

2. Kapitel

Planung und Auswertung von Experimenten*)

Willi Hager und Rainer Westermann

Vorbemerkungen

Ein wesentliches Ziel jeder empirischen Wissenschaft besteht darin, zu fundierten Kausalaussagen zu gelangen, also zu Aussagen über die Ursachen oder Bedingungen der jeweils interessierenden beobachtbaren Sachverhalte. Die bevorzugte Methode zur Überprüfung entsprechender Vermutungen oder *Kausalhypothesen* ist das Experiment, auch in der Psychologie und in benachbarten Sozial- und Verhaltenswissenschaften (Abschn. 1.2). Das Ausmaß, in dem ein bestimmtes Experiment zur Prüfung einer solchen Kausalhypothese geeignet ist, werden wir als dessen *Validität* bezeichnen (Abschn. 1.3).

Soll im Experiment eine Kausalhypothese geprüft werden, müssen die in ihr vorkommenden (theoretischen) Begriffe in beobachtbare Variablen „übersetzt“ werden. Fehler bei dieser „Operationalisierung“ beeinträchtigen die „*Variablenvalidität*“ der Untersuchung (Teil 2). Ob mit Hilfe eines Experiments überhaupt Aussagen über Ursachen möglich sind, hängt von seiner „*internen Validität*“ ab (Teil 3). Inwieweit ein Experiment eine Prüfung für die betrachtete Kausalhypothese ist, wird auch dadurch beeinflusst, mit welchen Personen und in welcher Situation es durchgeführt wird. Diese in ihrer Bedeutung für die wissenschaftliche Theorienbildung häufig unterschätzte „*Populations- und Situationsvalidität*“ wird im Teil 4 besprochen. Die bisher erwähnten Aspekte der experimentellen Validität können sich in verschiedener Weise gegenseitig beeinflussen, und zwar je nach Art der geprüften Hypothese fördernd oder hemmend (Teil 5). Im Teil 6 zeigen wir, daß über die Gültigkeit einer wissenschaftlichen Hypothese entschieden werden kann, indem über die

*) Wir danken den Kollegen Marcus Hasselhorn (Heidelberg), Falk Leichsenring (Göttingen) und Werner Wippich (Trier) für ihre kritischen Anmerkungen zu einigen früheren Fassungen dieser Arbeit und Frau Gabriele Reimann für das Anfertigen der Zeichnungen.

Gültigkeit von aus ihr abgeleiteten *statistischen Hypothesen* entschieden wird. Diese Entscheidung erfolgt mit Hilfe von *Signifikanztests* (Teil 7). Die wichtigsten Fehler, die dabei gemacht werden können und die dann die „*statistische Validität*“ des Experiments herabsetzen, besprechen wir im Teil 8. Dabei werden sich wesentliche Hinweise für die Auswahl derjenigen Versuchspläne und Auswertungsmethoden ergeben, die für ein bestimmtes Experiment wahrscheinlich am besten geeignet sind. Auf zwei Aspekte der statistischen Validität gehen wir in den Teilen 9 und 10 besonders ein: auf Maße für die Größe des „experimentellen Effekts“ und auf die begründete Wahl des Stichprobenumfangs. Eine auf diesen Überlegungen basierende Planungs- und Entscheidungsstrategie stellen wir im Teil 11 dar.

Obwohl wir detailliert auf bestimmte Fragen und Probleme der statistischen Auswertung eingehen müssen, setzen wir die Kenntnis der gebräuchlichen Testverfahren voraus bzw. geben lediglich an, wo man sie sich verschaffen kann.

Im Laufe unserer Erörterungen werden wir immer wieder auf die verbreiteten sog. „Standard-Versuchspläne“ stoßen. Einige ihrer Vor- und Nachteile werden wir jeweils kurz vor dem Hintergrund der verschiedenen Aspekte der experimentellen Validität - dem zentralen Anliegen unseres Artikels - diskutieren; darüber hinausgehende Hinweise zur Anwendung und Auswertung dieser Pläne können dann jeweils den detaillierten Literaturangaben entnommen werden. Überhaupt sollen die zahlreichen Verweise den Leser anregen, sich mit einigen Originalarbeiten zu ihn interessierenden Problemen zu beschäftigen. Er wird dann feststellen, daß viele der hier angesprochenen Probleme noch kontrovers diskutiert werden, so daß man in zahlreichen Einzelfällen sicher auch andere Meinungen und Empfehlungen vertreten kann als die beiden Autoren dieser Arbeit. Insbesondere kann man zu anderen Ergebnissen kommen, wenn man einen anderen wissenschaftstheoretischen Ausgangspunkt wählt.

Insgesamt kommt es uns vornehmlich darauf an, dem „Produzenten“ experimenteller Ergebnisse konkrete Hinweise zu vermitteln, wie er sein Experiment so planen, durchführen und auswerten kann, daß die Resultate zur Entscheidung über die Gültigkeit seiner wissenschaftlichen Hypothese beitragen können.

Der „Konsument“ experimenteller Befunde soll angeregt werden, sich kritisch mit Ansatz, Durchführung und Auswertung ihn interessierender Untersuchungen zu befassen.

1. Einleitung

1.1 Einige Begriffsbestimmungen

Faßt man die wesentlichen Gemeinsamkeiten der zahlreichen Definitionen des Experiments (siehe z.B. Bredenkamp, 1969a; Zimmermann, 1972; Fietkau, 1973) zusammen, gelangt man etwa zu folgenden Kennzeichen:

- (1) Im Experiment werden bestimmte Bedingungen willkürlich (absichtlich) hergestellt.
- (2) Die hergestellten Bedingungen werden in bestimmter Weise (systematisch) variiert.
- (3) Durch den Vergleich verschiedener Beobachtungen wird der Einfluß dieser Variation auf bestimmte andere Merkmale festgestellt.
- (4) Alle anderen Bedingungen versucht man konstant zu halten oder zu kontrollieren.

Die vom Experimentator (abgekürzt: E) variierten Bedingungen bezeichnet man als die *unabhängigen Variablen* (UVn) oder *Faktoren* des Experiments. Im einfachsten Fall liegt eine unabhängige Variable (UV) mit zwei Ausprägungen vor. Diese Ausprägungen werden auch *Modalitäten*, Stufen oder (Behandlungs- bzw. Treatment-) Bedingungen genannt, oft auch nur kurz Treatments. Als Beispiel denke man an die UV „Art der Verstärkung“ (symbolisiert als A) mit den Modalitäten „Belohnung“ (abgekürzt als A_1) und „Bestrafung“ (A_2); allgemein bezeichnen wir die Modalitäten mit A_j , wobei gilt: $j = 1, 2, \dots, l, l', \dots > J$.

In der Psychologie sind die Modalitäten der UV i.a. wie im angeführten Beispiel qualitativ gestuft; nur in Ausnahmefällen besteht die Möglichkeit, eine UV mit quantitativen Abstufungen einzuführen (als Beispiel denke man an die Darbietungszeiten in einem Lernexperiment).

Werden unabhängig voneinander mehrere Merkmale variiert, spricht man von einem „*mehrfaktoriellen* (oder multidimensionalen) Experiment“. Als Beispiel für eine zweite UV sei die „Häufigkeit der Verstärkung“ (symbolisiert als B) genannt, und zwar mit den Ausprägungen „Nach jeder Reaktion“ (B_1) und „Nach jeder 3. Reaktion“ (B_a). Allgemein benennen wir die Modalitäten des Faktors B mit B_k , wobei gilt: $k = 1, 2, \dots, m, m', \dots, K$. In mehrfaktoriellen Experimenten wird i.a. jede Bedingung jeder UV mit jeder Bedingung jeder anderen UV kombiniert („gekreuzt“); vgl. jedoch Abschnitt 8.3.5.

Die Sachverhalte, die als Auswirkungen der vorgenommenen Bedingungsvariation betrachtet werden, stellen die *abhängigen Variablen* (AVn) des Experiments dar, z.B. die „Anzahl der verbalen Aggressionen gegen den Versuchsleiter“ (Y_u) oder die „Anzahl der Durchgänge bis zum Erreichen des Lernkri-

teriums“ (Y_v). Untersuchungen, in denen die Werte auf nur einer AV erhoben werden, heißen *univariat*, solche mit mehreren *multivariat*.

Diese Begriffsbestimmungen sind nur als erste Orientierung gedacht, Präzisierungen ergeben sich aus dem weiteren Inhalt dieser Arbeit. Insbesondere werden wir im Abschnitt 3.6 zu einer genaueren Definition des Experiments gelangen, die auch eine Abgrenzung gegenüber anderen Forschungsmethoden erlaubt.

1.2 Das Experiment als Methode zur Prüfung von Kausalaussagen

Um beschreiben zu können, wann Planung und Auswertung eines Experiments adäquat sind und welche Fehler dabei gemacht werden können, wollen wir von Ziel und Zweck des Experiments ausgehen: In der Fachliteratur wird das Experiment als die einzige Methode bezeichnet, die generell zur Überprüfung von Kausalhypothesen geeignet ist (Siebel, 1965; Aronson & Carlsmith, 1968; Miller, 1970; Gadenne, 1976; Cook & Campbell, 1979). Kausalhypothesen und alle anderen Arten von Vermutungen, vorläufigen Problemlösungen, theoretischen Ableitungen, Vorhersagen usw., die empirischen Untersuchungen vorgeordnet sind, werden wir zusammenfassend „wissenschaftliche Hypothesen“ (WH) nennen. Zur Verdeutlichung wollen wir von einer fiktiven Problemstellung ausgehen, die uns als Beispiel auch bei unseren weiteren Erörterungen zuweilen begegnen wird:

Da man befürchtet, daß durch die mangelnde Integration von Ausländern (Gastarbeitern) schwere soziale Konflikte auftreten, sei die politische Entscheidung gefallen, die Benachteiligung von Ausländern zu beenden und ihre Eingliederung in die Gesellschaft zu fördern. Da dieser Beschluß nicht einhellig von der Bevölkerung unterstützt wird, soll als einer der ersten Schritte ein Programm zur Veränderung der negativen Einstellung gegenüber Ausländern durchgeführt werden.

Wir wollen uns vorstellen, ein Team von Psychologen und Sozialwissenschaftlern habe den Auftrag bekommen, dieses Programm zu planen, durchzuführen und seine Wirksamkeit zu überprüfen.

In der Fachliteratur ist eine große Zahl von theoretischen Vorstellungen darüber zu finden, durch welche Maßnahmen Einstellungen verändert werden können. Eine der bekanntesten Theorien in diesem Zusammenhang ist die kognitive Dissonanztheorie von Festinger (1978; vgl. Irle, 1975; Frey, 1978).

Aus der Theorie der kognitiven Dissonanz läßt sich unter anderem die folgende Hypothese ableiten: „Wenn zwischen der Einstellung einer Person zu einem bestimmten Objekt und einem anderen kognitiven Element eine Dissonanz besteht, dann verändert sich die Einstellung so, daß diese Dissonanz

vermindert wird.“ Hier wird die Erzeugung von Dissonanz (D) als hinreichende Bedingung für die Änderung der Einstellung bezeichnet. Wir wollen diese spezielle wissenschaftliche Hypothese als WH, bezeichnen und durch „ $D \mapsto dE$ “ symbolisieren. Dabei soll dE eine Veränderung der betrachteten Einstellung bezeichnen. Falls die Hypothese gültig ist, falls sie also eine psychologische Gesetzmäßigkeit ausdrückt, gilt folgendes: Treten die Ereignisse „D“ und „ dE “ ein, kann „D“ als die *Ursache* von „ dE “ bezeichnet werden (Gadenne, 1976, 29). Von daher können wir unsere obige Vermutung als Kausalhypothese bezeichnen.¹⁾

Trotz einer oberflächlichen Ähnlichkeit wäre es inadäquat, wissenschaftliche Kausalhypothesen mit der materiellen Implikation („ $P \rightarrow Q$ “) der Logik gleichzusetzen. Man würde dann zu dem paradoxen Ergebnis kommen, daß eine Kausalhypothese notwendigerweise wahr sein muß, wenn die Prämisse (P) ganz sicher falsch ist und/oder wenn die Konklusion (Q) ganz sicher wahr ist. Etwas konkreter ausgedrückt hätte das u.a. zur Folge, daß man bei Nichtvorliegen von P keine Erwartungen hinsichtlich des Eintretens oder Nicht-Eintretens von Q ableiten kann. Dagegen ist in wissenschaftlichen Kausalhypothesen (implizit) meist auch die Aussage enthalten, daß die Folge Q nicht eintritt, wenn P nicht gegeben ist *und* wenn alle anderen Bedingungen gleichbleiben. Deshalb umfaßt die Prüfung von Kausalhypothesen immer den Vergleich von Beobachtungen unter mindestens zwei Bedingungen, die sich möglichst nur dahingehend unterscheiden, daß in einer P gegeben ist, in der anderen jedoch nicht. Wie dieses Ziel zu erreichen ist, wird das Hauptthema der folgenden Teile (insbesondere Teil 3) sein. (Näheres zum Begriff der Kausalität findet man u.a. bei Suppes, 1970; Brand, 1976; Cook & Campbell, 1979.)

Um Mißverständnisse zu vermeiden, sei betont: Eine solche Kausalhypothese bedeutet weder, daß - um im Beispiel zu bleiben - „Einstellungsänderung“ die einzige Folge des „Auftretens von Dissonanz“ sein muß, noch daß „Einstellungsänderungen“ ausschließlich durch Dissonanz hervorgerufen werden können. Aus dem meist hochkomplexen empirischen Zusammenhangsgefüge isoliert eine Kausalhypothese also in der Regel nur einen Teilaspekt.

Wir wollen uns in dieser Arbeit mit der Frage beschäftigen, wie man zu empirisch fundierten Aussagen über die „Wahrheit“ oder „Falschheit“ derartiger psychologischer Kausalhypothesen gelangen kann. Zuvor soll aber nicht versäumt werden, darauf hinzuweisen, daß Kausalaussagen der Form

¹⁾ Wir müssen uns in dieser Arbeit auf vereinfachende Andeutungen zur Struktur von Theorien und zur Ableitung von Hypothesen beschränken und verweisen im einzelnen auf Bunge (1967a, b), Stegmüller (1973c, 1974b, 1978, 1979a, b, 1980), Groeben & Westmeyer (1975), Suppe (1977a, b), Henning & Muthig (1979, 13-18) sowie ferner Abschnitt 2.1. Zu den Anforderungen, die eine Aussage erfüllen muß, um als wissenschaftliche Hypothese gelten zu können, siehe Bunge (1967a, 229, 280-290).

„Wenn . . . , dann . . . “ nicht die einzige Art von Hypothesen sind, die in der Psychologie interessieren. Betrachten wir als Beispiel die berüchtigte Aussage „Weiße sind (im Durchschnitt) intelligenter als Neger“. Hier wird etwas ausgesagt über die Unterschiede zwischen statistischen Parametern (Mittelwerten) zweier Populationen hinsichtlich einer bestimmten Variablen. Allgemein ausgedrückt sind solche Hypothesen Aussagen über den statistischen Zusammenhang mindestens zweier Variablen. Auf die Frage der Verursachung gehen sie nicht ein. Zur Unterscheidung von den Kausalhypothesen werden sie als *statistische Populationshypothesen* bezeichnet (Bredenkamp, 1979; Hager & Westermann, im Druck, a). Mit ihnen und anderen möglichen Arten von Hypothesen (s. Bunge, 1967a) werden wir uns gemäß unserer Hauptfragestellung nur am Rande beschäftigen.

Es gibt aber eine große Zahl möglicher *Störfaktoren*, die dazu führen können, daß eine konkrete Untersuchung oder eine bestimmte Art von Untersuchungen *nicht* als bestmögliche Prüfung einer kausalen psychologischen Hypothese bezeichnet werden kann. *Liegen in einem bestimmten Experiment derartige Störfaktoren vor, wollen wir davon sprechen, daß die Validität dieses Experiments zur Prüfung der interessierenden Kausalhypothese herabgesetzt ist.* Hieraus ergibt sich die Forderung, daß Experimente grundsätzlich so zu planen, durchzuführen und auszuwerten sind, daß ihre Validität möglichst hoch ist. Dieser Artikel befaßt sich mit den wesentlichsten Aspekten, die zu beachten sind, will man dieser Forderung annähernd nachkommen.

1.3 Die Validität eines Experiments

Wir werden im folgenden die angesprochenen Störfaktoren in vier Gruppen einteilen und dementsprechend vier Aspekte der experimentellen Validität behandeln:

- (1) Variablenvalidität (Teil 2)
- (2) interne Validität (Teil 3)
- (3) Situations- und Populationsvalidität (Teil 4)
- (4) statistische Validität (Teil 6)

Diese Einteilung entspricht der von Cook & Campbell (1976, 1979), die ihrerseits eine Erweiterung der Unterscheidung zwischen interner und externer Validität nach Campbell (1957, 1969) und Campbell & Stanley (1963) ist. Die folgenden Ausführungen weichen allerdings insofern grundsätzlich von diesen Ansätzen ab, als in ihnen ein Einwand berücksichtigt wird, den Gadenne (1976) gegenüber Campbell & Stanley (1963) geltend macht und der im wesentlichen auch auf die neueren Arbeiten von Cook & Campbell zutrifft: Gadenne (1976) weist nach, daß das Konzept der internen und der externen

Validität einen *induktivistischen Ansatz* darstellt, d.h. Schlüsse vom Besonderen auf das Allgemeine enthält. Da die logische Rechtfertigung induktiver Schlüsse (immer) noch aussteht, empfiehlt Gadenne (1976), das Problem der möglichen Störfaktoren im Experiment im Rahmen der Falsifikationstheorie Poppers (1976) zu behandeln, nach der zur Überprüfung von Hypothesen und Theorien ausschließlich deduktive Schlüsse anzuwenden sind.

Ohne an dieser Stelle detailliert auf wissenschaftstheoretische Probleme eingehen zu können, wollen wir im Anschluß an Gadenne (1976) in vereinfachter Weise schildern, welche Konsequenzen die Falsifikationstheorie für die Überprüfung von Kausalhypothesen wie etwa unserer WH, hat (vgl. Bredenkamp, 1980).

Das Ziel jeder empirischen Wissenschaft kann man in der Aufstellung von wahren Aussagen (Theorien, Hypothesen) mit möglichst hohem Informationsgehalt über die Realität sehen. Nach Popper (1976, 1979) kann dieses Ziel nur erreicht werden, indem man - erstens - nach Fakten sucht, die der betrachteten Theorie oder Hypothese widersprechen und - zweitens - „bessere“ Theorien aufzustellen versucht, die mit mehr empirischen Daten in Einklang stehen. Die als erstes angesprochene Überprüfung einer Theorie bzw. Hypothese erfolgt dadurch, daß man aus ihr Vorhersagen über empirisch beobachtbare Ereignisse ableitet. Diese Vorhersagen haben die Form „Wenn die Bedingung b (die sog. „Anfangsbedingung“) gegeben ist, dann tritt Ereignis e ein“. In einer experimentellen Untersuchung bezieht sich die Anfangsbedingung auf die Modalitäten der UV X, und das Ereignis e entspricht meist bestimmten Unterschieden dY auf der abhängigen Variablen. Beobachtet man nun X, ohne daß dY eintritt, ist man berechtigt, die Hypothese als „falsifiziert“ anzusehen, allerdings nur wenn der entsprechende *Basissatz* „ $X \wedge \neg dY$ “ („X hat vorgelegen, aber keine Veränderung auf Y ist eingetreten“) als reproduzierbare Tatsache akzeptiert werden kann. Praktisch bedeutet dies, daß man sich zur Falsifikation einer Hypothese nie allein aufgrund eines einzigen empirischen Ergebnisses entschließt und daß zur Prüfung einer Hypothese stets mehrere empirische Beobachtungen (konkreter: mehrere Experimente) notwendig sind. Glass (1976, 1978) und Pillemer & Light (1980) geben einen Überblick über Möglichkeiten, die Ergebnisse verschiedener Untersuchungen systematisch zusammenzufassen (s.a. Fricke, 1977; Rosenthal & Rubin, 1979; Vatz et al., 1980; Rosenthal, 1980; Cooper & Rosenthal, 1980; Bredenkamp, 1980; 35-37 und Abschn. 11.2). Zur Falsifikation wird man sich insbesondere dann entschließen, wenn eine neue Theorie aufgestellt werden kann, die auch die der alten Theorie widersprechenden Ergebnisse mit einbeziehen kann (vgl. die Beschreibung wissenschaftlicher Forschungsprogramme als Theorienketten von Lakatos (1974)). Dadurch ergibt sich ein Erkenntnisfortschritt. Solange man sich nicht für die Falsifikation einer Hypothese entschieden hat, gilt sie als „vorläufig bewährt“.

Nun sollte man von einer Bewährung der Hypothese aufgrund vorliegender Daten nur sprechen, wenn man tatsächlich versucht hat, die Hypothese auf eine „ernstzunehmende“ Weise zu widerlegen. „Ernstzunehmen“ ist ein Falsifikationsversuch, wenn er so angelegt wird, daß im Falle der Falschheit der Hypothese auch Daten zu erwarten sind, die ihr widersprechen. Es sollen also *ungerechtfertigte Bewährungen* der Hypothese vermieden werden. Ist diese Forderung erfüllt, spricht Popper (1976) von einem „strengen“ *Prüfversuch*. Allerdings besteht nach dieser Definition die Möglichkeit, eine strenge Prüfung dadurch zu erreichen, daß man die Untersuchung so plant, daß ausschließlich hypothesenkonträre Ergebnisse eintreten können. Dies wäre aber offensichtlich ein ungeeigneter Weg, Erkenntnisse über die Realität zu gewinnen. Deshalb soll der Verweis auf das Kriterium eines strengen Prüfversuchs immer auch folgendes bedeuten: Wenn statt der zu prüfenden Kausalhypothese WH ihr logisches Gegenteil falsch ist, soll die Wahrscheinlichkeit eines der WH widersprechenden empirischen Ergebnisses gering sein. Eine strenge Prüfung soll u.E. also sowohl fälschliche Bewährungen der WH als auch *fälschliche Falsifikationen* vermeiden.

Da diese Forderung nach strengen Prüfungen von Hypothesen und Theorien grundlegend ist für die weiteren Erörterungen in dieser Arbeit, wollen wir jetzt genauer überlegen, unter welchen Umständen eine Untersuchung zur Prüfung einer Hypothese als mehr oder weniger streng bezeichnet werden kann. Wir erinnern uns: Aus den (u.U. relativ komplexen) Anfangsbedingungen X wird mit Hilfe der zu überprüfenden WH das Ereignis dY prognostiziert. Nehmen wir nun an, es gäbe eine andere gut bewährte Hypothese HS, die aus den vorliegenden Anfangsbedingungen X (oder Teilaspekten davon) das gleiche Ereignis dY prognostiziert. Eine solche Hypothese soll als *Störungshypothese* bezeichnet werden. Auch wenn die WH falsch sein sollte, wäre wegen HS in diesem Fall nicht das der WH widersprechende Ereignis $\neg dY$ zu erwarten. Deshalb kann eine Untersuchung nicht als strenger Prüfversuch bezeichnet werden, wenn in ihr die Anfangsbedingungen einer bewährten Störungshypothese vorliegen. Man kann demnach allgemein formulieren:

Die Prüfung einer wissenschaftlichen Hypothese durch ein Experiment ist streng, wenn das aus dieser Hypothese vorhergesagte Ereignis nicht auch mit Hilfe anderer Theorien oder Hypothesen (sog. Störungshypothesen) aus Bedingungen, die im Falle dieses Experiments vorliegen, abgeleitet werden kann.

