M.G. & A. STUART (1967): The advanced theory of statistics. Vol. 2: Inference and relationship. London: Griffin

KIESLER, D.J. (1977): Experimentelle Untersuchungspläne in der Psychotherapieforschung. In F. PETERMANN & C. SCHMOOK (Hrsg.): Grundlagentexte der klinischen Psychologie. Bern: Huber, 106 - 147.

KLAUS, G. & M. BUHR (1972): Wörterbuch der Philosophie, Bd. II. Reinbek bei Hamburg: Rowohlt

KLEITER, W. (1969): Krise des Signifikanztests in der Psychologie. Jahrb. f. Psychol., Psychother. & med. Anthropol., 144 - 164

KONDAKOW, N.I. (1978): Wörterbuch der Logik. Berlin: Das europäische Buch

KREYSZIG, E. (1974⁴): Statistische Methoden und ihre Anwendungen. Göttingen: Vandenhoek & Ruprecht

KRIZ, J. (1978³): Statistik in den Sozialwissenschaften. Reinbek bei Hamburg: Rowohlt TB

KURTZ, K.H. (1965): Foundations of psychological research. Boston: Allyn & Bacon INC.

LEISER, E. (1978): Einführung in die statistischen Methoden der Erkenntnisgewinnung. Köln: Pahl-Rugenstein

LIENERT, G.A. (1973 & 1978): Verteilungsfreie Methoden in der Biostatistik. Meisenheim: Hain

ŁUKASIEWICZ, J. (1930): Philosophische Bemerkungen zu mehrwertigen Systemen des Aussagenkalküls. Compte-redus de la Societé des Sciences et de Lettres de Varsovie. Warschau: Cl. III 23

ŁUKASIEWICZ, J. (1913): Zitiert nach BOCHEŃSKI & MENNE (1973)

MEEHL, P.E. (1967): Theory-testing in psychology and physics: a methodological paradox. Philosophy of Science, 34, 103-115

MENGES, G. (1972²) Grundriß der Statistik. Teil 1: Theorie. Opladen: Westdeutscher Verl.

MERTENS, W. & G. FUCHS (1978): Krise der Sozialpsychologie? München: Ehrenwirth

NEYMAN, J. & E.S. PEARSON (1933): Zitiert nach SPIELMAN (1974)

OPP, K.-D. (1970): Methodologie der Sozialwissenschaften. Reinbek bei Hamburg: Rowohlt

POPPER, K.R. (1966): Logik der Forschung. Tübingen: Mohr

POST, E.L. (1921): Zitiert nach ŁUKASIEWICZ (1930)

PRIM, R. & H. TILMANN (1973): Grundlagen einer kritisch-rationalen Sozialwissenschaft. Heidelberg: Quelle & Meyer

REICHENBACH, H. (1957⁵): Experience and prediction. Chicago: Univ. Press

RENN, H, (1975): Nichtparametrische Statistik. Stuttgart: Teubner

ROSENTHAL, R. & J. GAITO (1963): The interpretation of levels of significance by psychological researchers. J. of Psychol 55, 33-38

SCHAICH, E.(1977): Schätz- und Testmethoden für Sozialwissenschaftler. München: Vahlen

SEGETH, W. (1973⁸): Elementare Logik. Berlin: VEB Dt. Verl. d. Wissensch.

SIEGEL, S. (1976): Nichtparametrische statistische Methoden. Frankfurt: Fachbuchhandlg. f. Psychol.

SINOWJEW, A.A. (1968): Über mehrwertige Logik: Ein Abriß. Berlin: VEB Dt. Verl. d. Wissensch.

SOMMER, J. (1966): Die Quantifizierung der Perspektive-Theorie der geometrisch-optischen Täuschungen. Ber. 25. Kongr. d. Dt. Ges. f. Psychol. in Münster, 317-323

SPIELMAN, S. (1974): The logic of tests of significance. Philosophy of Science 41, 211 - 226

STEGER, J.A. (1971): Readings in Statistics. New York: Holt, Rinehart & Winston

STEGMÜLLER, W. (1970): Probleme und Resultate der Wissenschaftstheorie und analytischen Philosophie. Band II: Theorie und Erfahrung. 1. Halbband. Berlin: Springer

STEGMÜLLER, W. (1973a): Probleme und Resultate der Wissenschaftstheorie und analytischen Philosophie. Band IV: Personelle und statistische Wahrscheinlichkeit. 2. Halbband: Statistisches Schließen, statistische Begründung, statistische Analyse. Berlin: Springer

STEGMÜLLER, W. (1973b): Probleme und Resultate der Wissenschaftstheorie und analytischen Philosophie. Band IV: Personelle und statistische Wahrscheinlichkeit. 1. Halbband: Personelle Wahrscheinlichkeit und rationale Entscheidung. Berlin: Springer

STELZL, I. (1982): Fehler und Fallen der Statistik. Bern: Huber

STEVENS, S.S. (1968): Zitiert nach BREDENKAMP (1972)

THOLEY, P. (1982): Signifikanztest und BAYESsche Hypothesenprüfung. Arch. f. Psychol. 134, 319-342

VETTER, H. (1967): Wahrscheinlichkeit und logischer Spielraum. Tübingen: Mohr

WELLENREUTHER, M. (1982): Grundkurs: Empirische Forschungsmethoden. Königstein: Athenäum

WOLINS, L. (1976): Secondary analysis of published research in the behavioral sciences. Proc. Am. Stat. Ass., 109-117

WOTTAWA, H. (1977): Psychologische Methodenlehre. München: Juventa

ZAHLEN, J.-P. (1966): Über die Grundlagen der Theorie der parametrischen Hypothesentests. Stat. Hefte 7, 148-174

ZAWIRSKI, Z. (1935): Über das Verhältnis der mehrwertigen Logik zur Wahrscheinlichkeitsrechnung. Studia Philosophica (Krakau) 1, 407 - 442

ZECHA, G. & H. LUKESCH (1982): Die Methodologie der Aktionsforschung. Analyse, Kritik, Konsequenzen. In PATRY, J.-L (Hrsg.): Feldforschung. Bern: Huber, 365-387