Diejenigen Merkmale des Experiments, aus denen mit Hilfe einer Störungshypothese die gleiche Prognose abgeleitet werden kann wie aus der zu prüfenden wissenschaftlichen Hypothese, werden als *Störfaktoren* oder *Störbedingungen* der experimentellen Validität bezeichnet. *Liegt in einer Untersuchung eine potentielle Störbedingung nicht vor, heißt diese Bedingung kontrolliert.*

Eine Präzisierung des Begriffs der Strenge einer Prüfung, die es erlauben würde, jeder Untersuchung einen Zahlenwert zuzuordnen, der ein Maß dafür

ist, wie streng diese Untersuchung als Prüfung einer bestimmten Hypothese ist, liegt noch nicht vor. Wir können vielmehr nur folgendes sagen:

Die Prüfung einer wissenschaftlichen Hypothese durch eine Untersuchung U_1 ist strenger als durch eine Untersuchung U_2 , wenn alle potentiellen Störfaktoren, die in U_1 kontrolliert sind, auch in U_2 kontrolliert sind, und wenn in U_1 zusätzlich mindestens ein weiterer Störfaktor kontrolliert ist (Gadenne, 1976, 64).

über den Begriff der Strenge eines Prüfversuchs wollen wir jetzt den Begriff der Validität eines Experiments (bzw. allgemeiner einer Untersuchung) definieren (vgl. Hager & Westermann, im Druck, a):

Die Validität einer Untersuchung zur Überprüfung einer wissenschaftlichen Hypothese ist um so größer, je größer die Wahrscheinlichkeit ist, daß die Untersuchung Daten erbringt, die der Hypothese widersprechen, falls diese tatsächlich falsch ist, bzw. je größer die Wahrscheinlichkeit ist, daß die Untersuchung Daten erbringt, die der Hypothese nicht widersprechen, falls das logische Gegenteil dieser Hypothese falsch ist.

Wir werden in den folgenden Abschnitten ausführlich besprechen, welche konkreten Maßnahmen ein Experimentator zu treffen hat, um die so definierte Validität seines Experiments möglichst groß werden zu lassen (vgl. dazu auch die nicht vorwiegend statistisch orientierten Ausführungen zur Planung des Experiments von Campbell & Stanley, 1963; Cochran, 1968a; Bredenkamp, 1969a, 1980, 1-40; Armitage & Remington, 1970; Stanley, 1973; Cook & Campbell, 1979; Henning & Muthig, 1979). Ganz allgemein laufen diese Maßnahmen darauf hinaus, möglichst viele der potentiellen Störungshypothesen auszuschließen, indem man die Untersuchung so plant, durchführt und gestaltet, daß die für die Anwendung dieser Störungshypothesen notwendigen Anfangsbedingungen nicht gegeben sind, daß also die entsprechenden Störfaktoren kontrolliert sind. Die dabei im folgenden zu besprechende Einteilung der potentiellen Störfaktoren in verschiedene Gruppen ist - das sei ausdrücklich betont-weder die einzig mögliche noch ist unser „Katalog“ von Störfaktoren abgeschlossen und umfassend.)

²⁾ Unberücksichtigt bleiben bei der folgenden Diskussion triviale Störungen der experimentellen Validität durch falsches Verhalten des Versuchsleiters, falsches Aufzeichnen der Antworten, Fehler bei der Berechnung, bewußte Fälschung der Daten usw. (siehe Mosteller, 1968; Barber, 1976).

2. Variablenvalidität (VV)

In unserer exemplarischen Hypothese WH, wird eine Aussage über die Beziehung zwischen zwei Begriffen gemacht: „Dissonanz“ und „Einstellung“. Beide Begriffe sind keine empirischen Begriffe, sondern gehören zur theoretischen Sprache. In diesem Abschnitt wollen wir zunächst erläutern, was diese Einordnung bedeutet, und dann überlegen, welche Konsequenzen sie für die Validität von Experimenten zur Überprüfung von Kausalhypothesen hat.

Nach Carnap (1960) kann man die wissenschaftliche Sprache in zwei Stufen einteilen, indem man zwischen einer *Beobachtungssprache* und einer *theoretischen Sprache* unterscheidet (zur Kritik dieses Ansatzes siehe Suppe, 1977a). Die Beobachtungssprache umfaßt dabei ausschließlich Begriffe, die sich auf beobachtbare Objekte, Eigenschaften und Relationen beziehen sowie solche Begriffe, die sich durch explizite Definitionen vollständig auf diese zurückführen lassen (Stegmüller, 1974b). Wir wollen das Vokabular der Beobachtungssprache kurz als *Beobachtungsbegriffe* bezeichnen. Alle Begriffe, die nicht zur Beobachtungssprache gehören, sind *theoretische Begriffe*. Diese theoretischen Begriffe sind also nicht vollständig auf Beobachtbares zurückzuführen, sie sind nach MacCorquodale & Meehl (1948) *hypothetische Konstrukte* mit einer *Überschußbedeutung*.

Dieses Konzept ist insofern eine Idealisierung, als man inzwischen klar erkannt hat, daß keine Beobachtung voraussetzungs-, d.h. theoriefrei ist (Herrman, 1973; Lakatos, 1974; Suppe, 1977a,b). Deshalb spricht Hempel (1974) statt von einer Beobachtungssprache von einem „vorgängig verfügbaren Vokabular“.

Von besonderer Bedeutung ist nun der Umstand, daß ohne theoretische Begriffe keine Kausalaussagen möglich sind.

Zu Beginn des Abschnittes 1.2 hatten wir bereits erwähnt, daß ein Ereignis (z.B. D) dann als Ursache eines anderen Ereignisses (z.B. dE) bezeichnet werden kann, wenn beide zusammen auftreten und wenn es eine *allgemeine Gesetzmäßigkeit* gibt, wonach D (regelmäßig) dE zur Folge hat (Gadenne, 1976). Eine Aussage wie „ $D \rightarrow dE$ “ kann aber nur dann als allgemeines Gesetz bezeichnet werden, wenn es über einen bestimmten raumzeitlichen Zusammenhang hinaus gilt. Damit können die in der Aussage enthaltenen Begriffe aber nicht mehr ganz bestimmten beobachtbaren Sachverhalten entsprechen, sondern sie beziehen sich auf unbegrenzte Mengen von möglichen „Realisierungen“, die sich zumindest dadurch unterscheiden, daß sie zu unterschiedlichen Zeiten und an unterschiedlichen Orten auftreten. Die in Kausalaussagen auftretenden Begriffe müssen also notwendigerweise theoretische Begriffe sein. Zur Prüfung einer psychologischen Kausalhypothese müssen von daher stets den in ihr enthaltenen theoretischen Begriffen beobachtbare Variablen zugeordnet werden. Aussagen, die einem theoretischen Begriff einen empiri-

schen Begriff zuordnen, bezeichnet man als *Zuordnungsregeln* (Stegmüller, 1974b, 308-319).

Welche Anforderungen sind nun an diese Zuordnungsregeln zu stellen, wenn wir vom Ziel einer strengen Prüfung der Hypothese ausgehen? Oder anders ausgedrückt: Durch welche Mängel bei der Zuordnung von empirischen und theoretischen Begriffen kann die Validität einer Untersuchung als Prüfung einer Kausalhypothese eingeschränkt werden?

Wir wollen die wichtigsten dieser Mängel zu fünf Störfaktoren zusammenfassen. Da sie die Beziehung zwischen empirischen und theoretischen Variablen betreffen, wollen wir diesen Teilaspekt der Validität einer Untersuchung als „Variablenvalidität“ (W) bezeichnen und entsprechend von „Störfaktoren (VV)“ sprechen, um sie von den später noch zu besprechenden anderen Arten von Störfaktoren zu unterscheiden.

2.1 Mangelnde Eindeutigkeit der Zuordnung als Störfaktor (VV)

Aus einer Kausalhypothese wie WH_u ist eine empirisch prüfbare Prognose der oben beschriebenen Form $X \wedge dY$ nur ableitbar, wenn man den theoretischen Begriffen der Hypothese ganz bestimmte empirisch beobachtbare Variablen (*empirische Realisierungen* oder *Operationalisierungen*) zuordnet. Danach muß für jede empirische Variable eindeutig entscheidbar sein, ob sie einem bestimmten theoretischen Begriff zugeordnet ist oder nicht. Liegt diese eindeutige Zuordnung nicht vor, ist es also z.B. unsicher, ob durch bestimmte Maßnahmen X „Dissonanz“ erzeugt wird und/oder ob das Ergebnis Y eines gegebenen Tests eine Entsprechung des theoretischen Begriffs „Einstellung“ darstellt, kann ein auf den ersten Blick hypothesenkonträres Untersuchungsergebnis $X \wedge \neg dY$ darauf zurückgeführt werden, daß Y eine inadäquate Operationalisierung für die „Einstellung“ ist und/oder daß in X gar keine Entsprechung der Dissonanzbedingung D vorgelegen hat. Da in diesem Falle keine der Hypothese widersprechenden empirischen Ergebnisse auftreten können, kann von einer echten Prüfung der Hypothese gar keine Rede sein.

Die „Erklärung“ (erwartungswidriger) empirischer Ergebnisse durch Verweisen auf inadäquate Operationalisierungen findet man sehr häufig in den mit „Diskussion“ überschriebenen Teilen von Forschungsberichten. Entsprechende Argumentationen stellen nicht unbedingt eine besonders lobenswerte „kritische Würdigung“ der eigenen Forschungsbefunde dar, sondern weisen (fast) stets auf (häufig gravierende) Fehler in der Konzeption der Untersuchung hin.

Wie kann man diesen Fehler vermeiden, d.h. wie kommt man zu einer eindeutigen Zuordnung von empirischen zu theoretischen Variablen? Um es gleich vorwegzunehmen: Einen routinemäßig beschreitbaren Weg zu diesem Ziel

gibt es nicht. Wir können deshalb nur exemplarische Überlegungen skizzieren, die dem Leser als Anregungen für eigene Problemlösungen dienen mögen.

Betrachten wir eine Hypothese wie WH., isoliert, ist eine Zuordnung empirischer Variablen zu den theoretischen Begriffen recht willkürlich möglich, wenn wir einmal davon absehen, daß sie nicht unserem Vorverständnis über die verwendeten theoretischen Begriffe widersprechen sollte. Um diese Beliebigkeit einschränken zu können, benötigen wir eine *Theorie*, d.h. eine Menge von Sätzen, die die in unserer Hypothese auftretenden Begriffe dadurch näher spezifiziert, daß sie sie mit anderen theoretischen Begriffen in Beziehung setzt. Sollte durch diese Explikation noch keine eindeutige Ableitung von Operationalisierungen möglich sein, muß man ein Vorgehen wählen, das Ähnlichkeit hat mit der *Konstruktvalidierung* psychologischer Tests (Cronbach & Meehl, 1955; Campbell & Fiske, 1959):

Durch eine Menge theoretischer Sätze entsteht - um ein Bild Hempels (1974) zu benutzen - ein Netz theoretischer Begriffe, zu dem ein „passendes“ Netz empirischer Begriffe gefunden werden muß. Man ordnet empirische und theoretische Begriffe bei einer solchen Betrachtungsweise also nicht einzeln einander zu, sondern versucht, für eine Menge theoretischer Begriffe empirische Entsprechungen so zu finden, daß die nach der Theorie bestehenden Verbindungen zwischen den theoretischen Begriffen sich in den statistischen Assoziationen zwischen den entsprechenden empirischen Variablen wiederfinden. Durch einen Rekurs auf einen größeren theoretischen Zusammenhang kann also die Eindeutigkeit in der Zuordnung von empirischen zu theoretischen Begriffen erhöht werden. Allerdings sind sehr viele psychologische Theorien noch zu unpräzise formuliert, als daß man die jeweilige Menge der möglichen Operationalisierungen für ihre Begriffe tatsächlich genau spezifizieren kann. Damit sind solche Theorien und die aus ihr abgeleiteten Hypothesen aber nur in begrenztem Maße einer strengen Prüfung zugänglich.

2.2 Mangelnde konzeptuelle Replikation als Störfaktor (VV)

Nach der hier zugrunde gelegten Zweisprachenkonzeption von Carnap (1960) haben theoretische Begriffe gegenüber der Beobachtungssprache eine Überschußbedeutung, d.h. die Bedeutung eines theoretischen Begriffs kann durch endlich viele Beobachtungsbegriffe nicht vollständig erfaßt werden. Nun wird für einen theoretischen Begriff kaum je eine optimale empirische Entsprechung zu spezifizieren sein, in der Regel werden verschiedene mögliche empirische Realisierungen zur Auswahl stehen. So kann „Dissonanz“ in ganz unterschiedlichen praktischen Situationen auftreten, und Einstellungen werden sowohl aus verbalen Antworten wie aus dem beobachtbaren Verhalten in natürlichen Situationen ermittelt (vgl. Cook & Selltitz, 1964). Wird eine Untersuchung lediglich mit anderen Operationalisierungen wiederholt, spricht man

von einer *konzeptuellen Replikation* (Carlsmith, Ellsworth & Aronson, 1976, 64-81; Bredenkamp, 1979).

Wir wollen die konzeptuelle Replikation als um so stärker bezeichnen, je größer die Anzahl der berücksichtigten empirischen Entsprechungen ist und je verschiedenartiger die Bereiche sind, aus denen diese Realisierungen stammen.

Es kann durchaus der Fall auftreten, daß die zu prüfende Hypothese nur für einen Teil der möglichen empirischen Realisierungen gültig ist, für den anderen dagegen nicht. So mag unsere Hypothese WH_u bspw. zutreffen, wenn „Dissonanz“ durch einstellungskonträre Handlungen operationalisiert wird, nicht aber wenn die empirische Entsprechung einstellungskonträre Information ist. In ihrer allgemeinen Formulierung wäre die WH_u dann falsch. In einem solchen Falle ist die Wahrscheinlichkeit, bei Falschheit der Hypothese auch ein ihr widersprechendes empirisches Ergebnis zu erhalten, um so höher, je stärker die konzeptuelle Replikation ist. *Deshalb ist allgemein die Prüfung einer Hypothese um so strenger, je stärker konzeptuell repliziert wird.* Konzeptuelle Replikationen sind insbesondere dann notwendig, wenn es bereits bewährte Störungshypothesen gibt, aus denen hervorgeht, daß die geprüfte Hypothese nur für bestimmte Realisierungen gelten könnte.

Für den theoretischen Begriff, der der UV der geprüften Hypothese entspricht, kann in einer Untersuchung meist nur eine empirische Entsprechung einbezogen werden. Die konzeptuelle Replikation der AV ist zwar durch multivariate Untersuchungen mit theoretisch unbegrenzt vielen Entsprechungen der AV leichter möglich, versuchs- und auswertungstechnische Probleme setzen aber auch hier i.a. enge Grenzen (siehe Teil 8). Deshalb sind für eine strenge Prüfung einer Hypothese mehrere Untersuchungen notwendig, die sich nur hinsichtlich der vorgenommenen Operationalisierungen unterscheiden.³⁾ Erst danach wäre eine Falsifikation der Hypothese zu rechtfertigen.

Ergeben sich bei der konzeptuellen Replikation unterschiedliche Resultate im Hinblick auf die Gültigkeit der Hypothese, braucht das nicht unbedingt zu einer generellen Verwerfung der Hypothese zu führen, sondern kann auch Anlaß sein für eine Neuabgrenzung oder Differenzierung des entsprechenden theoretischen Begriffs (Carlsmith, Ellsworth & Aronson, 1976; Bredenkamp, 1979) oder für eine Änderung der Zuordnungsregel.

Auf jeden Fall machen die bisherigen Ausführungen zur Variablenvalidität deutlich, daß kaum je eine wissenschaftliche Hypothese schon aufgrund des Ergebnisses einer einzigen empirischen Untersuchung als falsifiziert oder gut bewährt bezeichnet werden sollte.

³⁾ Von den möglichen Operationalisierungen stehen dem Experimentator aus technischen oder ethischen Gründen meist nicht alle zur Verfügung (zu den ethischen Problemen beim Experimentieren siehe Klauer, 1973, 149-160; Carlsmith, Ellsworth & Aronson, 1976, 93-117; Schuler, 1980).

2.3 Mangelnde Entsprechung im Variationsbereich von theoretischen und empirischen Variablen als Störfaktor (VV)

Diese mögliche Beeinträchtigung der Variablenvalidität betrifft fast ausschließlich die unabhängigen Variablen. Für die ihnen entsprechenden theoretischen Variablen kann jeweils ein bestimmter Variationsbereich umschrieben werden. So kann unsere Variable „Dissonanz“ von „fehlender Dissonanz“ in - idealiter - unendlich vielen verschiedenen Abstufungen bis zu einer „extrem starken Dissonanz“ schwanken. Für eine adäquate empirische Realisierung ist ein entsprechender Variationsbereich zu fordern. Wird eine theoretische Variable im Experiment als UV operationalisiert, werden aus der großen Zahl von möglichen Ausprägungen meist nur relativ wenige berücksichtigt. Nun ist es aber durchaus möglich, daß die geprüfte Hypothese nur für bestimmte Teilmengen der Ausprägungen gültig ist (z.B. könnten nur mittlere Dissonanzen zu Einstellungsänderungen führen). Ein Experiment ist folglich um so valider, je mehr Ausprägungen der UV berücksichtigt sind bzw. je vollständiger die verwendeten Ausprägungen dem möglichen Variationsbereich der theoretischen UV entsprechen. Dabei kommt es nicht nur darauf an, daß empirische Entsprechungen für die extremen Bereiche der theoretischen Variablen vorhanden sind, vielmehr muß der dazwischenliegende Bereich repräsentiert sein, um auch U-förmige oder noch kompliziertere Beziehungen zwischen unabhängiger und abhängiger Variable entdecken zu können (vgl. Wormser, 1974).

Da die Menge der möglichen Ausprägungen der UV meist entweder unendlich oder nicht genau zu spezifizieren ist, ist die angestrebte Entsprechung nicht durch eine Zufallsauswahl der zu verwendenden Ausprägungen zu erreichen, sondern nur durch eine systematische Auswahl (z.B. je eine Behandlungsbedingung mit großer, mittlerer und fehlender Dissonanz). Dann besteht auch die Möglichkeit, neben der Hypothese über den generellen Einfluß der UV auf die AV noch spezifischere Hypothesen über die Beziehung in einzelnen Teilbereichen der UV zu prüfen. Gerade bei der Durchführung von Experimenten in der Sozialpsychologie ist es oft gar nicht so einfach, extreme Ausprägungen der theoretischen unabhängigen Variablen zu realisieren. Die im Experiment möglichen Behandlungen sprechen nämlich wegen der Künstlichkeit der Situation in der Regel die Probanden viel zu wenig an, um beispielsweise eine starke Dissonanz zu erzeugen (Carlsmith, Ellsworth & Aronson, 1976). Von daher stellen Experimente nicht unbedingt immer die strengstmögliche Prüfung einer Hypothese dar (vgl. Teil 4 und 5). Vielmehr müssen wir gerade im Experiment mit einer Störung der Variablenvalidität dadurch rechnen, daß die experimentelle Manipulation nur geringen Unterschieden auf der theoretischen UV entspricht und deshalb u.U. nicht die von der Hypothese vorhergesagte Wirkung auf die AV zeigt. In Abhängigkeit von der Art der wissenschaftlichen Hypothese (vgl. Abschn. 8.1) kann eine solche Störung der Variablenvalidität zu ungerechtfertigten Falsifikationen oder Bestätigungen führen.

2.4 Zu geringes Skalenniveau als Störfaktor (VV)

Wir beginnen mit einer These, die wir im folgenden erläutern und begründen:

Die Prüfung einer wissenschaftlichen Hypothese ist um so strenger, je besser das Skalenniveau der empirischen Variablen der Struktur der theoretischen Begriffe entspricht.

Die Struktur eines theoretischen Merkmals oder Begriffs ist festgelegt durch die Art der Relationen, die auf der Menge aller Merkmalsausprägungen definiert sind (vgl. dazu Stegmüller, 1974b). Ein Merkmal hat die einfachste Form einer Struktur, wenn man lediglich mehrere einander ausschließende und erschöpfende Ausprägungen unterscheidet, bei Einstellungen beispielsweise „konservative“, „liberale“ und „sonstige“. Geht man davon aus, daß zwischen den Ausprägungen eine Rangordnung besteht, berücksichtigt man schon ein Strukturmerkmal mehr. Man gelangt dadurch zu *komparativen Begriffen*. Die höchste uns interessierende Form der Strukturiertheit haben *metrische Begriffe*: Bei ihnen sind zwischen den einzelnen Ausprägungen auch *Abstände* definiert, man will also z.B. eine Aussage darüber machen können, ob der Unterschied in der Einstellung zu Gastarbeitern zwischen Alfred und Bruno größer ist als zwischen Claus und Dirk. Nehmen wir nun an, wir hätten dem uns interessierenden theoretischen Einstellungsbegriff als empirische Entsprechung das numerische Ergebnis Y eines genau definierten Skalierungsverfahrens zugeordnet. Bei einem komparativen Einstellungsbegriff würde das etwa der folgenden Zuordnungsregel entsprechen:

- (1) Wenn Person 1 eine positivere Einstellung zu Gastarbeitern hat als Person 2, ist Y_1 größer als Y_2 .

Bei einem metrischen Einstellungsbegriff muß zusätzlich noch die Entsprechung von Abständen auf der theoretischen Variablen und Differenzen auf der empirischen Variablen festgelegt werden:

- (2) Wenn der Unterschied in den Einstellungen der Person 1 und 2 größer ist als der bei den Personen 3 und 4, dann gilt $|Y_1 - Y_2| > c |Y_3 - Y_4|$.⁴⁾

Die Zuordnungsregel (1) ist aber nur dann sinnvoll, wenn Y eine Messung mindestens auf Ordinalskalenniveau ist (vgl. Suppes & Zinnes, 1963). Haben die Y-Werte nämlich kein Ordinal-, sondern nur Nominalskalenniveau, kön-

⁴⁾ Zuordnungsregeln dieser Art sind keine „operationalen Definitionen“ im Sinne Bridgmans (1927), weil sie eine andere logische Struktur aufweisen. Ferner gehen wir davon aus, daß einem theoretischen Begriff durch mehrere Zuordnungsregeln mehrere empirische Variablen entsprechen können. Siehe zur Inadäquatheit von operationalen Definitionen u.a. Bunge (1967a), Herrmann (1973) und Stegmüller (1974b).

nen die von Personen zugeordneten Zahlenwerte insofern beliebig transformiert werden, als lediglich gewährleistet sein muß, daß genau die Personen, die vor der Transformation *gleiche* Zahlenwerte auf der Variablen Y hatten, auch nach der Transformation *gleiche* Zahlenwerte zugeordnet bekommen. Drückt man diese Forderung formal aus, ergibt sich: Hat die Zahlenzuordnung nur Nominalskalenniveau, sind beliebige eindeutige Transformationen erlaubt. Damit sind aber auch nicht-monotone Transformationen zulässig, und Aussagen über die Rangordnung von Zahlenwerten sind empirisch nicht sinnvoll, da ihr Wahrheitswert sich unter diesen zulässigen Transformationen ändert.

Entsprechend kann begründet werden, daß Zuordnungsregel (2) nur dann sinnvoll ist, wenn Y Intervallskalenniveau aufweist. Nur in diesem Fall bleibt nämlich bei erlaubten Transformationen der Zahlen die in der Zuordnungsregel (2) getroffene Aussage über die Größenordnung von Zahlendifferenzen in ihrem Wahrheitswert erhalten.

Welche Konsequenzen hat es für die Prüfung einer Kausalhypothese, wenn das durch die Struktur der theoretischen Begriffe und die Art der Zuordnungsregel geforderte Skalenniveau nicht gegeben ist? Nehmen wir als Beispiel an, unsere Einstellungsvariable Y habe kein Ordinalskalenniveau. Dann ist die Zuordnungsregel (1) nicht mehr sinnvoll, und die Aussage der Hypothese WH_u kann nicht mehr in die Aussage „übersetzt“ werden, daß unter der Dissonanzbedingung der Wert der Variablen Y größer ist als unter der Bedingung „keine Dissonanz“. über diese Prognose ist WH_u also nicht mehr prüfbar. Wir können diesen Sachverhalt auch anders formulieren: Wird aus einer wissenschaftlichen Hypothese eine Aussage über die Rangordnung von Einzelwerten oder Medianen auf einer empirischen Variablen abgeleitet, muß diese Variable mindestens Ordinalskalenniveau haben. Sind in der Hypothese metrische Begriffe enthalten und wird aus ihr eine Vorhersage abgeleitet, die zum Beispiel die Rangordnung arithmetischer Mittelwerte oder die Größe von Produkt-Moment-Korrelationskoeffizienten betrifft, muß mindestens Intervallskalenniveau vorliegen, da diese Aussagen ihren Wahrheitswert verändern können, wenn alle monotonen Transformationen erlaubt sind - zur Begründung siehe Suppes & Zinnes (1963) und Orth (1974).

Allgemein gilt: Liegt das durch die Struktur der theoretischen Begriffe und die Art der aus der wissenschaftlichen Hypothese abgeleiteten Vorhersage geforderte Skalenniveau nicht vor, kann die Hypothese über diese Vorhersage nur einer weniger strengen Prüfung unterzogen werden.

Diese Überlegungen haben wichtige Konsequenzen für die Anwendung statistischer Testverfahren. Wie im Teil 6 noch zu zeigen sein wird, werden zur Überprüfung wissenschaftlicher Kausalhypothesen aus ihnen statistische Hypothesen abgeleitet. Die üblichen parametrischen Testverfahren wie der t- und

der F-Test (s. dazu im einzelnen Teil 7) prüfen vor allem statistische Hypothesen über Gleichheit und Rangordnung von arithmetischen Mittelwerten bzw. über Produkt-Moment-Korrelationen. Diese statistischen Hypothesen stellen nach dem oben Gesagten nur sinnvolle Aussagen dar, wenn die jeweilige empirische Variable mindestens Intervallskalenniveau hat. Außerdem müssen zur mathematischen Begründung dieser Tests u.a. Annahmen über Normalverteilung und Varianzgleichheit gemacht werden, die ebenfalls nicht sinnvoll sind, wenn nur Ordinalskalenniveau vorliegt, da sich diese Verteilungsaspekte bei bestimmten monotonen Transformationen ändern (s. Abschn. 8.2). Obwohl ein bestimmtes Skalenniveau nicht zu den *mathematischen* Voraussetzungen der parametrischen Testverfahren gehört (Lord, 1953; Gaito, 1960b, 1980; Anderson, 1961; McNemar, 1962; S. 375), ist deren Anwendung und Interpretation aus den genannten Gründen nur sinnvoll, wenn mindestens Intervallskalenniveau vorliegt.⁵⁾ Wie wir am Ende dieses Abschnitts sehen werden, wird dadurch ihr Anwendungsgebiet aber weniger eingeschränkt, als es auf den ersten Blick scheinen mag.

Gehen wir zunächst auf die Frage ein, wie man das Skalenniveau einer Variablen bestimmt. Nach der Definition von Suppes & Zinnes (1963) stellt eine Zuordnung von Zahlen zu Objekten dann eine Messung dar, wenn durch die Beziehungen zwischen den Zahlen empirisch beobachtbare Beziehungen zwischen den Objekten widergespiegelt werden. Sind diese empirischen Beziehungen unabhängig von jeder Zahlenzuordnung beobachtbar, kann man von einer *Repräsentationsmessung* sprechen (s. Dawes, 1977). Die hinreichenden Bedingungen dafür, daß eine solche Zahlenzuordnung Ordinal- bzw. Intervallskalenniveau hat, sind in den Axiomen sogenannter *Meßstrukturen* formuliert (Krantz et al., 1971; Orth, 1974). Um in einem konkreten Anwendungsfall das Skalenniveau zu bestimmen, muß geprüft werden, ob die empirisch beobachtbaren Beziehungen zwischen den Meßobjekten, die durch die Zahlen repräsentiert werden sollen, diese Bedingungen erfüllen (zur praktischen Durchführung dieser Prüfung siehe Westermann, 1980, im Druck, b).

Für die meisten in der Psychologie verwendeten Zahlenzuordnungen lassen sich jedoch gar keine unabhängig beobachtbaren Beziehungen zwischen Objekten auffinden, die eventuell repräsentiert werden können. Wir sprechen dann nicht von einer Repräsentations-, sondern von einer *Indexmessung* (vgl. Suppes & Zinnes, 1963; Dawes, 1977; Allerbeck, 1978). Solchen Variablen

⁵⁾ Zur kontroversen Diskussion über die Beziehung zwischen Skalenniveau und Statistik siehe z.B. Stevens, 1951; Lord, 1953; Suppes & Zinnes, 1963; Adams, Fagot & Robinson, 1965; Baker, Hardyck & Petrinovich, 1966; Pfanzagl, 1968, S. 34-56; Gardner, 1975; Lantermann, 1976; Dawes, 1977, Kap. 7; Steinfatt, 1977; Allerbeck, 1978 sowie ferner die Reader von Heermann & Braskamp, 1970; Liebermann, 1971; Steger, 1971 und Kirk, 1972.

kann jedes gewünschte Skalenniveau zugeschrieben werden, falls sie sich nur irgendwie auf eine Abzähloperation zurückführen lassen („*Messung per fiat*“).

Postuliert man z.B. Intervallniveau für das Ergebnis eines Fragebogentests (gegeben durch die Anzahl der positiven Antworten), und ordnet man diese Variable einem metrischen Einstellungsbegriff zu, muß man sich darüber im klaren sein, daß damit die Annahme verbunden ist, daß z.B. zwischen Personen, die 4 und 6 positive Antworten gegeben haben, der gleiche Einstellungsunterschied besteht wie zwischen Personen mit 22 und 24 positiven Antworten. Man definiert damit praktisch die Abstände hinsichtlich des theoretischen Begriffs in Abhängigkeit von einer konkreten Zahlenzuordnung. Dieses Vorgehen ist nicht von vornherein unberechtigt, denn der Wissenschaftler ist ja in der Definition seiner theoretischen Begriffe relativ frei, wenn er mit Herrmann (1973) davon ausgeht, daß diese theoretischen Begriffe keine realen Entitäten mit einer zu entdeckenden Struktur bezeichnen.

Für einen Wissenschaftler, der empirisch gehaltvolle und überprüfbare Theorien aufstellen will, wird es aber immer das Ziel sein müssen, repräsentationale Messungen mindestens auf Intervallskalenniveau zu verwenden. Bei der Indexmessung drücken die Zahlen nur die unterschiedlichen Antworthäufigkeiten in einem gegebenen Erhebungsverfahren aus. Bei einer Repräsentationsmessung auf Intervallskalenniveau dagegen repräsentieren Ordnung und Abstände zwischen Zahlen empirisch beobachtbare Relationen, wie sie z.B. entstehen, wenn Probanden persönliche Präferenzen und subjektive Unterschiede hinsichtlich eines definierten Merkmals bei verschiedenen Objekten ausdrücken. Von daher ist die Zuordnung von theoretischen und empirischen Variablen bei einer Repräsentationsmessung wesentlich weniger beliebig als bei einer Indexmessung, wodurch die entsprechenden Hypothesen einer stärkeren Prüfung zugänglich werden (siehe exemplarisch für den Einstellungsbegriff Westermann, 1982).

Daß psychologische Variablen auf Intervallskalenniveau gemessen werden, ist unter anderem deshalb wichtig, weil bei Ordinalskalenniveau lediglich Aussagen über monotone Zusammenhänge zwischen Variablen gemacht werden können und Aussagen über die Art der Funktion (linear, S-förmig usw.) erst sinnvoll sind, wenn mindestens Intervallskalenniveau vorliegt.

Welche neuen Aspekte ergeben sich nun für die Prüfung statistischer Hypothesen durch parametrische Testverfahren? Zwar sollte bei subjektiven Variablen wie „Einstellung zu Gastarbeitern“, „Arbeitszufriedenheit“ oder „Angst“ die Annahme des Intervallskalenniveaus durch die empirische Prüfung meßtheoretischer Axiome fundiert werden, doch betreffen viele empirische Voraussagen, die zur Überprüfung wissenschaftlicher Hypothesen abgeleitet werden, physikalische Variablen wie „Reaktionszeit“ oder „Hautwiderstand“. Diese haben i.d. R. Verhältnisskalenniveau (zur Begründung Orth, 1974, 47-49). Andere Vorhersagen betreffen Variablen, die als einfache Abzählungen interpretiert werden können und damit zwangsläufig auf Absolutskalenni-

veau liegen. So könnte - als Beispiel zum letzten Punkt - aus einer Kausalhypothese abgeleitet werden, daß die Anzahl der positiven Antworten zu Gastarbeitern im Fragebogen FB um so größer ist, je stärker die erzeugte Dissonanz war. Falls keine zu starken Abweichungen von den mathematischen Voraussetzungen für die parametrischen Tests vorliegen (s. Abschn. 8.2), können sie auch in diesem Fall verwendet werden, um die abgeleitete statistische Hypothese zu überprüfen. Überhaupt kein Hindernis für die Anwendung parametrischer Tests stellen Probleme des Skalenniveaus dar, wenn man keine wissenschaftliche Hypothese prüft, die in theoretischen Begriffen formuliert ist, sondern vielmehr die statistische Hypothese selbst als wissenschaftliche Hypothese betrachtet (s. Abschn. 1.2).

Dies ist der Fall, wenn man z.B. nicht prüfen will, ob bei „Dissonanz“ die „Einstellungen“ positiver werden, sondern nur wissen will, ob unter einer bestimmten Bedingung A die Werte im Einstellungstest ET höher sind als unter der Bedingung nicht-A. Da hier nur der Zusammenhang von ganz bestimmten, genau definierbaren empirischen Variablen untersucht wird, sind bezüglich der AV keinerlei Transformationen sinnvoll, und man kann diese Variable zur Intervall- oder Verhältnisskala deklarieren (Messung per fiat).

2.5 Konfundierung von theoretischen Begriffen als Störfaktor (VV)

Bisher haben wir für eine strenge Prüfung u.a. gefordert, daß die Menge der möglichen empirischen Realisierungen für einen theoretischen Begriff genau abzugrenzen ist. Das schließt nicht aus, daß eine bestimmte empirische Variable als Operationalisierung für mehrere theoretische Begriffe betrachtet werden kann. Man bezeichnet diese Begriffe dann als *konfundiert*. In diesem Fall ist es leichter möglich, gut bewährte Störungshypothesen zu finden, die die gleiche Prognose erlauben wie die zu prüfende Hypothese. Dadurch kann die Möglichkeit einer strengen Prüfung dieser Hypothese stark gefährdet werden. Wir wollen dies anhand von vier Fallgruppen näher untersuchen.

1. Recht häufig wird in der Psychologie für eine gegebene Menge experimenteller Bedingungen, die Operationalisierungen unterschiedlicher Ausprägungen eines theoretischen Begriffs darstellen sollen, eine *alternative theoretische Interpretation* gegeben.

So führt beispielsweise Bern (1972, 1975) die Einstellungsänderungen in den Experimenten zur Prüfung von Festingers Dissonanztheorie nicht auf Unterschiede in der erzeugten Dissonanz zurück, sondern geht davon aus, daß experimentelle Bedingungen, die nach Festinger unterschiedlichen Ausmaßen an „kognitiver Dissonanz“ entsprechen, zumeist für die Probanden unterschiedliche „Selbstwahrnehmungen“ mit sich bringen, die dann zu unterschiedlichen „Einstellungen“ führen.

Eine solche Konfundierung theoretischer Begriffe schränkt aber die Möglichkeit strenger Prüfungen aller beteiligten Hypothesen und Theorien nicht ein, vielmehr wird durch sie überhaupt erst die Möglichkeit eröffnet, bei empirischer Nichtbewährung einer Hypothese oder Theorie T_1 diese empirischen Ergebnisse im Rahmen einer anderen Theorie T_2 mit anderen theoretischen Begriffen zu erklären. Diese Art von Konfundierung ist also kein Mangel, sondern Voraussetzung für Erkenntnisfortschritt (vgl. Abschn. 1.3).

2. Auch die abhängige Variable einer Untersuchung entspricht kaum je eindeutig dem theoretischen Begriff, dem sie zugeordnet ist. In unterschiedlichen Werten der Personen auf der AV können sich vielmehr nicht nur unterschiedliche Ausprägungen der entsprechenden theoretischen Variablen ausdrücken, sondern auch Unterschiede auf einer ganzen Reihe von anderen theoretischen Variablen. So können nicht nur positivere Ausprägungen der Variablen „Einstellung zu Gastarbeitern“ zu höheren Werten im entsprechenden Einstellungstest führen, sondern beispielsweise auch stärkere Ausprägungen von Variablen wie „Antworten gemäß sozialer Erwünschtheit“, „Ja-Sage-Tendenz“ und „Tendenz zu extremen Antworten“. Daß solche sog. *Reaktionsstile* („response sets“) mit der in der Untersuchung interessierenden theoretischen abhängigen Variablen konfundiert sind, ist typisch für die in der Sozialforschung verbreiteten verbalen Skalierungsmethoden. Man kann den störenden Einfluß dieser Art von Konfundierung auf die Validität der Untersuchung verringern, indem man im Rahmen einer konzeptuellen Replikation (siehe Abschn. 2.2) auch „nicht-verbale“ Realisierungen des entsprechenden theoretischen Begriffs einbezieht und/oder indem man das Erfassungsinstrument so konstruiert, daß diese Antworttendenzen sich möglichst wenig auswirken können (siehe dazu Edwards, 1957a,b, 1970; Cook & Selltitz, 1964; Berg, 1967; Scott, 1968; Holm, 1975, 1976, 1977; Koch, 1976).

Auch wenn die *Reliabilität* der empirischen Variablen gering ist, liegt eine Störung der Variablenvalidität durch Konfundierung theoretischer Begriffe vor: Im Sinne der klassischen Testtheorie (vgl. Lord & Novick, 1968; Fischer, 1974; Kranz, 1979; Wottawa, 1980) sind dann in den beobachteten Werten der empirischen Variablen Meßfehler mit den sog. „wahren“ Werten in einem zu großen Maße vermengt.

3. Nach einer hauptsächlich auf Orne (1962) zurückgehenden Auffassung verhalten sich Teilnehmer eines Experimentes gemäß ihren Hypothesen darüber, was im Experiment von ihnen erwartet wird, was Ziel und Sinn dieses Experimentes ist.⁶⁾ Die Art der von den Versuchspersonen (Vpn) gebildeten Hypo-

⁶⁾ Zu Einzelheiten über Forschungen zur Sozialpsychologie des Experimentes siehe z.B. Rosenthal & Rosnow, 1969a; Klauer, 1973; Timaeus, 1974, 1975; Kruglanski, 1975; Mertens, 1975; Barber, 1976; Gniech, 1976; Rosnow & Davis, 1977; Silverman, 1977; Rosenthal & Rubin, 1978; Bungard, 1980; Callaway, Nowicki & Duke, 1980.

thesen hängen von gewissen „*demand characteristics*“ der experimentellen Situation ab. Eine dieser Determinanten der Probanden-Hypothesen sind die Erwartungen des Versuchsleiters, deren Auswirkungen ausführlich von Rosenthal (1969, 1977) untersucht wurden. Werden nun die Ausprägungen einer theoretischen Variablen durch unterschiedliche experimentelle Bedingungen operationalisiert, haben diese Bedingungen unterschiedliche „*demand characteristics*“ und können folglich zu unterschiedlichen Hypothesen bei den Vpn führen. Unter der Gültigkeit der Auffassung von Orne wären also die theoretische Variable, die der experimentellen UV entspricht, und die Variable „Hypothesen der Vpn“ zwangsläufig konfundiert. Nun ist zwar die Annahme, daß das Verhalten im Experiment generell von den Hypothesen der Vpn bestimmt wird, empirisch kaum zu stützen (Bredenkamp 1980, 41-47), doch muß man in einzelnen Untersuchungen immer mit solchen Effekten rechnen. Um eine möglichst strenge Prüfung einer Hypothese zu erreichen, muß man sich folglich bemühen, die Bildung entsprechender VP-Hypothesen zu erschweren, bzw. man muß verhindern, daß eine eventuelle Hypothesenbildung bei den Vpn die gleiche empirische Auswirkung hat, die auch von der zu prüfenden Kausalaussage vorhergesagt wird. In unserem Fall muß etwa verhindert werden, daß die Vpn unter der Dissonanz-Bedingung leicht erkennen können, daß von ihnen eine Einstellungsänderung erwartet wird.

Es gibt eine Reihe von Techniken, die verhindern sollen, daß Probanden unter verschiedenen experimentellen Bedingungen unterschiedliche Hypothesen über das von ihnen erwartete Verhalten bilden (siehe Carlsmith, Ellsworth & Aronson, 1976, 280-301 und Rosnow & Davis, 1977):

- man vermeidet, daß Personen Unterschiede zwischen den Behandlungsbedingungen überhaupt wahrnehmen können (Blindversuch, ähnlich den Placebo-Experimenten in der Pharmaforschung),
- man verleitet alle Personen dazu, falsche Hypothesen zu bilden,
- man erhebt die AV in einer anderen Situation,
- man täuscht die Personen über ihre Rolle als Versuchspersonen, z.B. indem man sie glauben macht, sie seien die Versuchsleiter,
- man führt ein Feldexperiment (s. Abschn. 3.6) durch, in dem die Personen nicht bemerken, daß sie Teilnehmer an einer wissenschaftlichen Untersuchung sind,
- man untersucht als AV statt leicht zu gebender und folgenloser verbaler Antworten offene Verhaltensweisen, die für die Probanden auch subjektiv bedeutsam sind,
- man vermeidet wiederholte Erfassungen der AV an den gleichen Personen (s.a. Abschn. 3.2).
- man prüft Hypothesen, die schwierig zu „erraten“ sind,
- man bittet die Probanden um ehrliches und unvoreingenommenes Antwortverhalten.

Vom Experimentator ausgehende störende Beeinflussungen des Versuchspersonen-Verhaltens können vor allem dadurch eingeschränkt werden, daß der Versuchsleiter die geprüfte Hypothese nicht kennt (naiver Versuchsleiter) und/oder daß er nicht weiß, welcher experimentellen Bedingung die jeweils untersuchte Person zugeordnet ist (Doppelblindversuch). Auch eine Standardisierung der Kommunikation zwischen Versuchsleiter und -personen kann die Gefahr einer solchen Validitätsstörung vermindern.

4. Wir wollen jetzt eine Art von Konfundierung ansprechen, deren Ausmaß davon abhängt, wie der Experimentator die experimentellen Modalitäten definiert. Stellen wir uns dazu zwei unterschiedliche empirische Realisierungen von „starker Dissonanz“ (X_1) und „fehlender Dissonanz“ (X_2) für eine Person mit einer negativen Einstellung zu Gastarbeitern vor. (Im folgenden beziehen sich die eingeklammerten Indices auf die beiden Fälle 1 und 2.)

- Fall 1: $X_{(1)1}$: Einen Abend mit einer sehr angenehmen Gastarbeiterfamilie verbringen.
 $X_{(1)2}$: Einen negativen Text über Gastarbeiter lesen.
- Fall 2: $X_{(2)1}$: Einen sehr positiven Text über Gastarbeiter lesen.
 $X_{(2)2}$: Einen negativen Text über Gastarbeiter lesen.

Die Bedingungen $X_{(1)1}$ und $X_{(1)2}$ unterscheiden sich nicht nur hinsichtlich der „Dissonanz“, sondern auch noch hinsichtlich der empirischen Entsprechungen einer großen Zahl anderer theoretischer Variablen. Demzufolge sind im Fall 1 eine Fülle von Störungshypothesen denkbar, die die gleiche Prognose erlauben wie die WH_u . So könnte eine positivere Einstellung unter $X_{(1)1}$ als unter $X_{(1)2}$ durch die recht gut bewährte Hypothese vorhergesagt werden, daß Einstellungsänderungen durch persönliche Kontakte eher erfolgen als durch verbale Information. Auch im Fall 2 können sich die Bedingungen noch hinsichtlich anderer Variablen als „Dissonanz“ unterscheiden (z. B. „Verständlichkeit“ oder „Glaubwürdigkeit“ etc.), ihre mögliche Anzahl ist aber doch weit geringer als im Fall 1. Daher sind auch weniger bewährte Störungshypothesen zu erwarten. Im Fall 2 liegt also eine strengere Prüfung unserer Hypothese WH , vor als im Fall 1.

In diesem Beispiel wird deutlich, daß zur strengen Prüfung einer Hypothese die experimentellen Bedingungen, die empirische Realisierungen von verschiedenen Ausprägungen der theoretischen unabhängigen Variablen darstellen, sich so weit ähneln sollen, daß hinsichtlich möglichst weniger anderer Variablen systematische Unterschiede bestehen. Dieses Ziel ist allerdings besonders in sozialpsychologischen Untersuchungen mit relativ komplexen theoretischen Variablen recht schwer zu erreichen (siehe Carlsmith, Ellsworth & Aronson, 1976, 61-64, sowie ferner Underwood & Shaughnessy, 1975, 28-36).

2.6 Zusammenfassung

Wir haben in diesem Abschnitt eine erste Gruppe von Faktoren kennengelernt, die dazu führen können, daß eine Untersuchung keine strenge Prüfung der gegebenen Hypothese darstellt. Da diese Faktoren die Beziehung zwischen den in der Untersuchung auftretenden beobachtbaren Begriffen (Variablen, Bedingungen) und den entsprechenden theoretischen Variablen aus der zu prüfenden Hypothese betreffen, haben wir sie als Störfaktoren der Variablen-Validität - Störfaktoren (VV) - bezeichnet.

Nach unseren Überlegungen ist eine Kausalhypothese um so strenger prüfbar,

- je eindeutiger den theoretischen Begriffen der Hypothese empirische Variablen („Operationalisierungen“) zugeordnet sind,
- je mehr Operationalisierungen der theoretischen Begriffe berücksichtigt werden und je unterschiedlicher diese Operationalisierungen sind,
- je mehr die berücksichtigten Ausprägungen der UV den möglichen Ausprägungen der entsprechenden theoretischen Variablen entsprechen,
- je eher das Skalenniveau der empirischen Variablen der Struktur der zugehörigen theoretischen Begriffe entspricht,
- je weniger bei den berücksichtigten Operationalisierungen andere theoretische Variablen mit den Begriffen der Hypothese konfundiert sind.

(Die Konstanz aller anderen Bedingungen ist jeweils vorausgesetzt.)

3. Interne Validität

Wie wir im Abschnitt 1.2 gesehen haben, folgt aus der Bedeutung von Kausalaussagen, daß zu ihrer Überprüfung Beobachtungen unter mindestens zwei Bedingungen nötig sind, die sich möglichst nur dahingehend unterscheiden, daß in einer Bedingung eine Entsprechung der in der Kausalaussage spezifizierten Ursache vorliegt, in der anderen jedoch nicht. Wir wollen das am Beispiel unserer Hypothese WH_u kurz erläutern.

Um diese WH_u prüfen zu können, kann im Rahmen eines einfachen Versuchesplanen (Vpl.) oder Designs wie folgt verfahren werden: Zunächst wird bei der Person die AV ohne Dissonanz (X_1) gemessen (Variable Y_{11}), anschließend wird Dissonanz erzeugt (X_2) und daraufhin wird die AV erneut gemessen (Y_{12}). Bei einem derartigen Vorgehen spricht man von „intraindividueller Bedingungsvariation“, die folgendermaßen symbolisiert werden kann:

(Vpl.) 1) X_1 Y_{11} X_2 Y_{12}

Wählt man dagegen die sog. „interindividuelle Bedingungsvariation“, ist wie folgt vorzugehen: Bei einer Person wird Dissonanz erzeugt (X_2), bei einer

anderen jedoch nicht (X_1), und die Einstellung der Personen wird gemessen (Y_{22} und Y_{11}). Der resultierende Versuchsplan ergibt sich wie folgt:

$$\begin{array}{cc} \text{(Vpl.2)} & 2) & \begin{array}{cc} X_1 & Y_{11} \\ X_2 & Y_{22} \end{array} \end{array}$$

Unter Berücksichtigung der Zuordnungsregel (1) aus Abschnitt 2.4 wird bei Gültigkeit der Hypothese WH_u erwartet, daß Y_{12} größer als Y_{11} ist. Die gleiche Vorhersage läßt sich jedoch auch mit Hilfe anderer Hypothesen ableiten, falls sich die Messungen Y_{11} und Y_{12} nicht nur dadurch unterscheiden, daß vor Y_{12} die Dissonanzbedingung realisiert wurde, vor Y_{11} dagegen nicht.

Wir wollen diese Bedingungen, die ebenfalls zu unterschiedlichen Werten auf der AV Y führen können, als „Störfaktoren der internen Validität“ bezeichnen.⁷⁾ Sie können zusätzlich zu den experimentellen Bedingungen (X_1 versus X_2) wirksam werden oder aber an ihrer Stelle. Die wichtigsten dieser Störfaktoren haben wir zu zwei Gruppen zusammengefaßt, auf die wir im folgenden eingehen - eine etwas andere Einteilung findet sich bei Campbell & Stanley (1963) und bei Cook & Campbell (1979).

3.1 Variation personaler und situationaler Merkmale als Störfaktoren (IV)

3.1.1 Variation situationaler Merkmale

Beim Vpl. 1 kann zwischen erster und zweiter Messung neben X_2 noch ein anderes Ereignis aufgetreten sein, das nach einer bewährten Störungshypothese die gleiche Auswirkung hat wie nach der zu prüfenden Hypothese die experimentelle Behandlung X_2 .

Bspw. könnte der Proband während der Untersuchung zur Prüfung unserer Hypothese WH_u sehr attraktive Gastarbeiterinnen kennengelernt haben, woraus nach der bewährten Störungshypothese (vgl. Amir, 1969) „angenehme persönliche Begegnungen mit Menschen aus einer bestimmten Gruppe verändern die Einstellung zur entsprechenden Gruppe zum Positiven hin“ die gleiche Einstellungsänderung folgt wie aus der WH_u ,

Entsprechend können beim Vpl. 2 sich die Situationen, in denen die beiden Messungen Y_{11} und Y_{22} erhoben wurden, über den Unterschied zwischen X_1 und X_2 hinaus noch hinsichtlich weiterer Merkmale unterscheiden. Liegt für

⁷⁾ Wir können diese Bezeichnung von Campbell & Stanley (1963) übernehmen, weil dieser Aspekt der experimentellen Validität auch nach unserer Betrachtungsweise insofern „intern“ ist, als (wie wir noch sehen werden) zur Kontrolle dieser Störfaktoren nur das gerade durchgeführte Experiment betrachtet werden muß, während z.B. eine hohe Variablenvalidität in aller Regel mehrere Untersuchungen erfordert.

mindestens eines dieser Merkmale eine bewährte Störungshypothese vor, die die gleiche Prognose erlaubt wie die WH_u , ist die interne Validität der Untersuchung nicht gewährleistet.

3.1.2 Variation personaler Merkmale

Zwischen den beiden Messungen beim Vpl. 1 kann es unabhängig vom Eintreten von X_2 zur Veränderung eines Merkmals des Probanden kommen, aus der sich die Einstellungsänderung vorhersagen läßt. Bspw. kann das allgemeine Aggressivitätsniveau des Probanden sinken.

Beim Vpl. 2 ist die interne Validität gestört, wenn sich die Probanden unter den beiden experimentellen Bedingungen hinsichtlich eines Merkmals unterscheiden und wenn dieser Unterschied nach einer bewährten Hypothese das auch aus der WH_u vorhergesagte empirische Ereignis zur Folge hat.

Beispiele für mögliche Störungshypothesen sind etwa: In der Bedingung X_2 liegt schon vor der Untersuchung eine positivere Einstellung vor; der Proband unter X_2 tendiert stärker dazu, sozial erwünschte Antworten zu geben.

3.2 Störfaktoren (IV) bei Meßwiederholung

Neben den bereits erwähnten, bei allen Designs möglichen Störfaktoren (IV), die in Unterschieden zwischen X_1 und X_2 hinsichtlich situationaler und personaler Merkmale bestehen, sind bei Designs mit wiederholter Messung der AV unter den verschiedenen experimentellen Bedingungen an den gleichen Personen (intraindividuelle Bedingungsvariation, s. Vpl. 1) weitere Gruppen von möglichen Störfaktoren (IV) zu beachten (nach Campbell & Stanley, 1963; Namboodiri, 1972; Greenwald, 1976; Cook & Campbell, 1979; s.a. Edwards, 1971, 225-226; Keppel, 1973, 395-400).

(1) Veränderungen beim Meßinstrument

Bei der zweiten Messung der AV kann sich das Meßinstrument verändert haben. Zum Beispiel können sich bei Erfassung der AV durch menschliche Beurteiler (Rater) deren Beurteilungskriterien mit der Zeit (oder zunehmender Beurteilungspraxis) verschieben.

(2) Ausscheiden von Versuchspersonen

Besonders wenn die Untersuchung sich über einen längeren Zeitraum hinzieht, können zwischen erster und zweiter Messung einige Versuchspersonen ausscheiden.

(3) Sensitivierung

Vpn können durch die erste Behandlung X_1 so sensitiviert werden, daß sie auf X_2 anders reagieren, als wenn sie allein X_2 ausgesetzt worden wären.

(4) übungs-, Ermüdungs- und Erinnerungseffekte

Bei der zweiten Messung kann der Wert der AV bspw. durch die größere Vertrautheit mit den Aufgaben, durch eine von der ersten Messung herrührende Ermüdung oder auch durch die Erinnerung an die Antworten bei der ersten Messung mitbestimmt werden.

(5) Vermengung von Behandlungswirkungen

Der Wert der AV kann bei der zweiten Messung außer durch X_2 auch durch die zurückliegende experimentelle Behandlung X_1 beeinflusst werden.

Die beiden zuletzt angesprochenen Störfaktoren (IV) werden oft als „Carry-over-“ oder „Übertragungseffekte“ zusammengefaßt. Zur Erläuterung und Unterscheidung diene ein Beispiel von Greenwald (1976, 318): Zur Untersuchung der Wirkung zweier Drogen auf die Reaktionszeit verabreicht ein Experimentator (E) den Vpn nacheinander beide Drogen und läßt anschließend jeweils einen Reaktionstest durchführen. Übungseffekte können dabei auftreten, wenn die im Test geforderten motorischen Tätigkeiten beim zweiten Mal besser beherrscht werden als beim ersten Mal. Eine Vermengung von Behandlungswirkungen liegt dagegen vor, wenn die Wirkung der zuerst verabreichten Droge bei der zweiten Einnahme noch nicht vollständig abgeklungen ist.

Zusammenfassend wollen wir die 5 Störfaktoren (IV) bei Meßwiederholung als *Sequenzeffekte* bezeichnen. Liegt ein Sequenzeffekt vor, kann über eine gut bewährte Störungshypothese die gleiche Prognose abgeleitet werden wie aus der zu prüfenden Kausalhypothese. In diesem Fall ist die interne Validität der Untersuchung gestört.

Diese Störfaktoren gelten nicht nur für den bisher betrachteten Fall einer zweimaligen Messung, sondern auch wenn eine Person s unter K experimentellen Bedingungen beobachtet wird:

(VPl. 3) $X_1 \quad Y_{s1} \quad X_2 \quad Y_{s2} \quad \dots \quad X_K \quad Y_{sK}$

Hier entstehen i.a. zudem Sequenzeffekte höherer Ordnung, d.h. Beeinflussungen von weiter auseinanderliegenden Behandlungen und Messungen.

Ein weiterer und möglicherweise gravierender Nachteil von Designs mit wiederholten Messungen soll hier ebenfalls erwähnt werden, obwohl er eher zu den Störungen der Variablenvalidität zu rechnen ist: Eine wiederholte Messung, d.h. eine Beobachtung einer Person unter allen experimentellen Bedingungen, erleichtert die Bildung von Hypothesen über das Ziel des Experi-

ments, und es besteht daher eine erhöhte Gefahr, daß die unabhängige Variable der zu prüfenden Kausalhypothese nur konfundiert mit diesen Vermutungen des Probanden realisiert werden kann (siehe auch Abschnitt 2.5).

Weshalb werden in der Psychologie trotz dieser zusätzlichen Gefahren für die Validität der Untersuchung wiederholte Messungen der Personen unter allen experimentellen Bedingungen durchgeführt?

Zum einen bieten sich Meßwiederholungen an, wenn die hier als störend klassifizierten Übungs- und Übertragungseffekte selbst Gegenstand der Hypothesenprüfung sind, wie dies etwa in Untersuchungen zum Lernfortschritt oder zu Adaptationsphänomenen der Fall ist (vgl. Greenwald, 1976, 316-318). Zum anderen werden Designs mit wiederholten Messungen wegen ihrer Ökonomie angewendet: Eine Versuchsperson „liefert“ mehr Daten, und die für Instruktion, Übungsphase usw. notwendige Zeit ist geringer, als wenn jedes Beobachtungsdatum an einer anderen Versuchsperson erhoben werden würde.

Auf weitere Vor-, aber auch Nachteile werden wir ausführlicher im Abschnitt 8.4.6 zu sprechen kommen.

3.3 Zur Kontrolle der Störfaktoren (IV) bei interindividueller Bedingungsvariation

Glücklicherweise brauchen wir bei der Durchführung von Untersuchungen die jeweilige Untersuchungssituation nicht nach den unendlich vielen potentiellen Störfaktoren (IV) abzuklopfen, um die möglichen Störungshypothesen einzeln auszuschalten. Vielmehr kann man diese Störfaktoren schon durch eine geeignete Planung der Untersuchung kontrollieren: durch Konstanthaltung bzw. Elimination und durch die Zufallsordnung von Probanden zu den einzelnen Behandlungsbedingungen (Randomisierung).

Wir können in diesem Artikel nur allgemein aufzeigen, durch welche Arten von Maßnahmen die Störfaktoren (IV) zu kontrollieren sind, und dies durch Beispiele illustrieren. Konkrete Handlungsanweisungen in bezug auf die Gestaltung der Untersuchungssituation (von der Auswahl der Stimuli über ihre Darbietungsform und die physikalische Ausschaltung anderer Variablen bis hin zu den Verhaltensvorschriften für den Versuchsleiter) können aus diesen allgemeinen Prinzipien für jedes Einzelexperiment abgeleitet werden. Diesbezügliche Hinweise bieten darüber hinaus die eher praktisch orientierten Einführungen in die experimentelle Psychologie (insbesondere Zimny, 1961; Selg, 1975; Carlsmith, Ellsworth & Aronson, 1976; außerdem z.B. Bugelski, 1960; Underwood, 1966; Heckhausen, 1969; Matheson, Bruce & Beauchamp, 1970; Sheridan, 1971; Arnold, 1972; Runkel & McGrath, 1972; Traxel, 1974; Wormser, 1974; Massaro, 1975; Fromkin & Streufert, 1976; Preiser, 1977; McGuigan, 1979, sowie Kazdin, 1980).

3.3.1 Konstanthaltung und Elimination

Wird im konkreten Fall vermutet, daß ein bestimmter Faktor zu einer Beeinträchtigung der internen Validität führt, so kann man diese mögliche Störung dadurch ausschalten, daß man den betreffenden Faktor konstant hält, und zwar sowohl *während* der gesamten Untersuchungszeit als auch - bei interindividueller Bedingungsvariation - *zwischen* den Behandlungsbedingungen (Zimny, 1961; Bredenkamp, 1969a).

Die Konstanthaltung als Technik zur Erhöhung der internen Validität ist besonders nutzbringend bei Faktoren einsetzbar, die zu den vom Experimentator hergestellten situativen Bedingungen gehören und die nicht die unabhängige Variable im engsten Sinne darstellen, also etwa bei der Instruktion, der Darbietungszeit und -form, dem Lärmpegel, dem Versuchsleiter usw. Variieren diese Faktoren systematisch zusammen mit den Modalitäten der UV, wird die interne Validität beeinträchtigt, weil die Variation der Werte auf der AV nicht mehr *ausschließlich* auf die *systematische Variation der UV* (und ggf. einen unsystematischen Fehleranteil) zurückgeführt werden kann. Man kann im übrigen diese Rahmenbedingungen durch Automatisierung des Versuchsablaufes recht problemlos konstant halten, also etwa durch elektronische Zeitsteuerung, Instruktion vom Tonband etc.

Die Elimination einer potentiellen Störbedingung kann als Spezialfall der Konstanthaltung angesehen werden. Sie ist etwa dann indiziert, wenn eine Konstanthaltung (etwa des Ausmaßes und der Qualität der sozialen Kontakte) nicht erreichbar ist.

Wird eine Untersuchung mit Personen in deren natürlicher Umgebung durchgeführt, gibt es eine Fülle von möglichen Störfaktoren. Ihre Konstanthaltung oder Elimination ist in den meisten Fällen unmöglich. Damit wäre aber auch eine strenge Prüfung der Hypothese nicht möglich. Von daher empfiehlt es sich, Untersuchungen in streng kontrollierten Situationen durchzuführen, also in einem gegen möglichst viele potentielle Störfaktoren abgeschirmten „Laboratorium“.

(Daß das *ausschließliche* Experimentieren im psychologischen Labor die strenge Prüfung einer Hypothese verhindern kann, wird später in Zusammenhang mit der Situationsvalidität erörtert.)

Was dabei unter Abschirmung konkret zu verstehen ist, hängt vom Bereich ab, aus dem die untersuchte Hypothese stammt. So sind in wahrnehmungspsychologischen Experimenten in erster Linie äußere Sinnesreize (plötzliche Geräusche, Beleuchtungsschwankungen usw.) konstantzuhalten oder zu eliminieren (z. B. durch einen schallisolierten Raum), in sozialpsychologischen Experimenten müssen eher die sozialen Interaktionen mit anderen Versuchsbeteiligten gleichförmig gestaltet werden und Kontakte mit Außenstehenden verhindert werden.

3.3.2 Randomisierung

Durch Konstanthaltung oder Elimination kann man nur solche potentiellen Störfaktoren kontrollieren, die bekannt und beobachtbar sind oder die durch so grobe Isolierungsmaßnahmen wie die Durchführung einer Untersuchung im Laboratorium ausgeschaltet werden können. Außerdem ist die Konstanthaltung oder Elimination eines Störfaktors relativ aufwendig, so daß solche Maßnahmen nur für eine relativ geringe Anzahl potentieller Störfaktoren durchführbar sind. Nun können sich die Personen, die bei einer interindividuellen Bedingungsvariation unter den verschiedenen Ausprägungen der UV untersucht werden, aber hinsichtlich einer großen und unbekannten Zahl individueller Merkmale unterscheiden.

Es gibt nur einen Weg, diese möglichen Störfaktoren (IV) zu kontrollieren: Zu jeder Modalität der UV werden *mehrere* „Untersuchungseinheiten“ (kurz: „Vpn“ für „Versuchspersonen“) beobachtet, und Vpn und Modalitäten werden *zufällig* einander zugeordnet. Dieses Vorgehen bezeichnet man als *Randomisierung*.

Nur wenn eine Randomisierung durchgeführt wird, können Kausalhypothesen streng geprüft werden: Nach einer Zufallszuordnung von Vpn und Bedingungen gibt es keine plausible Begründung für Störungshypothesen, nach denen sich die unter den verschiedenen Bedingungen beobachteten Vpn hinsichtlich eines ihrer Merkmale systematisch unterscheiden. Die Wahrscheinlichkeit, daß auf einer mit den Vpn zusammenhängenden Variablen zwischen den Behandlungsgruppen ein Unterschied von einer bestimmten Mindestgröße auftritt, ist nämlich um so geringer, je größer die Anzahl n der jedem Treatment zufällig zugeordneten Vpn ist. Zwischen den Behandlungsgruppen sind also auch nach einer Randomisierung systematische Unterschiede möglich; sie sind jedoch um so unwahrscheinlicher, je mehr Vpn verwendet werden.⁸⁾

Der Versuchsplan 2 wird mit Randomisierung zum weitverbreiteten Versuchsplan mit einer Experimental- und einer Kontrollgruppe, der i.a. als *Zufallsgruppenversuchsplan* bezeichnet wird:

(Vpl.	4)	R	X_1	Y_{i1}	
		R	X_2	Y_{i2}	(„R“ bedeutet „Randomisierung“)

Diesen Versuchsplan kann man auch in der folgenden Weise darstellen, die sich für die weiteren Erörterungen als nützlich erweisen wird (Vpl. 5):

⁸⁾ Diese zufällig möglichen Unterschiede werden bei der Versuchsauswertung über Signifikanztests berücksichtigt (vgl. Teil 7). Wie die nach bestimmten Kriterien optimale Anzahl der Vpn zu bestimmen ist, stellen wir im Teil 10 dar.

(Vpl. 5)

R – UV X	
X ₁	X ₂
Y ₁₁	Y ₁₂
Y ₂₁	Y ₂₂
⋮	⋮
Y _{n₁1}	Y _{n₂2}

Dabei bezeichnen die Y_{ik} -Werte die Größe der AV bei den verschiedenen Vpn. Aus dem Vpl. 5 wird ersichtlich, daß in der sog. „Kontrollgruppe“ (X_1) die Anzahl der Vpn gleich n_1 und in der „Experimentalgruppe“ (X_2) gleich n_2 ist, so daß also insgesamt $n_1 + n_2 = N$ Vpn untersucht werden.

Bei multifaktoriellen Experimenten werden die Vpn nach dem Prinzip der Randomisierung zufällig den möglichen *Kombinationen* von Modalitäten zugewiesen.

Beispiel: In einem Experiment zur Prüfung unserer WH_u könnte eine UV B vier Dissonanzstärken entsprechen (z.B. „fehlende“ (B_1), „leichte“ (B_2), „mittlere“ (B_3), „starke“ (B_4) Dissonanz), und diese könnten auf zwei verschiedene Arten erzeugt werden (z.B. „einstellungskonträre Information“ (A_1), „forced compliance durch Rollenspiel“ (A_2)). Den entstehenden acht Behandlungskombinationen werden dann zufällig jeweils n Vpn zugewiesen (siehe Vpl. 6).

		R – UV B (Stärke der Dissonanz)			
		B ₁ (keine)	B ₂ (leichte)	B ₃ (mittlere)	B ₄ (starke)
(Vpl. 6)					
R – UV A (Art der Disso-	A ₁	Y ₁₁₁	Y ₁₁₂	Y ₁₁₃	Y ₁₁₄
	(Informa-	Y ₂₁₁	Y ₂₁₂	Y ₂₁₃	Y ₂₁₄
	tion)	⋮	⋮	⋮	⋮
		Y _{n11}	Y _{n12}	Y _{n13}	Y _{n14}
nanz- erzeug- ung)	A ₂	Y ₁₂₁	Y ₁₂₂	Y ₁₂₃	Y ₁₂₄
	(Rollen-	Y ₂₂₁	Y ₂₂₂	Y ₂₂₃	Y ₂₂₄
	spiel)	⋮	⋮	⋮	⋮
		Y _{n21}	Y _{n22}	Y _{n23}	Y _{n24}

Zur Notation: Y_{ijk} bezeichnet bei solchen Versuchsplänen den Wert der AV Y mit $i = 1, \dots, s, s', \dots, n$; $j = 1, \dots, l, l', \dots, J$ und $k = 1, \dots, m, m', \dots, K$. Y_{slm} ist also der Wert der beliebig herausgegriffenen Vp s unter der beliebigen Bedingungskombination $A_l B_m$ oder (AB),..

Noch ein Wort zur Definition der Untersuchungseinheiten („Vpn“). In der Regel sind das tatsächlich einzelne Personen (oder Tiere). Werden jedoch Merkmale ganzer sozialer Gruppen als AV untersucht, stellen diese Gruppen die Untersuchungseinheiten dar. Häufig bezieht sich die AV zwar auf einzelne Personen, diese können aber nicht zufällig den Modalitäten zugewiesen werden, weil sie nur in vorgegebenen Gruppen (z.B. Schulklassen) untersucht werden können. In diesem Fall besteht die einzige Möglichkeit zur Sicherung der internen Validität darin, daß man möglichst viele dieser Gruppen zufällig den Treatmentkombinationen zuordnet - vgl. hierzu insbesondere Glass & Stanley (1970, 501-509) sowie Abschnitt 8.4.5.

Die praktische Durchführung einer Zufallszuordnung erfolgt am besten mit Hilfe von Zufallszahlen. Entsprechende Tabellen finden sich z. B. in Pearson & Hartley (1954, 1972), Fisher & Yates (1963), Edwards (1971), Kreyszig (1973) und Kriz (1978).

Natürlich muß man das Ergebnis einer Zufallszuordnung nicht vorbehaltlos akzeptieren, wenn sich eine offensichtlich „unwahrscheinliche“ Gruppeneinteilung ergeben hat, die die interne Validität stören könnte; dies ist bspw. der Fall, wenn alle männlichen Vpn der Experimental- und alle weiblichen Vpn der Kontrollgruppe zugewiesen worden sind. Nach Möglichkeit ist unter diesen Umständen die Randomisierung zu wiederholen (Mosteller, 1968, 115 f.).

Damit durch Randomisierung die Störfaktoren (IV) ausgeschaltet werden, müssen die dadurch entstandenen Gruppen natürlich bis zur Messung der AV am Ende der Untersuchung erhalten bleiben. Bei längerdauernden Untersuchungen ist das nicht immer der Fall, denn hier kommt es häufiger vor, daß Personen im Laufe der Untersuchung ausscheiden. Hängt diese sog. „experimentelle Mortalität“ mit der UV zusammen, beeinträchtigt sie die interne Validität (Cook & Campbell, 1979). Dies muß immer vermutet werden, wenn eine Behandlungsbedingung für die Probanden unangenehmer ist als eine andere.

Diese Störung der experimentellen Validität ist bei der Interpretation von Untersuchungsergebnissen stets zu berücksichtigen, und zwar unabhängig davon, welches der speziellen statistischen Auswertungsverfahren (siehe Abschn. 10.4.2) gewählt wird.

Bisher haben wir von der Kontrolle individueller Merkmalsunterschiede als Störfaktoren (IV) durch zufällige Zuordnung der Personen (oder anderer Untersuchungseinheiten) zu den Behandlungsgruppen gesprochen. Das Randomisierungsprinzip kann darüber hinaus auch zur Kontrolle anderer Störfaktoren (IV) dienen. Hier nur einige Beispiele für diese sehr wichtige Vorgehensweise:

- Kann die Untersuchung nicht für alle Probanden zum gleichen Zeitpunkt durchgeführt werden, können Unterschiede hinsichtlich der Tageszeit (vor

allem bei allgemeinspsychologischen Untersuchungen) und/oder des Datums der Untersuchung (vor allem in der Sozialpsychologie) Störfaktoren (IV) darstellen. Diese können dadurch ausgeschaltet werden, daß man die zur Verfügung stehenden Untersuchungstermine zufällig den experimentellen Bedingungen zuordnet. Dies erfolgt zum Beispiel, indem man jede ins Laboratorium kommende Versuchsperson zufällig einer Behandlungsgruppe zuweist.

- Wird die Untersuchung mit verschiedenen Versuchsleitern, an unterschiedlichen Orten, mit verschiedenen Apparaten usw. durchgeführt, so sind die Elemente der jeweils zur Verfügung stehenden Menge zufällig den Untersuchungseinheiten oder direkt den Behandlungsgruppen zuzuordnen, damit sie nicht zu Störfaktoren (IV) werden können.

3.3.3 Einführung eines Kontrollfaktors

Eine Verallgemeinerung der Konstanthaltung einer potentiellen Störbedingung ist ihre systematische Variation (Bredenkamp, 1969 a). Statt beispielsweise die Darbietungszeit der Reize für alle Probanden konstantzuhalten, kann man verschiedene Darbietungszeiten wählen und diese Variable als sog. „Kontrollfaktor“ (KF) mit Q Modalitäten in das Design einfügen - siehe Vpl. 7. Abgesehen vom hohen praktischen Aufwand lassen sich natürlich auch gleichzeitig mehrere Kontrollfaktoren einführen, auch zusätzlich zu mehreren bereits vorhandenen experimentellen Faktoren (unabhängige Variablen). Auf jeden Fall werden die Probanden wieder den möglichen Behandlungskombinationen zufällig zugewiesen. In jeder Zelle des Versuchsplanes 7 ergeben sich dann n Werte auf der AV Y , nämlich $Y_{lqm}, \dots, Y_{sqm}, \dots, Y_{nqm}$, wobei gilt: $q = 1, \dots, r, \dots, Q$ für den Kontrollfaktor KF und $k = 1, \dots, m, \dots, K$ für den experimentellen Faktor B.

(Vpl. 7)

		R – UV B (experimenteller Faktor)				
		B_1		B_m		B_K
R – UV KF (Kontroll- faktor)	KF_1					
	$\vdots$					
	KF_r			Y_{srm}		
	$\vdots$					
	KF_Q					

Der Vorteil dieses Vorgehens besteht darin, daß man prüfen kann, inwieweit der Einfluß der UV auf die AV vom jeweiligen Niveau des Kontrollfaktors abhängt (siehe dazu Abschn. 4.3 und 8.5). Zur Kontrolle der Störfaktoren der internen Validität ist die explizite Berücksichtigung des Kontrollfaktors allerdings nicht notwendig. Dazu genügt - um im Beispiel zu bleiben - die Konstanzhaltung der Darbietungszeit oder ihre zufällige Variation über alle Modalitäten der interessierenden UV.

Hier sei schon darauf hingewiesen, daß andere Arten von Kontrollfaktoren eingeführt werden können, um die Populations- oder Situationsvalidität oder die statistische Validität zu erhöhen (s. Abschn. 4.3 und 8.4.1).

3.4 Zur Kontrolle der Störfaktoren (IV) bei intraindividueller Bedingungsvariation (Meßwiederholung)

Die Ausführungen im Abschnitt 3.3.1 über die Konstanzhaltung als Methode zur Kontrolle von Störfaktoren (IV) können unmittelbar auf Experimente mit wiederholten Messungen übertragen werden. Wir wollen uns hier nur mit den für wiederholte Messungen spezifischen Störfaktoren (den sog. Sequenzeffekten) beschäftigen. Diese lassen sich nur kontrollieren, indem man die Abfolge der Behandlungen und Messungen variiert. Da bei einer intraindividuellen Variation dieser Reihenfolge (etwa nach der Spiegelbildmethode, siehe Selg, 1975) jede Person unter jeder Behandlungsbedingung mindestens zweimal beobachtet werden muß, können dabei leicht zusätzliche Sequenzeffekte entstehen. Deshalb kommt nur eine interindividuelle Variation der Reihenfolgen in Frage. Dafür gibt es drei Gruppen von Techniken (nach Zimny, 1961, 158-186; Selg, 1975, 50-55; Bredenkamp, 1969a):

(1) Zufällige Reihenfolgen

Man bestimmt für jede Untersuchungseinheit die Reihenfolge der Bedingungen zufallsmäßig. Dies geschieht am leichtesten über Tabellen mit Zufallsreihenfolgen einer gegebenen Menge von Zahlen (am ausführlichsten bei Moses & Oakford, 1963; außerdem bei Cochran & Cox, 1957; Fisher & Yates, 1963; Underwood & Shaughnessy, 1975; John & Quenouille, 1977). Dann kann man zwar davon ausgehen, daß man mit zunehmender Zahl der Untersuchungseinheiten der angestrebten Kontrolle der Sequenzeffekte beliebig nahe kommt; besonders bei wenigen Personen können sich aber doch noch beträchtliche Abweichungen ergeben.

(2) Vollständiges Ausbalancieren

Bei K Modalitäten der UV sind $K! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot K$ verschiedene Reihenfolgen möglich. Wird jeder dieser Reihenfolgen eine gleich große Zahl von Versuchspersonen zufällig zugeordnet, so sind alle Sequenzeffekte (auch die höheren) kontrolliert.

(3) Unvollständiges Ausbalancieren

Bei K Modalitäten werden K verschiedene Reihenfolgen so gewählt, daß jede Bedingung gleich häufig an jeder Stelle steht. Schreibt man die K Reihenfolgen untereinander, nennt man die entstehende Anordnung ein „lateinisches Quadrat“ (Lindquist, 1953). Ist K eine gerade Zahl, können die Reihenfolgen darüber hinaus so gewählt werden, daß jede Bedingung gleich häufig vor und hinter jeder anderen Bedingung steht. Ist $K+1$ eine Primzahl, lassen sich Quadrate konstruieren, die auch alle Sequenzeffekte höherer Ordnung kontrollieren (siehe dazu Cochran & Cox, 1957, 133-135; Cox, 1958, 272-274; Alimena, 1962; Edwards, 1971, 227-228; Raghavarao, 1971; Namboodiri, 1972; John & Quenouille, 1977, 196-214; zu den mathematischen Grundlagen siehe Dénes & Keedwell, 1974).

Allgemein können lateinische Quadrate stets verwendet werden, um ausgewogene Kombinationen dreier Faktoren mit der gleichen Anzahl von Modalitäten zu erreichen, wenn diese Ausprägungen aus irgendeinem Grunde nicht vollständig kombiniert („gekreuzt“) werden können (siehe Myers, 1972, 259-281; Winer, 1971, 685-751).

Welche der drei Strategien zur Kontrolle der Sequenzeffekte sollte man anwenden?

Das unvollständige Ausbalancieren stellt die am wenigsten befriedigende Kontrolltechnik dar. Allerdings gibt es zu ihr dann keine Alternative, wenn die Anzahl der Untersuchungseinheiten relativ klein ist oder wenn aus technischen Gründen nur relativ wenige verschiedene Reihenfolgen realisiert werden können.

Das vollständige Ausbalancieren ist nur dann möglich, wenn die Anzahl N aller Untersuchungseinheiten gleich $K!$ oder einem ganzzahligen Vielfachen davon ist. Da die Stichprobengröße N jedoch aufgrund ganz anderer Gesichtspunkte festgelegt werden sollte (siehe Teil 10), dürfte sie nur in Ausnahmefällen gleich $c \cdot K!$ sein.

Bei der Planung von Untersuchungen mit wiederholten Messungen sollte man daher in der Regel die Reihenfolgen der Behandlungen nach dem Zufallsprinzip bestimmen.

Die genannten Kontrolltechniken sind allerdings nur mit einer sehr wesentlichen Einschränkung wirksam: Durch Variation der Reihenfolgen können Sequenzeffekte nur dann ausgeglichen werden, wenn sie in dem Sinne symmetrisch sind, daß das Ausmaß der von jeder Behandlungsbedingung ausgehenden Sequenzeffekte gleich dem Ausmaß der von anderen auf sie entfallenden Effekte ist (vgl. Zimny, 1961, 165). In vielen Kontexten ist es aber z.B. durchaus wahrscheinlich, daß eine der Behandlungen besonders stark für andere Behandlungen sensibilisiert oder daß unter einer Behandlung der Übungseffekt besonders ausgeprägt ist. Diesen Übungseffekten versucht man oft durch eine ausgedehnte *prä-experimentelle Übungsphase* entgegenzuwirken, durch die erreicht werden soll, daß während des Experiments kaum noch Übungsfortschritte eintreten. Daneben wird die Gefahr von Übertragungseffekten natürlich im allgemeinen um so geringer, je größer die *zeitliche Distanz* der verschiedenen Behandlungen ist (etwa im Drogenbeispiel aus Abschnitt 3.2). Auch ist es zur Vermeidung von Übertragungs- und Sensibilisierungseffekten gerade bei Meßwiederholungen besonders empfehlenswert, den Vpn möglichst wenig Hinweise auf den Zweck der Untersuchung zu geben (vgl. Abschn. 2.5), z.B. indem man in unsystematischer Weise auch andere als die interessierenden unabhängigen Variablen verändert (vgl. Greenwald, 1976, 317).

3.5 Versuchspläne mit interindividueller Bedingungsvariation und Vortest

Möchte man auf die Vorteile einer interindividuellen Bedingungsvariation nicht verzichten, besteht aber aus irgendeinem Grund ein Interesse an den Werten der Probanden auf der AV vor der experimentellen Behandlung, kann man die Versuchspläne 4 und 2 wie in Vpl. 8 zu einem Kontrollgruppenexperiment mit Meßwiederholung kombinieren (vgl. Campbell & Stanley, 1963).

(Vpl. 8)	R	Y_{i0}	X_1	Y_{i1}
	R	Y_{i0}	X_2	Y_{i2}

Nachteilig bei diesem Versuchsplan ist, daß die Hypothese ausschließlich an Personen geprüft wird, bei denen in den sog. Vortests die AV bereits einmal vor der experimentellen Behandlung erhoben wurde. Dies setzt die Populationsvalidität des Experiments herab (siehe Abschn. 4.1). Falls man an Vortestwerten interessiert ist, sollte man deshalb den Versuchsplan 8 mit dem Versuchsplan 4 zum sog. „Solomon-Vier-Gruppen-Design“ (s. Vpl. 9) kombinieren (Solomon, 1949; Campbell & Stanley, 1963; Bredenkamp, 1969a; Huck & Sandler, 1973; Oliver & Berger, 1980). Bei diesem Versuchsplan läßt sich auch prüfen, inwieweit die Tatsache, daß ein Vortest durchgeführt worden ist, einen Einfluß auf die Beziehung zwischen der UV und der AV hat (vgl. Abschn. 8.5).

(Vpl. 9)	R	Y_{i0}	X_1	Y_{i11}
	R	Y_{i0}	X_2	Y_{i12}
	R		X_1	Y_{i21}
	R		X_2	Y_{i22}

3.6 Zur Definition des Experiments und anderer Untersuchungsmethoden

Wir hatten im Abschnitt 3.3 gesehen, daß ohne eine zufällige Zuordnung von Untersuchungseinheiten (Vpn) und experimentellen Bedingungen stets mit einer Störung der internen Validität gerechnet werden muß. Ohne Randomisierung können also Hypothesen über kausale Beziehungen zwischen Variablen nicht streng geprüft werden. Deshalb stellt die Randomisierung die wesentliche Bedingung dafür dar, daß eine Untersuchung als Experiment bezeichnet werden kann (Bredenkamp, 1969a, 1980; Carlsmith, Ellsworth & Aronson, 1976; s.a. unsere unten gegebene Definition).

Wird ein Experiment nicht unter künstlichen („Labor-“) Bedingungen durchgeführt, sondern in „natürlichen“ Situationen, spricht man von einem „*Feldexperiment*“ (s. Bredenkamp, 1969a; Redding, 1970; French, 1972; Kerlinger, 1979; Patry, 1979, 1982; Westmeyer, 1982).

Werden aus irgendeinem Grunde die Vpn *nicht* zufällig den Modalitäten der UV zugeordnet, sprechen Campbell & Stanley (1963) von einem „*Quasi-Experiment*“. Bis zu einem gewissen Grade kann man sich auch bei diesem dem Ideal einer strengen Prüfung annähern, und zwar indem man für möglichst viele relevante Störungshypothesen (IV) nachzuweisen versucht, daß für sie in der konkreten Untersuchungssituation keine entsprechenden Anfangsbedingungen vorliegen (Näheres siehe Cook & Campbell, 1976, 1979).

Alle Untersuchungen, die weder Experimente noch Quasi-Experimente sind, werden als „*Korrelationsstudien*“ bezeichnet. Sie können zwar zur Prüfung von Theorien noch insofern eingesetzt werden, als aus diesen ja stets auch Vorhersagen über korrelative Zusammenhänge ableitbar sind, diese Korrelationen lassen sich in der Regel aber auch durch bewährte Störungshypothesen vorhersagen, so daß keine strenge Prüfung möglich ist. Annäherungen an dieses Ziel sind allenfalls im Rahmen der Prüfung kausalanalytischer Modelle möglich (s. Namboodiri, Carter & Blalock, 1975; Hummell & Ziegler, 1976; Opp & Schmidt, 1976; Kenny, 1979).

Genaugenommen können wir eine Unterscheidung zwischen Experimenten, Quasi-Experimenten und Korrelationsstudien gar nicht für Untersuchungen als Ganzes treffen, sondern nur für einzelne unabhängige Variablen. Häufig finden sich nämlich innerhalb einer Untersuchung neben experimentellen Variablen im strengen Sinne, bei denen die Vpn zufällig den Modalitäten(kombinationen) zugeordnet werden, andere UVn, deren Ausprägungen bei den einzelnen Vpn gar nicht zufällig festgelegt werden können, sondern anderweitig „vorgegeben“ sind (z. B. Geschlecht, Intelligenzquotient, Art der psychischen

Störung).⁹⁾ Kausalaussagen sind in bezug auf derartige Variablen natürlich in aller Regel nicht möglich.

Auf der Grundlage dieser Überlegungen gelangen wir zu folgender Definition der *Experiments*:

Eine Untersuchung ist bezüglich einer unabhängigen Variablen X ein Experiment, wenn die gleichen Sachverhalte unter verschiedenen Bedingungen $X_1, \dots, X_K$ systematisch beobachtet werden und wenn Untersuchungseinheiten und Bedingungen einander zufällig zugeordnet werden bzw. wenn die Reihenfolgen zufällig bestimmt werden, in denen die Untersuchungseinheiten unter diesen Bedingungen beobachtet werden.

Wir können damit - abweichend von Wundt (1913) - auch Untersuchungen als Experimente auffassen, bei denen die Behandlungsbedingungen nicht völlig *willkürlich* hergestellt werden (können) und/oder bei denen die Bedingungen von anderen Instanzen als dem Experimentator *variiert* werden und/oder die nicht beliebig *wiederholbar* sind. Die entscheidende Frage für die Abgrenzung des Experiments von den anderen angesprochenen Untersuchungsmethoden ist vielmehr, ob eine Randomisierung erfolgt, denn diese stellt eine notwendige Voraussetzung für die strenge Prüfung einer Kausalhypothese dar.

4. Populations- und Situationsvalidität¹⁰⁾

4.1 Populationsvalidität (PV)

Unsere Beispielhypothese WH_u ist in mehrfacher Hinsicht typisch für psychologische Hypothesen. Uns interessieren hier zwei Punkte:

1. Die in der Hypothese vorkommenden theoretischen Begriffe bezeichnen Eigenschaften oder Merkmale, die einzelnen Personen zukommen.
2. Die in der Hypothese getroffene Aussage soll für alle Personen einer unendlichen Population gültig sein.

Beginnen wir bei der Erläuterung dieser beiden Punkte mit dem letzten! Die kognitive Dissonanztheorie enthält keine Einschränkungen ihres Geltungsbe-

⁹⁾ Dazu gehören auch Variablen, die sich auf etwas Vergangenes beziehen (z.B. frühkindliche Erfahrungen, Seminarbesuch im vergangenen Semester). Werden solche Variablen mit gegenwärtig beobachtbaren in Beziehung gesetzt, spricht man von „*ex-post-facto*-Untersuchungen“ (s. Meehl, 1970 und Kerlinger, 1979, 579-597).

¹⁰⁾ Wir vermeiden es, Populations- und Situationsvalidität zusammenfassend als „externe Validität“ zu bezeichnen (Campbell & Stanley, 1963; Bracht & Glass, 1968), weil gerade der Begriff der externen Validität in einer ausgesprochen induktivistischen Weise gebraucht wird (vgl. Gadenne 1976, und Abschn. 1.3).

reichs, d.h. wir müssen davon ausgehen, daß sie den Anspruch erhebt, für alle Personen gültig zu sein. Die aus dieser Theorie von unserem imaginären Forschungsteam abgeleitete Hypothese WH_u beansprucht zwar nicht unbedingt Allgemeingültigkeit für alle Menschen, wohl aber zumindest für alle Bürger eines bestimmten Staates. Dabei soll sie nicht nur für die endliche Menge der zu einem ganz bestimmten Zeitpunkt lebenden Bürger gelten, sondern auch für die mehr oder minder später existierenden, durch Geburt und Tod, Einwanderung und Fortzug veränderten Populationen. Der Geltungsbereich der Hypothese ist also eine unendliche Menge von Individuen. In der Terminologie der Logik bezeichnete man solche Aussagen als „*unbegrenzte Allsätze*“. Die meisten psychologischen Hypothesen und Gesetze entsprechen derartigen unbegrenzten Allsätzen. Dies gilt auch für solche Hypothesen, die deterministische Aussagen der Form „Wenn . . ., dann . . .“ vermeiden und statt dessen nur Wahrscheinlichkeitsaussagen beinhalten.

Ein Beispiel hierfür ist die folgende Hypothese WH_z : „Wenn zwischen der Einstellung einer Person zu einem bestimmten Objekt und einem anderen kognitiven Element eine Dissonanz besteht, dann verändert sich mit einer Wahrscheinlichkeit q die Einstellung dahingehend, daß diese Dissonanz vermindert wird“, oder kurz und formal ausgedrückt: „ $Prob(dE | D) = q$ “.

Die Formulierung von Hypothesen als unbegrenzte Allsätze hat einen entscheidenden Vorteil: Sollte die Hypothese in strengen Prüfversuchen nicht falsifiziert werden und somit als gut bewährte Gesetzesaussage gelten, kann sie zur *wissenschaftlichen Erklärung* einzelner Sachverhalte herangezogen werden. Nach Hempel & Oppenheim (1948; vgl. auch Hempel, 1965; Stegmüller, 1974a; Groeben & Westmeyer, 1975; Küttner, 1979; zur Kritik siehe Suppe, 1977b, 624-632) besteht nämlich die wissenschaftliche Erklärung in der logischen Ableitung der Aussage über den zu erklärenden Sachverhalt (z.B. „Person s hat ihre Einstellung zum Positiven hin verändert.“) aus einem gut bewährten Gesetz (im Beispiel ist dies unsere WH_u) und einigen Anfangsbedingungen (im Beispiel u.a. „Bei Person s ist Dissonanz erzeugt worden.“). Eine derartige Erklärung kann aber nur dann adäquat sein, wenn das darin verwendete Gesetz eine Aussage ist, die sich auf unendlich viele Anwendungsfälle (d.h. in der Regel: Personen) bezieht (Stegmüller, 1974a). Andernfalls könnte unser exemplarischer Sachverhalt nämlich aus einer Aussage wie „Alle Personen in diesem Raum haben ihre Einstellung zum Positiven hin geändert“ abgeleitet werden, womit sicherlich keine befriedigende wissenschaftliche Erklärung gegeben wäre.

Als *psychologische* Hypothesen oder Gesetze enthalten diese unbegrenzten Allsätze nun Aussagen über einzelne Individuen (im Gegensatz etwa zu soziologischen Hypothesen, in denen zumeist Merkmale von Personengruppen betrachtet werden). Trotzdem können diese Hypothesen nicht für jede einzelne Person falsifiziert werden, da Störungen der internen Validität in aller Regel

nicht mit genügender Sicherheit ausgeschlossen werden können, wenn wir einzelne Personen betrachten, sondern nur wenn mehrere Personen zufällig unterschiedlichen experimentellen Bedingungen zugeordnet werden (s. Abschn. 3.3.2). Die möglichst strenge Prüfung einer Hypothese muß also normalerweise an einer *Gruppe* von Personen erfolgen.

Gehen wir jetzt wieder vom Kriterium einer strengen Prüfung der wissenschaftlichen Hypothese aus! Danach muß eine Untersuchung so beschaffen sein, daß die Wahrscheinlichkeit eines hypothesenkonträren Ergebnisses hoch ist, wenn die Hypothese falsch ist. Im Zusammenhang mit der untersuchten Personengruppe kann das Ziel einer strengen Prüfung in zwei Fällen gefährdet sein: Erstens können keine hypothesenkonträren Ergebnisse eintreten, wenn die in der Untersuchung verwendete Personengruppe gar keine Stichprobe aus der Population darstellt, für die die Hypothese gelten soll (vgl. Bredenkamp, 1979). Wir wollen dies als Störung der *Populationsvalidität erster Art* (PV_1) bezeichnen. Sie liegt z.B. dann vor, wenn eine humanpsychologische Hypothese an Tieren oder eine für Schizophrene Gültigkeit beanspruchende Aussage an „Normalen“ überprüft wird.“) Zweitens wird eine Untersuchung in aller Regel nur an einer bestimmten Untermenge der Personen (oder anderer Untersuchungseinheiten) durchgeführt, für die die Hypothese gelten soll. Eine Störung der *Populationsvalidität zweiter Art* (PV_2) liegt dann vor, falls es eine bewährte Störungshypothese HS, gibt, nach der die Hypothese WH_u zwar für die Subpopulation gilt, aus der die untersuchten Personen stammen, für andere Untermengen des beanspruchten Geltungsbereiches aber nicht. Dementsprechend ist jede Variable, die Unterpopulationen, für die die zu prüfende Hypothese gilt, von solchen trennt, für die sie nicht gilt, ein möglicher Störfaktor der Populationsvalidität zweiter Art (vgl. Gadenne, 1976).

Beispiele: Die Gültigkeit einer wissenschaftlichen Hypothese kann beschränkt sein auf Subpopulationen von Freiwilligen (Rosenthal & Rosnow, 1969b), von Probanden, die mit psychologischen Experimenten vertraut sind, von Probanden, die motiviert genug sind, ein langdauerndes Experiment durchzuhalten, von Personen mit hoher Bewertungsangst (Weber & Cook, 1972), von Arbeitern eines bestimmten Betriebes, von Personen, die durch Messung der AV vor der Behandlung sensibilisiert worden sind (Lana, 1969) usw. In all diesen Fällen ist die PV gestört, wenn eine Untersuchung nur mit Personen aus der jeweiligen Unterpopulation durchgeführt wird.

Zusammenfassend können wir also formulieren: Eine Untersuchung zur Prüfung einer Kausalhypothese ist um so strenger, je weniger sie sich auf Probanden aus bestimmten Untermengen der Population beschränkt, für die die Hypothese gelten soll.

¹¹⁾ Oft sind Untersuchungen an Personen aus dem Gültigkeitsbereich der Hypothese aus technischen oder ethischen Gründen gar nicht durchführbar - man denke nur an die Erprobung neuer Medikamente. Werden in einem solchen Fall Tiere als Modelle menschlicher Organismen benutzt, so kann in diesen Untersuchungen die höchste zu diesem Zeitpunkt realisierbare Populationsvalidität gegeben sein.

Die untersuchten Probanden brauchen aber keine Zufallsstichprobe aus irgendeiner (Sub-)Population zu sein, denn jede Kausalhypothese kann (weil sie sich - wie begründet - notwendigerweise auf eine *unendliche* Population von Anwendungsfällen bezieht) an *beliebigen* Teilgruppen dieser Population falsifiziert werden (Holzkamp, 1964; Bredenkamp, 1972; Gadenne, 1976; s.a. Abschn. 8.2.6).

4.2 Situationsvalidität (SV)

Bei der Formulierung wissenschaftlicher Hypothesen und Theorien sollte stets angegeben werden, für welche raum-zeitlichen Bedingungskonstellationen (kurz: für welche Situationen) sie gültig sein sollen (Westmeyer, 1982). Dieser für eine Theorie beanspruchte Geltungsbereich wird meist aus einer sehr großen, in der Regel sogar unendlichen Menge von Situationen bestehen. Im Zusammenhang mit der bei einer empirischen Prüfung einer Theorie vorliegenden Situation können wir zwei Fälle unterscheiden, in denen das Ziel einer strengen Prüfung gefährdet ist: erstens wenn die Hypothese oder Theorie gar nicht für die Situation gelten soll, in der der Prüfversuch durchgeführt wird (Störung der SV erster Art); zweitens wenn die Vermutung berechtigt ist, daß die zu prüfende Hypothese zwar unter den gegebenen Umständen erfüllt ist, nicht aber unter (einigen oder vielen) anderen Kombinationen von situationalen Bedingungen, für die ihre Gültigkeit ebenfalls beansprucht wird (Störung der SV zweiter Art). Entsprechend der Definition von Störfaktoren der Populationsvalidität wollen wir eine Variable als Störfaktor der Situationsvalidität - Störfaktor (SV) - bezeichnen, wenn sie nach einer bewährten Störungshypothese Situationen definiert, in denen die zu prüfende Hypothese nicht gilt.

Störfaktoren der Situationsvalidität können von sehr unterschiedlicher Art sein. So mag eine Hypothese WH_v nur gültig sein, wenn den Probanden bewußt ist, daß sie an einer Untersuchung teilnehmen (Hawthorne- oder Placebo-Effekt, vgl. auch den von Klauer (1973, 558) beschriebenen „Novitätseffekt“ in der pädagogischen Forschung), wenn die abhängige Variable auch bereits vor der Behandlung gemessen wurde (Vortest-Sensitivierung), wenn die AV sofort nach der Behandlung erfaßt wird (mangelndes Überdauern des Effekts), wenn die Prüfung durch Versuchsleiter mit bestimmten Eigenschaften und Verhaltensweisen, in Räumen eines Forschungsinstituts, unter den Bedingungen eines Laborexperiments, unter Verwendung ganz bestimmter Apparaturen, bei einer bestimmten Dauer der experimentellen Behandlung, zu einer bestimmten Tageszeit, zeitlich kurz nach einer bestimmten Fernsehsendung usw. stattfindet (vgl. Bracht & Glass, 1968). Solche und viele ähnliche Faktoren können die Validität einer Untersuchung einschränken.

Daraus ergibt sich allgemein: Eine Untersuchung zur Prüfung einer Kausalhypothese ist um so strenger, je weniger sie sich auf ganz bestimmte zeitliche, räumliche und situationale Umstände aus dem Geltungsbereich der Hypothese beschränkt.

Umfaßt der Geltungsbereich einer Hypothese nicht nur Laborsituationen, müssen im Interesse einer möglichst strengen Prüfung dieser Hypothese also auch Feldexperimente durchgeführt werden (vgl. Abschn. 3.6 und 5).

4.3 Zur Kontrolle der Störfaktoren (PV und SV)

Liegt eine Vermutung darüber vor, daß eine Variable V ein Störfaktor der Populations- oder Situationsvalidität ist, kann man dies stets prüfen, indem man die Untersuchung unter mindestens zwei Ausprägungen dieser Variablen durchführt, diese Variable also als Kontrollfaktor einführt (vgl. Abschn. 3.3.3). Betrachten wir dazu ein Beispiel! In einem Experiment zur Prüfung unserer Hypothese WH_u wird die experimentelle Bedingung „keine Dissonanz“ (A_1) der Bedingung „Dissonanz“ (A_2) gegenübergestellt, und es interessiert uns, inwieweit die Zugehörigkeit zu einer bestimmten Berufsgruppe (Faktor B) die Resultate beeinflußt. Faktor B umfasse drei Modalitäten: „Student“ (B_1), „Beamter“ (B_2) und „freier Beruf“ (B_3). Es seien nun unter den 6 Bedingungskombinationen folgende Mittelwerte der AV aufgetreten (je höher die Werte sind, desto positiver ist die Einstellung):

Tabelle 4.1:

	B_1	B_2	B_3
A_1	0	5	10
A_2	2	4	6

Offensichtlich ist die Vorhersage der Hypothese WH_u bei Studenten eingetreten, jedoch nicht bei Beamten und Freiberuflern. Nach unserer Definition ist B also ein Störfaktor der Populationsvalidität. In statistischer Terminologie spricht man davon, daß eine „*disordinale Interaktion*“ bezüglich des Faktors A vorliegt (siehe im einzelnen Abschn. 8.5).

Dies gilt allgemein: *Eine Variable ist genau dann ein Störfaktor der Populations- oder Situationsvalidität, wenn bezüglich des Behandlungsfaktors eine disordinale Interaktion besteht.* Wird eine derartige Wechselwirkung wiederholt festgestellt, sollte zumindest der Geltungsbereich der Hypothese eingeschränkt werden. Wie man feststellen kann, ob zwei Variablen interagieren und ob ggf. diese Interaktion disordinal ist, werden wir im Abschnitt 8.5 noch ausführlicher besprechen.

Innerhalb einer Untersuchung wird sich stets nur eine sehr kleine Zahl von potentiellen Störfaktoren der SV oder PV auf diese Weise untersuchen lassen. Zur Klärung der Frage, in welchem Ausmaß die Strenge der Prüfung einer

Hypothese durch solche Faktoren eingeschränkt ist, muß man sich deshalb zum einen auf entsprechende Ergebnisse aus Untersuchungen zur Prüfung anderer Hypothesen stützen, zum anderen kann eine einigermaßen fundierte Beurteilung der Gültigkeit einer Kausalhypothese nur erfolgen, nachdem sie durch mehrere Untersuchungen in verschiedenen Subpopulationen und Situationen überprüft wurde. Da aus den unendlichen Mengen der Personen und Situationen, für die die Hypothese gelten soll, keine repräsentativen Stichproben zu ziehen sind, wird das Ziel einer strengen Prüfung der Hypothese durch ein Forschungsprogramm aus mehreren Untersuchungen am besten erreicht, wenn aus dem Geltungsbereich Personengruppen und Situationen bewußt so ausgewählt werden, daß sie möglichst verschiedenartig sind (vgl. Cook & Campbell, 1979, 74-80; s.a. Dipboye & Flanagan, 1979) und daß in anderen Kontexten bewährte oder (aus theoretischen Überlegungen) besonders plausible Störungshypothesen direkt geprüft werden können.

5. Beziehungen zwischen den Validitätsarten

Zwischen interner Validität auf der einen und Populations- und Situationsvalidität auf der anderen Seite besteht im allgemeinen eine gegenläufige Beziehung. Zunächst einmal sind Störungen der SV und PV häufig bedingt durch Maßnahmen zur Kontrolle möglicher Störfaktoren der internen Validität. Ein Beispiel möge dies verdeutlichen: Die IV ist gefährdet, wenn jede Experimentalgruppe einen anderen Versuchsleiter hat. Diese Gefahr kann z.B. dadurch ausgeräumt werden, daß man für alle Gruppen das Experiment durch den gleichen VL durchführen läßt, die Variable „Versuchsleiter“ also konstant hält (vgl. Abschn. 3.3.1). Dann besteht aber die Gefahr, daß das erhaltene Ergebnis spezifisch ist für Experimente unter Leitung dieser Person und daß sich bei anderen VL mit anderen Eigenschaften andere Resultate eingestellt hätten. Durch Konstanzhaltung von Bedingungen kann also die Situationsvalidität eingeschränkt werden. Umgekehrt wird durch eine Erhöhung der SV oft die interne Validität herabgesetzt. So ist unter natürlich auftretenden situationalen Bedingungen die Konstanzhaltung potentieller Störfaktoren (IV) oder die Randomisierung der Untersuchungseinheiten, -zeitpunkte und -räume meist wesentlich schwieriger. Oft sind diese Maßnahmen gar nicht durchzuführen, so daß allenfalls Quasi-Experimente möglich sind. Wegen dieser Gegenläufigkeit der betrachteten Validitätsaspekte sollte eine Entscheidung über eine Hypothese, deren Geltungsbereich auch „natürliche“ Situationen umfaßt, erst aufgrund der Ergebnisse von mehreren Untersuchungen Verschiedenster Art getroffen werden.

Läßt sich eine Hypothese jedoch tatsächlich nur in Laborsituationen mit einer ausreichenden internen Validität überprüfen, muß ihr Geltungsbereich strenggenommen auf diese Situationen beschränkt werden. Aussagen über andere

Situationen lassen sich dann zwar logisch nicht ableiten, es spricht jedoch nichts dagegen, aus Theorien und Hypothesen, die sich in strengen Prüfungen *gut bewährt* haben, sog. *technologische Prognosen* zu gewinnen (Gadenne, 1976; Brocke, 1978, 1979). Mit Hilfe von Gesetzmäßigkeiten, die (u.U. nur im Labor) relativ gut gesichert worden sind, werden also Voraussagen über Variablenzusammenhänge in komplexen praktischen Situationen wie z.B. einer Psychotherapie gemacht. Anders als bei der isolierten Bedingungsvariation des intern validen Experiments sind dort aber ganz sicher auch andere Variablen wirksam.

Im Zusammenhang mit der Variablenvalidität hatten wir im Abschnitt 2.5 erwähnt, daß mit der UV des Experiments Aspekte des Versuchsleiterverhaltens und Hypothesen der Versuchspersonen über die von ihnen erwarteten Verhaltensweisen konfundiert sein können und daß dadurch fälschliche Falsifikationen und Bestätigungen entstehen können. Wir erinnern hier an diesen Punkt, weil die Höhe der Gefahr einer solchen Konfundierung abhängig sein kann von der Situation, in der das Experiment durchgeführt wird. Je „unauffälliger“ und „natürlicher“ die Herstellung der Behandlungsbedingungen und die Registrierung der abhängigen Variablen erfolgt, je weniger Hinweise auf Hypothesen und Erwartungen des Experimentators die Situation den Probanden gibt, desto geringer ist die Gefahr derartiger Störungen der Variablenvalidität.

Von den möglichen Maßnahmen zur Vermeidung solcher Forschungsartefakte seien nur einige der wichtigeren genannt:

- Die Täuschung der Probanden über ihre Rolle als Versuchspersonen und/oder über die Erwartungen des Versuchsleiters;
- Durchführung von *Feldexperimenten* und Simulationsstudien (Rollen-spiele);
- Erfassung der abhängigen Variablen durch *nicht-reaktive* Methoden (vgl. auch Abschnitt 2.5; zu weiteren Einzelheiten siehe etwa Bredenkamp, 1969a; Summers, 1970; Bungard & Lück, 1974; Webb et al., 1975; Carlsmith, Ellsworth & Aronson, 1976).

Wir wollen nun die Beziehung zwischen der Variablenvalidität und den anderen Arten der experimentellen Validität genauer betrachten. Diese Beziehung hängt ganz entscheidend von der zu prüfenden Hypothese ab (Gadenne, 1976, 99-103). Bezieht sich die Hypothese auf ideale, künstliche Situationen (z.B. weil die Annahme der Konstanz aller anderen Bedingungen notwendig ist), ist es meist relativ einfach, gleichzeitig alle Teilaspekte der experimentellen Validität in befriedigendem Ausmaß sicherzustellen. Anders ist es bei Theorien und Hypothesen, deren Begriffe nur in natürlichen Situationen realisiert werden können, wie z. B. bei vielen entwicklungspsychologischen Theorien oder bei Hypothesen über das Verhalten in sozialen Gruppen. Sollen Prüfungen

solcher Hypothesen eine hohe Variablenvalidität haben, müssen sie in Situationen durchgeführt werden, die häufig nur eine ungenügende Kontrolle der Störfaktoren der internen Validität erlauben. Hier stellen also Variablenvalidität und interne Validität konträre, wenn nicht sogar unvereinbare Ziele dar. Typisch für die Psychologie sind jedoch Hypothesen und Theorien, die sowohl im Labor wie in natürlich auftretenden Situationen Gültigkeit beanspruchen. Diese können zwar durchaus im Labor-Experiment einer Prüfung unterzogen werden; wie schon im Zusammenhang mit der Situationsvalidität erwähnt, erfordert eine strenge Prüfung dieser Hypothesen aber auch ihre Überprüfung in anderen Umgebungen, also z.B. die Durchführung von Feld- oder Quasi-Experimenten.

Insgesamt gesehen sind Variablenvalidität und interne Validität von grundlegender Bedeutung als Populations- und Situationsvalidität. Bei unzureichender Variablenvalidität oder interner Validität ist ein Experiment auf keinen Fall eine strenge Prüfung der betrachteten Kausalhypothese, bei mangelhafter Populations- oder Situationsvalidität dagegen ist immerhin noch eine strenge Prüfung der in ihrer Gültigkeit auf die entsprechende Population oder Situation eingeschränkten Hypothese möglich.

6. Statistische Validität

Wir haben im Abschnitt 4.1 gesehen, daß psychologische Hypothesen typischerweise als Kausalaussagen formuliert sind, die für jede einzelne Person (oder Gruppe oder Schulklassse etc.) einer bestimmten Population gelten sollen. Eine empirische Überprüfung derartiger Hypothesen wäre dann am strengsten, wenn man ihre Gültigkeit für jede einzelne Person (oder Gruppe etc.) untersuchen könnte. Nun gibt es zwar durchaus Methoden zur Einzelfallanalyse (siehe etwa Hersen & Barlow, 1976; Kratochwill, 1978; Petermann & Hehl, 1979; Petermann, 1981), doch ist für eine hinreichend strenge Prüfung von Kausalhypothesen in aller Regel die Zusammenfassung mehrerer untersuchter Personen (oder allgemein: mehrerer experimenteller Untersuchungseinheiten) notwendig, und zwar in erster Linie aus den folgenden zwei Gründen, die sich aus unseren bisherigen Ausführungen ergeben (vgl. auch Gadenne, 1976, 88f.):

- (1) Die interne Validität einer Untersuchung kann nur in ausreichendem Maße gesichert werden, wenn mehrere experimentelle Einheiten den Ausprägungen der unabhängigen Variablen zufällig zugeordnet werden (vgl. Abschnitt 3.3.2; bei Meßwiederholungen entsprechend - siehe Abschnitt 3.4).
- (2) Ordnen wir einem theoretischen Begriff eine empirische Variable zu, so ist diese i. a. keine eindeutige oder fehlerfreie Entsprechung des theoretischen Begriffs.

schen Begriffs. Dies wird am Beispiel leicht deutlich: Das Ergebnis eines Probanden in einem Intelligenztest kann man nicht als fehlerfreies Maß für die Ausprägung des theoretischen Begriffs „Intelligenz“ ansehen, weil das verfügbare „Hintergrundwissen“ die Information enthält, daß es z.B. auch durch den augenblicklichen Ermüdungszustand beeinflusst wird. Hat daher eine Person ein unterdurchschnittliches Ergebnis in diesem Test erzielt, so werden wir sie trotzdem nicht sofort als „unterdurchschnittlich intelligent“ bezeichnen, wenn wir z.B. wissen, daß dieser Test nach einer anstrengenden Nachtschicht durchgeführt worden ist. Zur Notwendigkeit derartiger „Ausweichklauseln“ und ihren Konsequenzen für die Zuordnung von theoretischen und empirischen Begriffen siehe insbesondere Herrmann (1973).

Beobachtet man - um beim einfachsten Fall zu bleiben - jeweils n Versuchspersonen unter zwei experimentellen Bedingungen, wird man praktisch hinsichtlich aller in Psychologie und anderen Sozial- und Biowissenschaften interessierenden abhängigen Variablen feststellen, daß auch bei Personen aus der gleichen Behandlungsbedingung verschiedene Werte auftreten. Diese Variation entsteht sowohl durch Unterschiede zwischen den Probanden hinsichtlich der verschiedensten Persönlichkeitsmerkmale (z. B. Alter, Geschlecht, Intelligenz) und hinsichtlich der vorausgegangenen Erfahrungen (z.B. mit ähnlichen Aufgaben) als auch durch - nicht ganz vermeidbare - Unterschiede in den Durchführungsbedingungen bei den verschiedenen Probanden sowie durch geringe Zuverlässigkeit der Meßinstrumente und/oder Beobachter.

Betrachten wir wieder unsere Beispielhypothese WH_u , aus der ja die empirisch prüfbare Aussage folgt, daß unter der (Dissonanz-)Bedingung X_2 die Einstellungsvariable Y größer ist als unter der (Kontroll-)Bedingung X_1 . Auch angesichts der stets anzutreffenden Fehlervarianz der Variablen Y wäre die Prüfung dieser Hypothese völlig unproblematisch, wenn bei intraindividuellem Bedingungsvariation alle Individuen unter X_2 einen höheren Y -Wert hätten als unter X_1 bzw. wenn bei interindividueller Bedingungsvariation die Y -Werte aller unter X_2 beobachteten Personen größer wären als bei den Personen unter X_1 . Leider ist das in der Psychologie (und in verwandten Wissenschaften) praktisch nie der Fall, vielmehr überlappen sich die Verteilungen der Werte auf der abhängigen Variablen unter den verschiedenen Bedingungen mehr oder minder stark. Deshalb lassen sich hier Kausalhypothesen nur dadurch überprüfen, daß man aus ihnen *statistische Hypothesen* ableitet.¹²⁾ Wir wollen dies etwas näher erläutern (nach Bredenkamp, 1972): Aus WH_u folgt beispielsweise die statistische Hypothese, daß der Mittelwert der Verteilung der Y -Werte unter

¹²⁾ Der Gedanke einer implikativen Beziehung zwischen wissenschaftlichen statistischen Hypothesen geht u.W. auf Meehl (1967) zurück und wurde von Bredenkamp (1969b, 1972, 1979, 1980) aufgegriffen und weiterentwickelt.

der Bedingung X_2 größer ist als der entsprechende Mittelwert unter X_1 , oder kurz $H_1: \mu_2 > \mu_1$ bzw. anders geschrieben $\mu_2 - \mu_1 > 0$.¹³⁾ Logisch gesehen besteht zwischen der wissenschaftlichen Hypothese WH_u und der statistischen Hypothese H_1 eine Implikationsbeziehung: $WH_u \rightarrow H_1$.¹⁴⁾ Diese Beziehung eröffnet die Möglichkeit, wissenschaftliche Hypothesen über die Prüfung von statistischen Hypothesen zu falsifizieren. Denn kann man feststellen, daß die statistische Hypothese falsch ist, muß bei Gültigkeit der Implikationsbeziehung auch die übergeordnete wissenschaftliche Hypothese falsch sein. Zeigt sich dagegen, daß die statistische Hypothese richtig ist, hat sich auch die wissenschaftliche Hypothese in dem betr. Experiment bewähren können.

Diese Verfahrensweise scheitert in der dargestellten einfachen Form jedoch daran, daß statistische Hypothesen selbst grundsätzlich weder beweisbar noch widerlegbar sind, weil sie Aussagen über unendliche Populationen enthalten (Bredenkamp, 1972). Man kann lediglich Kriterien festlegen, die Angaben über die Bedingungen enthalten, unter denen man sich für die Annahme oder für die Ablehnung einer statistischen Hypothese entscheiden will (Gadenne, 1976, 86).

Auch wenn man aufgrund der erhobenen Daten zu der Entscheidung gelangt, daß die H_1 falsch ist, kann man wegen der Unsicherheit über den Wahrheitswert von statistischen Hypothesen daraus nie logisch die Falschheit der übergeordneten Kausalhypothese WH_u ableiten. Man kann lediglich eine weitere methodologische Regel akzeptieren und befolgen, nach der die Entscheidung für die Falschheit von H_1 notwendige Voraussetzung für eine Falsifikation der Kausalhypothese WH_u ist und nach der die Kausalhypothese solange als bewährt gelten kann, wie man die von ihr implizierte statistische Hypothese für richtig hält.

Unabhängig von der Wahl des Verfahrens, mittels dessen man zu Entscheidungen über das Zutreffen der statistischen Hypothesen gelangen will, besteht stets die Möglichkeit, daß falsche Entscheidungen getroffen werden - wir kommen auf dieses Problem im Abschnitt 7.3 zurück.

¹³⁾ Zur genaueren Definition des Begriffs der statistischen Hypothese siehe Abschn. 7.1. Betont sei auch, daß aus der WH_u noch andere statistische Hypothesen ableitbar sind, etwa über Mediane („ $Md_2 > Md_1$ “) oder über die stochastische Größe von Zufallsvariablen („ $F_1(Y) > F_2(Y)$ für alle Werte der AV Y “).

¹⁴⁾ Eine Implikationsbeziehung zwischen wissenschaftlicher und statistischer Hypothese kann auch dann hergestellt werden, wenn die wissenschaftliche Hypothese eine Wahrscheinlichkeitsaussage der Form $\text{Prob}(dE \mid D) = q$ ist (vgl. Abschnitt 4.1). Auch wenn man Theorien als axiomatisch formulierte abstrakte Strukturen betrachtet (s. Groeben & Westmeyer, 1975, 71-75; Westmeyer, 1981), kann die Angemessenheit einer solchen Theorie für einen bestimmten empirischen Sachverhalt geprüft werden, indem man aus den Axiomen statistische Hypothesen ableitet und diese überprüft (vgl. Westermann, 1980).

Das Kriterium der strengen Prüfung von Kausalhypothesen erfordert nun, vereinfacht ausgedrückt (vgl. Abschnitt 1.3), daß die Wahrscheinlichkeit hypothesenkonträrer Ergebnisse hoch ist, wenn die Kausalhypothese falsch ist, und daß sie niedrig ist, wenn die Hypothese „wahr“ ist.

Daraus folgt: Die Prüfung einer Kausalhypothese über eine statistische Hypothese ist (alle anderen Bedingungen als konstant vorausgesetzt) um so strenger, je geringer die Wahrscheinlichkeiten für Fehler bei der Entscheidung über die abgeleitete statistische Hypothese sind.

Diejenigen Faktoren, die diese Fehlerwahrscheinlichkeiten erhöhen, wollen wir als „Störfaktoren der statistischen Validität (StatV)“ bezeichnen (in Anlehnung an Cook & Campbell, 1976). Wodurch die statistische Validität einer Untersuchung konkret beeinträchtigt werden kann und wie man diese Störungen zu vermeiden versuchen kann, werden wir im Teil 8 erläutern. Grundlage dafür ist der Teil 7, in dem die für unser Thema wichtigsten Aspekte des statistischen Hypothesentestens dargestellt sind. Die Teile 9 und 10 sind dann zwei Teilaspekten der statistischen Validität gewidmet: den Maßen für die Größe „experimenteller Effekte“ und der nach bestimmten Kriterien optimalen Wahl der Zahl N von Untersuchungseinheiten. Die Überlegungen in den Teilen 7 bis 10 führen uns zu Empfehlungen, wie man bei der Planung und Durchführung von statistischen Hypothesenprüfungen vorgehen sollte und wie man auf dieser Basis zu Entscheidungen über Falsifikation oder Beibehaltung wissenschaftlicher Hypothesen gelangen kann. Diese Planungs- und Entscheidungsstrategie ist im abschließenden Teil 11 dargestellt.

7. Eine Strategie zur Entscheidung zwischen statistischen Hypothesen: Der Signifikanztest

7.1 Überblick über verschiedene alternative Strategien

Die Entscheidung über das Zutreffen von statistischen Hypothesen, die i.a. über Verteilungen von Zufallsvariablen oder deren Parameter formuliert werden, kann grundsätzlich auf mehrere Arten erfolgen, die wir im folgenden nach Menges (1972) und Barnett (1973) kurz und summarisch aufführen.¹⁵⁾

1. Das Verfahren nach Bayes setzt weitgehende Kenntnisse über die o.a. Verteilungen voraus, die häufig nur unter Verwendung subjektiver Annahmen verfügbar gemacht werden können - siehe zu diesem Modell im einzelnen Edwards, Lindman & Savage (1963), Menges (1972, 272-274), Philips (1974), Rüppell (1977), Rützel (1979, 1980) sowie Kleiter (1981).

¹⁵⁾ Wir greifen die hier nicht erläuterten Begriffe im folgenden auf.

2. Das von Fisher (e. g. 1935) entwickelte Fiduzial-Modell ist anwendbar ohne Vorkenntnisse über die Verteilungen von Zufallsvariablen, dabei „aber an strenge Bedingungen geknüpft“ (Menges, 1972, 275). Es ist wenig gebräuchlich, hat sich aber als das bislang einzig exakte Modell zur Lösung des sog. „Behrens-Fisher-Problems“ erwiesen (vgl. dazu Abschnitt 8.2.4.1). Eine kurze Darstellung des Fiduzial-Modells findet man etwa bei Menges (1972, 275-279).
3. Das ebenfalls auf Fisher zurückgehende Likelihood-Modell erlaubt die Bestimmung der (relativen) „Plausibilität“ (Likelihood) von Parametern - siehe dazu im einzelnen u.a. Edwards (1972), Menges (1972, 279-283) und Witte (1980, 34-40).
4. Das Konfidenz-Modell ermöglicht die Angabe von Zufallsintervallen, die mit vorgegebener Wahrscheinlichkeit einen bestimmten Parameter umschließen oder überdecken. Die weite Verbreitung dieses Modells dokumentiert sich auch darin, daß es in fast jedem Lehrbuch der Statistik vglw. ausführlich erörtert wird; siehe zur Einführung etwa Menges (1972, 282-286), Hays (1977, Kap. 9) oder Witte (1980, 23-26).
5. Vor dem Hintergrund dieses Konfidenz-Konzepts wurden die „klassischen“ Theorien des Signifikanztests entwickelt, nämlich das Modell von Fisher (e.g. 1925, 1956) und das Modell von Neyman & Pearson (1933a, b, 1936, 1938; Neyman, 1952).
6. Das Modell der sequentiellen Tests von Wald (1947) geht über diese Ansätze hinaus und kann als Spezialfall einer allgemeinen Theorie der statistischen Entscheidungsfunktionen aufgefaßt werden (Wald, 1950); zur Darstellung siehe etwa Weber (1967, 395-482) und Wetherill (1975).

Vergleiche bzgl. der Leistungsfähigkeit einiger der aufgeführten Ansätze findet man etwa bei Bredenkamp (1972, 134-150), Barnett (1973) und Witte (1977, 1980); zur Verbindung verschiedener Verfahren zu einem mehrstufigen Inferenzmodell siehe Witte (1980).

Üblicherweise wird von Psychologen, Pädagogen und Sozialwissenschaftlern der Signifikanztest zur Entscheidung über das Zutreffen von statistischen Hypothesen herangezogen - vgl. zu dieser Behauptung u.a. Sterling (1959), Cohen (1962), Edgington (1964 a, 1974), Smart (1964), Bozarth & Roberts (1972), Bredenkamp (1972, 9) und Witte (1980, 17). Aus diesem Grunde beschränken wir unsere Ausführungen auf dieses Verfahren.

Dabei ist anzumerken, daß es „den“ Signifikanztest nicht gibt. Vielmehr bestehen in Theorie und Praxis divergierende Auffassungen über das Rationale, das der Gruppe von Verfahren zugrunde liegt, die unter der Bezeichnung „Signifikanztest“ zusammengefaßt werden, und über ihre adäquate Anwendung.

Die theoretische Kontroverse entstand in den dreißiger Jahren, als Jerzy Neyman und Egon S. Pearson (1933 a, b, 1936, 1938) das hauptsächlich in den

zwanziger Jahren von Sir Ronald A. Fisher (1925, 1956) entwickelte Modell der statistischen Hypothesenprüfung durch ihre Theorie der Fehler 1. und 2. Art zu erweitern suchten und Fisher diese Erweiterung strikt zurückwies. Wir können aus Raumgründen auf diese Kontroverse nicht näher eingehen, sondern müssen den Interessenten auf zusammenfassende Darstellungen verweisen, etwa auf den Reader von Morrison & Henkel (1970) sowie die Veröffentlichungen von Spielman (1974, 1978), Carlson (1976), Chase & Tucker (1976) sowie Witte (1980).

Ebensowenig können wir die Grundgedanken dieser Testtheorien hier vorstellen; die in den nachfolgenden Abschnitten enthaltenen skizzenhaften Darstellungen einiger für das Verständnis späterer Ausführungen notwendiger Aspekte des Signifikanztests setzen daher die Kenntnis der Grundgedanken voraus. Sofern der Leser mit diesen nicht vertraut ist, mag er sich in einem der zahlreichen Lehrbücher informieren, etwa in den eher mathematisch oder wissenschaftstheoretisch orientierten Einführungen von Kendall & Stuart (1961), Hacking (1965, 1976), Fisz (1970), Menges (1972), Stegmüller (1973a, b) oder Mood, Graybill & Boes (1974); eher für Psychologen und Sozialwissenschaftler verfaßte Texte stammen von Hays (1963, 1977), Stilson (1966), Kriz (1978), Leiser (1978), Bortz (1979), Haagen & Seifert (1979) sowie Witte (1980) - diese Zusammenstellung erhebt selbstverständlich keinen Anspruch auf Vollständigkeit.

Gegen bestimmte Aspekte der Anwendung von Signifikanztests im Bereich der Psychologie und verwandten Wissenschaften sind wiederholt und zu Recht teils schwerwiegende Bedenken ins Feld geführt worden (vgl. zu einigen Einzelheiten etwa die Abschnitte 7.4.1 und 7.4.2). Diese haben Autoren wie etwa Kleiter (1969), Harnatt (1975, 1979), Derrick (1976), Guttman (1977) und Carver (1978) zu der Empfehlung veranlaßt, auf Signifikanztests weitestgehend oder völlig zu verzichten.

Wir sind zwar mit Tukey (1977) der Auffassung, daß etwa einer exploratorischen Datenanalyse (EDA) im Forschungsprozeß eine wesentlich gewichtigere Rolle beigemessen werden muß, als es derzeit geschieht, meinen aber andererseits mit Bredenkamp (1972, u.a. 134-150; 1980) und Witte (1980), daß ein „richtig angewendeter“ Signifikanztest trotz aller immanenten Schwächen (derzeit noch) unverzichtbar ist.

Wenden wir uns nun einigen Aspekten der Prüfung von statistischen Hypothesen mittels Signifikanztests zu!

7.2 Kurzer Abriß einiger Charakteristika von Signifikanztests

Mit dem Terminus „statistische Hypothese“ belegt man jede Annahme über die Wahrscheinlichkeits- oder die (theoretische) Populationsverteilung einer beobachtbaren Zufallsvariablen oder aber über einen oder mehrere Parameter (vgl. Kendall & Stuart, 1961, 161; Hays, 1977, 335; Haagen & Seifert, 1979, 167-175).

Der Terminus „Parameter“ bezeichnet dabei eine endliche Menge von Konstanten, durch die eine Populationsverteilung näher spezifiziert wird - siehe zu Einzelheiten etwa Menges (1972, 214-216); zum hier nicht erläuterten Terminus „Zufallsvariable“ siehe u.a. Stilson (1966, 121-125), Menges (1972, 139-145) oder Hays (1977, 110-133).

Als Beispiel für eine Populationsverteilung sei die Gruppe der sog. „Normalverteilungen“ genannt, die durch eine einzige Verteilungsfunktion charakterisiert ist (siehe etwa Hays, 1977, 297); die einzelnen Vertreter dieser Gruppe unterscheiden sich darin, welche numerischen Werte die beiden Parameter dieser Verteilung annehmen, nämlich die „Varianz σ^2 “ und der „Mittel-“ oder „Erwartungswert μ “ der entsprechenden Zufallsvariablen Y .

Könnte man die theoretisch möglichen Realisierungen der Zufallsvariablen Y , also die Gesamtmenge der Daten, im konkreten Fall vollständig untersuchen, wäre eine sichere Entscheidung über das Zutreffen oder Nicht-Zutreffen der statistischen Hypothese möglich.

Der empirisch arbeitende Wissenschaftler ist jedoch fast stets darauf angewiesen, weniger sichere und weniger definitive Aussagen aufgrund einer nur beschränkten Menge von Daten zu treffen. Diese Untermenge der Population von Daten heißt „Stichprobe (von Daten)“; sie umfaßt eine mögliche Auswahl aus der Grundgesamtheit.

Man spricht von einer „Zufallsstichprobe“, wenn alle möglichen Auswahlvarianten der gleichen Größe n die gleiche Wahrscheinlichkeit haben, als Stichprobe zu fungieren (vgl. Fisz, 1970, 394; Hays, 1977, 72); über andere Stichprobentechniken informieren etwa Schwarz (1975), Cochran (1972, 1977) sowie Rasch et al. (1978). Innerhalb des Modells „Signifikanztest“ kommt dem Konzept der zufälligen Stichproben für die wesentlichen Ableitungen eine zentrale Bedeutung zu.

Die Verteilung der Daten in der Stichprobe wird durch die sog. „(empirische) Häufigkeitsverteilung“ angegeben. Wie die Verteilung der Rohwerte in der Population durch bestimmte Parameter gekennzeichnet ist, die die in den Daten enthaltene Information zusammenfassen und reduzieren, so sind auch die Stichprobendaten durch sog. „Stichprobenfunktionen“ oder „Statistiken“ beschreibbar, die die Häufigkeitsverteilung charakterisieren.

Statistiken stellen häufig das Stichprobenäquivalent bestimmter Parameter dar, ohne mit diesen wertmäßig übereinstimmen zu müssen. Denn durch zufallsbedingte Schwankungen weicht die Stichprobenzusammensetzung meist von der Populationszusammensetzung ab. Erhebt man nun sehr viele Stichproben der gleichen Größe n und berechnet die empirischen Werte einer Statistik, dann dokumentieren sich diese Zufallsschwankungen darin, daß die resultierende „Stichproben-Kennwerte- oder Prüf- oder Stichproben-Verteilung“ der betreffenden Statistik eine Varianz größer als Null aufweist.¹⁶⁾

Um die auf i. a. sehr verschiedenen Experimenten beruhenden Werte für eine Statistik problemlos auf nur eine Prüfverteilung beziehen zu können, werden die Statistiken linear transformiert, und man nennt derartige „transformierte Statistiken“ „Prüf- oder Test-Statistiken“. Häufig besteht diese Transformation in einer Standardisierung, wie es etwa beim t -Wert der Fall ist, der eine(n) standardisierte(n) Mittelwert(sdifferenz) darstellt; andere bekannte Teststatistiken sind etwa χ^2 , F oder H .

Um von den Stichproben- auf die Populationsdaten oder -kennwerte schließen zu können, muß eine Verbindung zwischen der beobachtbaren Stichprobe und der theoretischen Grundgesamtheit konstruiert werden. Dies geschieht durch die sog. „Schätzfunktionen“, d.h. Statistiken oder Stichprobenfunktionen „mit bestimmten Eigenschaften“ - vgl. dazu Menges (1972, 294f.). Auf diese Schätzfunktionen gehen wir nicht ein; der interessierte Leser sei auf die zu Beginn dieses Teils genannte einführende Literatur verwiesen.

Wie erwähnt, weichen die Stichproben(kenn)werte in aller Regel von den Populations(kenn)werten ab - diese Abweichung oder Verzerrung wird als „unsystematisch“ (oder „zufällig“) interpretiert und als eine Folge des Stichprobenprozesses angesehen (vgl. Abschnitt 8.2). Der Signifikanztest dient nun der Beantwortung der Frage, ob ein konkretes empirisches Resultat, d.h. etwa eine empirische Realisation der Zufallsvariable „Teststatistik“, als nur durch den Zufall zustande gekommen „erklärt“ werden kann oder ob die Annahme einer systematischen „Verursachung“ angemessener erscheint. Der Erklärung „durch Zufall“ wird dann der Vorzug gegeben, wenn die Realisation der Test- oder Prüfstatistik in eine Klasse von möglichen Realisationen fällt, der eine insgesamt hohe Wahrscheinlichkeit zukommt; letztere wird unter Verwendung der o. g. Prüfverteilung bestimmt.

Für die mathematische Ableitung der Prüfverteilung muß die Gültigkeit einer statistischen Hypothese und (sehr oft auch) eine bestimmte Populationssituation (vgl. Abschnitt 7.5.2) angenommen werden. Findet man dann, daß der

¹⁶⁾ Je größer diese Varianz der Stichprobenverteilung ist, desto geringer ist unter sonst gleichen Bedingungen die „Präzision“ eines Experiments - vgl. dazu insbesondere Abschnitt 8.4. - Mit dem sog. „Standardfehler (einer Teststatistik)“ ist die Wurzel aus dieser Varianz gemeint.

empirische Wert für eine Teststatistik in einen Bereich fällt, der unter diesen Annahmen insgesamt wenig wahrscheinlich ist, schließt man daraus, daß die als richtig unterstellte Hypothese nicht zutrifft, daß m.a. W. die untersuchte Stichprobe einer Population entstammt, deren Parameter andere Werte aufweisen bzw. deren Verteilung anders ist, als in der statistischen Hypothese angenommen.

Die statistische Hypothese, auf deren Zutreffen die Ableitung der Stichprobenverteilung einer beliebigen (Test-)Statistik beruht, heißt üblicherweise die „Null-Hypothese“ (H_0).

Eine H_0 wird in der Regel von inhaltlichen Aussagen der Form „Es besteht *kein* Unterschied zwischen verschiedenen experimentellen Behandlungen und damit deren Auswirkungen auf bestimmte Parameter bzw. Verteilungen“ impliziert.

Wir haben jedoch im Teil 6 gesehen, daß unsere Kausalhypothese WH_u zu anderen Vorhersagen führte, nämlich etwa der, daß „Unterschiede zwischen verschiedenen Kennwerten bestehen“. Man nennt die von einer derartigen inhaltlichen Aussage implizierte statistische Hypothese die „Alternativhypothese“ (H_1).

Entschließt man sich aufgrund der experimentellen Daten, die H_0 zurückzuweisen, nimmt man üblicherweise gleichzeitig die H_1 an, weil die beiden statistischen Hypothesen sich gegenseitig ausschließen und darüber hinaus in der Regel so formuliert sind, daß sie alle logisch möglichen Annahmen über die Verteilungen oder Parameter ausschöpfen. Wenn wir bei unserem Beispiel aus Teil 6 bleiben, wäre die dort über Populationsmittelwerte μ_k formulierte Hypothese H_1 : $\mu_2 - \mu_1 > 0$ um die H_0 : $\mu_2 - \mu_1 \leq 0$ zu erweitern. Entsprechend gehört zu einer H_0 : $\mu_1 - \mu_2 = 0$ die H_1 : $\mu_1 - \mu_2 \neq 0$ usf.

Man entscheidet sich also üblicherweise für die von einer WH implizierte H_1 , indem man die ihr entgegengesetzte H_0 zurückweist. Für diese Ablehnung der H_0 wird folgendes Kriterium festgesetzt: „Weise die H_0 dann zurück, wenn Du ein Ergebnis gefunden hast, das zu einer Ergebnisklasse gehört, die unter der Annahme der Gültigkeit der H_0 eine Gesamtwahrscheinlichkeit aufweist, die kleiner oder gleich einem bestimmten geringen Wert α ist.“ Man nennt α das „Signifikanzniveau“.

Durch dieses Signifikanzniveau α (auch „Umfang eines Tests“ oder „Irrtumswahrscheinlichkeit“ genannt) wird die Menge aller möglichen Realisationen einer Teststatistik in zwei unterschiedlich große Teilmengen zerlegt. Die größere dieser Teilmengen heißt „Annahmehereich“ und enthält diejenigen $100(1 - \alpha)\%$ aller möglichen Werte der Prüfstatistik, die als mit der H_0 vereinbar angesehen werden, deren Varianz also als zufallsbedingt angesehen werden kann. Die kleinere Teilmenge dagegen heißt „Ablehnungs- oder Rejektionsbe-

reich“, weil sie alle diejenigen Realisationen der Teststatistik enthält, die zwar unter Gültigkeit der H_0 durchaus auftreten können (!), deren Auftretenswahrscheinlichkeit insgesamt jedoch so gering ist, daß sie „auf lange Sicht“ nur (sehr) selten, nämlich in $100 \cdot \alpha$ % aller Fälle, zu erwarten sind, falls H_0 zutrifft.

Man bezeichnet ein Ergebnis in diesem Ablehnungsbereich als „statistisch signifikant“ und meint damit, daß es „bedeutsam“ von einem aufgrund des Zufalls zu erwartenden Resultat abweicht.

Wenn nun eine empirische Realisation der Zufallsvariablen „Teststatistik“ in diesen Ablehnungsbereich fällt, d.h. unter Gültigkeit der H_0 (sehr) unwahrscheinlich ist, dann kann dieses „seltene Ereignis“ grundsätzlich auf zwei verschiedene Arten *interpretiert* werden:

1. Die H_0 ist richtig, und es ist „rein zufällig“ ein Wert aus einer Ereignisklasse mit einer insgesamt nur sehr geringen Wahrscheinlichkeit aufgetreten.
2. Das Resultat ist sehr unwahrscheinlich, falls H_0 zutrifft. Es wird daher davon ausgegangen, daß H_0 falsch und H_1 richtig ist.

Zwischen beiden Möglichkeiten läßt sich keine objektiv richtige Entscheidung treffen; es entspricht daher lediglich einer auf Fisher (1925) zurückgehenden Konvention, wenn im Falle eines unter H_0 unwahrscheinlichen (wiewohl möglichen!) Resultates entschieden wird, die H_0 als falsch anzusehen - vgl. zu dieser Verfahrensweise auch Neyman (1952, 43).

An dieser Stelle wird deutlich, daß der Signifikanztest als eine spezielle Strategie aufgefaßt werden kann, mittels derer man zu Entscheidungen über das Zutreffen oder Nicht-Zutreffen von statistischen Hypothesen gelangen *kann*. Die Entscheidungskriterien sind dabei nicht „test-immanent“ oder sonstwie zwingend vorgegeben, sondern beruhen vorwiegend auf Vereinbarungen.

7.3 Mögliche Fehler beim statistischen Testen

7.3.1 Fehler unter Gültigkeit der Null-Hypothese (Fehler 1. Art)

Obwohl unter Gültigkeit der H_0 extreme Resultate (sehr) unwahrscheinlich sind, besteht doch grundsätzlich die Möglichkeit ihres Auftretens. Tritt nun in empirischen Daten ein Wert der Statistik im Ablehnungsbereich unter H_0 auf, obwohl H_0 tatsächlich zutrifft, wird man diese irrtümlich als nicht zutreffend zurückweisen. Mit dieser (Fehl-)Entscheidung begeht man einen sog. „Fehler 1. Art“, auch α -Fehler genannt.

Die bedingte Wahrscheinlichkeit für einen derartigen Fehler beträgt α , denn die Wahrscheinlichkeit der Klasse von sehr extremen Resultaten, die im Ab-

lehnungsbereich von H_0 liegen, ist bei stetigen Verteilungen genau und bei diskreten Verteilungen höchstens gleich dem Signifikanzniveau α :

$$p(\text{Zurückweisung von } H_0 \mid H_0 \text{ trifft zu}) = \alpha.$$

Diese Fehlerwahrscheinlichkeit kann durch die Wahl sehr kleiner Werte für α , etwa $\alpha = 0,05$ oder $\alpha = 0,01$ etc., vom E selbst kontrolliert und gering gehalten werden. Mit der Fixierung des numerischen Wertes für α expliziert der E seine maximale Bereitschaft, einen Fehler 1. Art zu begehen. Mit der Wahl des Wertes für α wird der Annahme- und Ablehnungsbereich für die betr. Stichprobenverteilung ebenfalls festgelegt.

Diese Festlegung hat in jedem Fall vor der Datenerhebung zu erfolgen, da durch sie der Signifikanztest „erst . . . eindeutig bestimmt ist“ (Haagen & Seifert, 1979, 203).

Das Komplement zur Wahrscheinlichkeit α ist $1 - \alpha$; dabei handelt es sich um die bedingte Wahrscheinlichkeit, eine zutreffende H_0 auch als richtig auszuweisen:

$$p(\text{Annahme von } H_0 \mid H_0 \text{ trifft zu}) = 1 - \alpha.$$

Die Gesamtwahrscheinlichkeit aller numerischen Werte der Statistik im Annahmebereich unter H_0 beträgt $1 - \alpha$; d.h. in $100 \cdot (1 - \alpha)$ % aller Fälle wird erwartet, daß der Signifikanztest eine in der Population wahre Null-Hypothese „entdeckt“.

Nun ist hier fortwährend von „der“ Null-Hypothese die Rede, obwohl wir im vorangegangenen Abschnitt 7.2 gesehen haben, daß unter „der“ Null-Hypothese auch eine ganze Klasse von Werten für den in Frage stehenden Parameter spezifiziert werden kann; dies kommt etwa in den folgenden (zulässigen) Hypothesenformulierungen zum Ausdruck:

$$H_0: \mu_2 - \mu_1 \leq 0 \text{ oder: } H_0: \mu \leq \mu_0 = 100 \text{ (siehe vorn).}$$

Man nennt derartige Hypothesen, die anstelle nur eines einzigen Wertes einen ganzen Wertebereich für den Parameter spezifizieren, „zusammengesetzt“ (auch „unspezifisch“ und „unexakt“). Dagegen heißt eine Hypothese, die nur genau einen Wert für den Parameter zuläßt, „einfach“ (auch „exakt“ und „spezifisch“). Die folgenden beiden Hypothesen sind bspw. exakt:

$$H_0: \mu_2 - \mu_1 = 0 \text{ oder: } H_0: \mu = \mu_0 = 100.$$

Näheres zu diesen Unterscheidungen findet man etwa bei Stilson (1966, 385f.), Weber (1967, 167f.) und Bortz (1979, 148f.); zu beachten ist hierbei, daß der Sprachgebrauch recht uneinheitlich ist, wie die Arbeiten von Hays (1977, 335f.) und Haagen & Seifert (1979, 206-225) belegen.

Die Frage stellt sich also, welche von diesen theoretisch beliebig vielen Null-Hypothesen dem statistischen Test unterzogen wird. Grundsätzlich wird auch dann, wenn H_0 eine zusammengesetzte Hypothese ist, nur eine einzige davon statistisch geprüft, und zwar diejenige einfache H_0 , die dem Parameterbereich am nächsten liegt, der von allen unter der Bezeichnung „ H_1 “ zusammengefaßten Hypothesen abgedeckt wird - vgl. dazu im einzelnen Neyman & Pearson (1933 a, b) und Haagen & Seifert (1979, 206-225).

In den o. gen. Beispielen werden also nur die folgenden Hypothesen geprüft:

$$H_0: \mu_2 - \mu_1 = 0 \text{ und } H_0: \mu = \mu_0 = 100.$$

Alle anderen unter der Klasse „ H_0 “ zusammengefaßten Hypothesen sind dann zu verwerfen, wenn auch die einfache Hypothese zurückgewiesen werden kann, die den relativ extremsten Wert für den interessierenden Parameter angibt.

Die festgelegte Wahrscheinlichkeit für einen Fehler 1. Art (a-Fehler) ist dabei maximal für die getestete einfache H_0 und geringer für alle anderen Elemente der Klasse „ H_0 “ (Bredenkamp, 1972, 21 f.; Hays, 1977, 363).¹⁷⁾

Die vorstehenden Ausführungen über den a-Fehler beruhen auf der Voraussetzung, daß in der Population eine einfache statistische Null-Hypothese zutrifft. Bei den folgenden Darstellungen wollen wir davon ausgehen, daß diese Annahme nicht richtig ist, sondern daß die unter der Alternativhypothese formulierten Annahmen der Populationssituation entsprechen.

7.3.2 Fehler unter Gültigkeit der Alternativhypothese (Fehler 2. Art)

Die statistische Alternativhypothese H_1 wird in der Regel von einer wissenschaftlichen Hypothese wie unserer Beispielhypothese WH_u impliziert, die wir u.a. auf den Seiten 68 und 69 bereits behandelt haben.

Mit der H_1 werden Annahmen über die Werte des interessierenden Parameters spezifiziert, die grundsätzlich von den unter H_0 formulierten abweichen. Trifft daher eine der unter der Klasse „ H_1 “ angegebenen einfachen Alternativhypothesen zu, resultieren (fast) stets Stichprobenverteilungen der benutzten Prüfstatistik, die von den unter H_0 geltenden Verteilungen quantitativ abweichen.

Der Grad dieser Abweichung von den tabellierten und zur Signifikanzbeurteilung benutzten sog. „zentralen“ Prüfverteilungen kann meist durch einen sog.

¹⁷⁾ Wenn im folgenden weiter von der Null-Hypothese die Rede ist, soll darunter ggf. die gesamte Hypothesenklasse „ H_0 “ verstanden werden, ohne daß explizit zwischen „einfachen“ und „zusammengesetzten“ Hypothesen unterschieden wird.

„Nicht-Zentralitätsparameter“ angegeben werden (vgl. dazu im einzelnen Abschnitt 9). Man bezeichnet die unter Gültigkeit der H_1 resultierenden Stichprobenverteilungen als die „nicht-zentralen Verteilungen“ der betreffenden Teststatistik, die für χ^2 und F erstmals Fisher (1928) näher erörtert hat.

Auch unter Gültigkeit der H_1 sind die Realisationen der Teststatistik bestimmten Zufallsschwankungen unterworfen, die im konkreten Einzelfall so ausgeprägt sein können, daß der Wert der Statistik im Annahmehereich unter H_0 liegt, die dann angenommen wird.

Da aber nach unserer Voraussetzung H_1 richtig ist, trifft der E damit eine falsche Entscheidung, er begeht einen „Fehler 2. Art“. Die Wahrscheinlichkeit für einen derartigen Fehler 2. Art, d.h. für eine irrtümliche Entscheidung für H_0 , wird formal mit β bezeichnet; entsprechend findet sich daher auch oft die Bezeichnung „ β -Fehler“.

Das Komplement zu β , also $1 - \beta$, heißt „Teststärke“ („Macht“, „Mächtigkeit“, „Güte“, „Trennschärfe“ oder „Power“) eines Tests und gibt die Wahrscheinlichkeit einer richtigen Entscheidung für H_1 an:

$$\begin{aligned} p(\text{Zurückweisung der } H_0 \text{ und Annahme der } H_1 \mid H_1 \text{ trifft zu}) &= \beta \\ p(\text{Annahme der } H_1 \mid H_1, \text{ trifft zu}) &= 1 - \beta \end{aligned}$$

Wir haben im vorigen Abschnitt festgestellt, daß mit jeder der unter „ H_0 “ zusammengefaßten einfachen Hypothesen eine „eigene“ (Gesamt-)Irrtumswahrscheinlichkeit α verbunden ist, deren Maximalwert üblicherweise vom E durch eine Festsetzung gering gehalten wird. In der gleichen Weise ist auch mit jeder der unter „ H_1 “ zusammengefaßten einfachen Hypothesen eine eigene Fehlerwahrscheinlichkeit β verknüpft (vgl. zu Einzelheiten etwa Menges, 1972, 328-335, oder Hays, 1977, 357-373). Die Wahrscheinlichkeit β für eine irrtümliche Annahme der H_0 läßt sich daher nur dann bestimmen, wenn angegeben werden kann, welche der zahlreichen einfachen Hypothesen unter „ H_1 “ zutrifft, wenn also m.a. W. bekannt ist, wie sehr der interessierende Parameter (resp. die Verteilung) von dem unter H_0 spezifizierten Wert abweicht. Wir bezeichnen diese Abweichungen im folgenden allgemein als „experimentelle Effekte“ (EE), deren Auftreten trotz dieser Benennung nicht an die Art der Untersuchung gebunden ist. Auf diese experimentellen Effekte gehen wir im Teil 9 ausführlich ein.

Bestimmt man unter Verwendung der nicht-zentralen Verteilungen die Werte für die Teststärke $1 - \beta$ in Abhängigkeit vom Signifikanzniveau α , dem experimentellen Effekt und der Stichprobengröße N für einen bestimmten Test, erhält man die sog. „Teststärke-“, „Güte-“ oder „Trennschärfefunktion“ des betr. Signifikanztests - zu Einzelheiten siehe etwa Menges (a. a. O.) und Hays (a. a. O.).

Diese Gütefunktionen sind für einige der gebräuchlichen (sog. „parametrischen“) Teststatistiken wie χ^2 , F und t seit geraumer Zeit in tabellarischer Form als „Power Tables“ sowie in graphischer Form als „Power Charts“ verfügbar, allerdings vornehmlich in mathematischen und statistischen Fachzeitschriften (vgl. Abschnitt 10.2 und 10.3). Entsprechend wurden sie von Psychologen kaum rezipiert; und auch in den meisten Lehrbüchern der Statistik für Psychologen und Sozialwissenschaftler wird das Konzept „Teststärke“ - wenn überhaupt - nur kurz behandelt.“) Es kann daher kaum verwundern, wenn in der üblichen Forschungspraxis die Wahrscheinlichkeit für einen Fehler 2. Art (β) nicht kontrolliert oder bestimmt wird.

Diese Unterlassung kann bei der Entscheidung zwischen den statistischen Hypothesen zu schwerwiegenden Konsequenzen führen.

1. Möglichkeit: Der E entscheidet sich aufgrund des Tests für H_0 . Diese Entscheidung kann deshalb zustande gekommen sein, weil H_0 tatsächlich zutrifft, oder aber weil zwar die H_1 richtig ist, dem E jedoch ein Fehler 2. Art unterlaufen ist. Da die Wahrscheinlichkeit β für diesen Fehler meist nicht bekannt ist, kann auch nichts über die Güte der Entscheidung ausgesagt werden. Zwar beträgt unter der Annahme der Richtigkeit der H_0 die Wahrscheinlichkeit eines Resultates im Annahmebereich $1 - \alpha$, aber unter der Richtigkeit der H_1 kann die Wahrscheinlichkeit, ein Ergebnis im Ablehnungsbereich von H_1 zu erhalten, also β , grundsätzlich numerisch genau so groß wie die Wahrscheinlichkeit $1 - \alpha$ werden.
2. Möglichkeit: Der E lehnt H_0 ab und entscheidet sich für die H_1 . Diese Entscheidung nun kann deswegen getroffen worden sein, weil in der Tat H_1 zutrifft - in diesem Fall ist die Wahrscheinlichkeit eines Resultates im Annahmebereich von H_1 gleich $1 - \beta$ -, oder aber, weil dem HE ein Fehler 1. Art unterlaufen ist. Zwar wird die Wahrscheinlichkeit dieses Fehlers vom E durch die Wahl eines kleinen Signifikanzniveaus gering gehalten; da aber andererseits die Wahrscheinlichkeit $1 - \beta$, eine zutreffende H_1 als richtig auszuweisen, nicht bekannt ist, kann prinzipiell nicht ausgeschlossen werden, daß ihre numerische Größe gleich der oder kleiner als die Wahrscheinlichkeit α ist.

Wenn wir uns die Implikationsbeziehung zwischen der wissenschaftlichen Hypothese WH_u und der statistischen Hypothese H_1 , also $WH_u \rightarrow H_1$, noch einmal vergegenwärtigen, ergeben sich aus den vorstehenden Erörterungen bestimmte (mögliche) Konsequenzen für die Beurteilung der WH_u :

Begehen wir einen Fehler 1. Art, besteht die Gefahr, daß die wissenschaftliche Hypothese fälschlicherweise als bewährt angesehen wird. Begehen wir dage-

¹⁸⁾ Der interessierte Leser mag selbst Belege für diese Behauptung beibringen; dies stellt jedoch erfahrungsgemäß keine Schwierigkeit dar.

gen einen Fehler 2. Art, besteht die Gefahr, daß die wissenschaftliche Hypothese irrtümlich falsifiziert wird.“)

Wie bereits im Abschnitt 6 angesprochen, ist es unabdingbar, eine zutreffende statistische Hypothese mit großer Wahrscheinlichkeit annehmen und eine nicht-zutreffende mit großer Wahrscheinlichkeit ablehnen zu können, da nur auf diese Weise - langfristig - gerechtfertigte Falsifikationen oder aber Bestätigungen von wissenschaftlichen Hypothesen zu erwarten sind.

Es ist daher zwingend erforderlich, nicht nur die Wahrscheinlichkeit α für einen Fehler 1. Art zu kontrollieren und gering zu halten - hierdurch wird der Gefahr fälschlicher Bestätigungen der WH_u entgegengewirkt -, sondern auch die Wahrscheinlichkeit β für einen Fehler 2. Art zu kontrollieren und gering zu halten - hierdurch wird die Gefahr ungerechtfertigter Falsifikationen in Grenzen gehalten.“)

Diese Forderung nach Kontrolle *beider* Fehlerwahrscheinlichkeiten sind nicht neu; sie wurde bereits von Sterling (1960), Smart (1964), Bakan (1966), Brendenkamp (1969b, 1972, 1975, 1979, 1980), Krause & Metzler (1978) und Witte (1977, 1980) erhoben, um nur einige Autoren zu nennen. Die aus ihr folgenden konkreten Handlungsanweisungen für die Praxis werden allerdings selten realisiert. Bevor wir uns mit diesen näher befassen, wollen wir zunächst die Größen zusammenstellen, die den Ausgang eines (beliebigen) Signifikanztests determinieren, und anschließend einen Blick auf die verbreitete Forschungspraxis werfen.

7.4 Die Determinanten eines Signifikanztests

Unseren bisherigen Darstellungen kann entnommen werden, daß der Ausgang eines Signifikanztests stets vornehmlich von den folgenden vier Größen abhängig ist:

1. dem Signifikanzniveau α ,
2. der Teststärke $1 - \beta$,
3. dem in den empirischen Daten enthaltenen experimentellen Effekt EE,
4. der Varianz der Stichprobenverteilung der gewählten Teststatistik, die ihrerseits stets von der Stichprobengröße N resp. n abhängt (vgl. zu Einzelheiten Abschnitt 8.3).

Die Beziehungen zwischen diesen vier Größen werden üblicherweise in den Grundlagen-Texten nur am Rande behandelt (vgl. jedoch u.a. Hays, 1963,

¹⁹⁾ Selbstverständlich treffen diese Ausführungen mutatis mutandis auch in den Fällen zu, in denen die wissenschaftliche Hypothese eine statistische Null-Hypothese impliziert: $WH_z \rightarrow H_0$ (Hager & Westermann, im Druck, b).

1977; Leiser, 1978; Haagen & Seifert, 1979) und in der Praxis (zu) wenig beachtet.

Um sich eine zumindest ungefähre Vorstellung über die Bedeutung zu verschaffen, der insbesondere der Stichprobengröße N für den Ausgang eines Signifikanztests, d.h. der Beurteilung eines empirischen Resultates auf statistische Signifikanz, zukommt, sollte der Leser sich mit den entsprechenden Beispielen bei Leiser (1978, 178-191) oder auch bei Hays (1977, e.g. 357-362) vertraut machen. Zur Verdeutlichung der für uns wichtigsten Beziehungen wollen wir uns kurz mit den Tabellen C 4 und C 5 in Gaensslen & Schubö (1973, 1976, 314-317) befassen (vgl. ersatzweise etwa Fisher & Yates, 1963, 63, oder Bortz, 1979, 832f.). Mittels dieser Tabellen kann der empirisch erhobene Wert für einen Produkt-Moment-Korrelationskoeffizienten auf statistische Signifikanz beurteilt werden.

Aus den Tabellen wählen wir einige beliebige Werte für r aus, stellen fest, ob diese Werte bei vorgegebener Stichprobengröße N und vorgegebenem Signifikanzniveau α als statistisch signifikant (s.) oder nicht signifikant (n.s.) beurteilt werden. Diese Informationen sind in den Tabellen 7.1 und 7.2 neu zusammengestellt und um eine Angabe über den jeweiligen experimentellen Effekt EE erweitert worden, der hier gleich dem Quadrat des Korrelationskoeffizienten ist.

Tabelle 7.1: Statistische Signifikanz von r bei $\alpha = 0.05$.

r	$EE = r^2$	Stichprobengröße N					
		5	8	16	32	400	1000
0,90	0,81	S.	S.	S.	S.	S.	S.
0,35	0,12	n.s.	n.s.	n.s.	S.	S.	S.
0,12	0,01	n.s.	n.s.	n.s.	n.s.	S.	S.
0,08	0,006	n.s.	n.s.	n.s.	n.s.	n.s.	S.

Tabelle 7.2: Statistische Signifikanz von r bei $\alpha = 0,01$.

r	$EE = r^2$	Stichprobengröße N					
		5	8	16	32	400	1000
0,90	0,81	n.s.	S.	S.	S.	S.	S.
0,35	0,12	n.s.	n.s.	n.s.	n.s.	S.	S.
0,12	0,01	n.s.	n.s.	n.s.	n.s.	n.s.	S.
0,08	0,006	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Die Schlüsse, die aus den in den o.a. Tabellen enthaltenen Informationen zu ziehen sind, können ungeachtet der jeweils benutzten Teststatistik Gültigkeit für alle Signifikanztests beanspruchen:

1. Ist ein Ergebnis statistisch signifikant geworden, bedeutet dies lediglich, daß ein überzufälliger experimenteller Effekt im Experiment aufgetreten ist.²⁰⁾ Wie groß dieser ist, kann durch die statistische Signifikanz nicht festgestellt werden.
2. Jeder beliebig kleine EE kann ungeachtet seiner inhaltlichen Bedeutung statistisch „signifikant gemacht“ werden (Kleiter, 1969, 150f.), indem man eine genügend große Stichprobe benutzt. Unter sonst gleichen Bedingungen ist die statistische Signifikanz ausschließlich von der Stichprobengröße abhängig. Die Beispiele, die etwa Nunnally (1960), Bakan (1966), Meehl (1967), Kleiter (1969) und Cowles (1974) angeben, belegen, daß diese Beziehung in der Praxis offenbar sehr wohl bekannt ist und eingesetzt wird, weil aus noch anzuspreekenden Gründen signifikante Resultate die einzig erwünschten Ausgänge von Signifikanztests darstellen (vgl. hierzu Abschnitt 7.4.1).
3. Je kleiner ein EE ist, desto größer muß der Stichprobenumfang N sein, um diesen EE als statistisch signifikant ausweisen zu können.
4. Je kleiner ein EE ist, desto geringer ist unter sonst gleichen Bedingungen die Teststärke $1 - \beta$, wie zwar nicht den o.a. Tabellen, wohl aber den Gütefunktionen zu entnehmen ist. Eine Erhöhung der Teststärke ist dann über die Erhöhung des Signifikanzniveaus und/oder des Stichprobenumfanges möglich oder aber durch die Wahl eines anderen Versuchsplanes und/oder Auswerteverfahrens.

Die weiteren Beziehungen mag sich der interessierte Leser selbst ableiten. Wir wollen uns statt dessen im folgenden kurz der Frage zuwenden, warum der statistischen Signifikanz gemeinhin eine überragende Bedeutung beigemessen wird.

7.4.1 Forschungs- und Publikationspraxis 1: Signifikanzniveau und p -Werte

Bei Durchsicht psychologischer Fachzeitschriften wird man sehr leicht feststellen können, daß bei der Darstellung von empirischen Resultaten meist weder von einer Irrtumswahrscheinlichkeit α noch von der Möglichkeit eines Fehlers 2. Art die Rede ist. Statt dessen findet man sog. „ p -Werte“, die die Wahrscheinlichkeit des gefundenen und aller (noch) weiter von H_0 abweichenden Resultate unter der Voraussetzung angeben, daß die H_0 zutrifft. Diese p -Werte werden nach der Berechnung des empirischen

²⁰⁾ Das Auftreten eines EE ist nicht an die statistische Signifikanz gebunden. Wie Hays (1963, 326) im einzelnen darlegt und begründet, stellt das völlige Fehlen eines EE in der Empirie eine Ausnahme dar.

Wertes der benutzten Teststatistik entsprechenden Tabellen (etwa Pearson & Hartley, 1954, 1972; Fisher & Yates, 1963) entnommen und entsprechen im Grunde einem „gleitenden Signifikanzniveau“. Sie werden dann ungeachtet geringfügiger Unterschiede bei einzelnen Autoren i. a. nach folgendem Schema „beurteilt“ - vgl. dazu Harnatt (1975, 603), Krause & Metzler (1978, 225), Bortz (1979, 146) und Haagen & Seifert (1979, 203):

- $p = 0,1$ „bedeutet“: „symptomatische Abweichung“;²¹⁾
 $p = 0,05$ „bedeutet“: „signifikante Abweichung“, oft nur durch einen Stern („*“) gekennzeichnet;
 $p = 0,01$ „bedeutet“: „sehr“ oder „hoch signifikante Abweichung“, meist durch zwei Sterne („**“) symbolisiert;
 $p = 0,001$ „bedeutet“: „äußerst“ oder „höchst signifikante Abweichung“, abgekürzt durch drei Sterne („***“) angegeben.

Wie Rosenthal & Gaito (1963, 1964) sowie Beauchamp & May (1964) experimentell glaubhaft machen konnten, ist das (subjektive) Vertrauen in die Ergebnisse von Signifikanztests um so größer, je kleiner die nachträglich ermittelten p -Werte sind.

Diese Interpretation der Wahrscheinlichkeiten bestimmter extremer Ereignisklassen unter Gültigkeit der statistischen Null-Hypothese ist dabei nicht ausschließlich charakteristisch für die untersuchten „graduate students“ und „Ph.D.'s“ (Rosenthal & Gaito, 1963, 33; Beauchamp & May, 1964, 272), sondern auch für Herausgeber von Zeitschriften. Offenbar ist nämlich die Chance, daß ein Artikel, der eine statistische Auswertung via Signifikanztest enthält, zur Publikation akzeptiert wird, um so größer, je kleiner der ermittelte p -Wert ist - man vergleiche hierzu insbesondere die durch die einflußreiche Arbeit von Bakan (1966) bekannt gewordene sog. „Herausgeber-Philosophie“ von A. Melton (1962), der im Zeitraum von 1951-1962 das „Journal of Experimental Psychology“ herausgab.

Melton stellt dabei keine Ausnahmeerscheinung, sondern eher einen „typischen“ Vertreter für die weitverbreitete Fehlinterpretation des nachträglich bestimmten p -Wertes dar, wie u.a. durch die (kritischen) Publikationen von Sterling (1959), Cohen (1962, 1965), Smart (1964), Bakan (1966), Skipper, Guenther & Nass (1967), Labovitz (1968), Bozarth & Roberts (1972) sowie Schulman, Kupst & Suran (1976) im einzelnen dokumentiert wird.

Auf die nicht zu unterschätzenden Folgen dieser offenbar unausrottbaren Fehlinterpretation der statistischen Fehlerwahrscheinlichkeit α beim Hypothesentesten als Maß für die Größe oder inhaltliche Bedeutsamkeit eines experimentellen Effektes haben neben den o. gen. Autoren insbesondere Bredenkamp (1972, 51-73), Greenwald (1975), Lane & Dunlap (1978) sowie Rosenthal (1979) eindringlich hingewiesen.

Besonders bemerkenswert ist bei diesem Phänomen allerdings, daß diesen speziellen Interpretationen durch die Darstellungen in vielen Einführungslehrbüchern zur Statistik und/oder von renommierten Autoren entschieden Vorschub geleistet wird; Beispie-

²¹⁾ Anstelle des in der Literatur gebräuchlichen Ausdrucks „Abweichung“ benutzen wir in dieser Arbeit die Bezeichnung „experimenteller Effekt“.

le für diese Behauptung finden sich u.a. in Bredenkamp (1972, 57), Harnatt (1975, 600f.) sowie Haagen & Seifert (1979, 202f., 243); siehe ferner auch Guttman (1977, 91 f.) und Bortz (1979, e.g. 143).

7.4.2 Forschungs- und Publikationspraxis II: Experimentelle Effekte und Teststärke

Aufgrund der Ausführungen im vorangegangenen Abschnitt und der unzureichenden Beachtung, die der Teststärke (und den experimentellen Effekten) in der überwiegenden Mehrzahl der Lehrbücher gewidmet wird²²⁾ nimmt es nicht wunder, daß üblicherweise Angaben über die Teststärke (und den EE) nicht publiziert werden, obwohl dies für die EE von einigen Zeitschriften-Herausgebern bereits gefordert wird (vgl. Bredenkamp & Feger, 1970; ferner Lane & Dunlap, 1978; Soderquist & Hussian, 1978).

Allerdings läßt sich die Teststärke im Rahmen von statistischen Reanalysen für zahlreiche experimentelle Anordnungen und Auswerteverfahren auch nach der Datenerhebung und -auswertung noch bestimmen, weil sie bei vorgegebenem Signifikanzniveau α , fester Stichprobengröße N und bekannter Stichprobenverteilung ausschließlich von der Größe des EE in den erhobenen Daten abhängt (s.o.; zur Bestimmung des EE aus den Daten siehe Abschnitt 9.3.2.3 und 9.4). Diese Tatsache haben sich zahlreiche Autoren zunutze gemacht und die Teststärke für publizierte Daten und Experimente analysiert, die in verschiedenen Bereichen der psychologischen Forschung und einigen verwandten Gebieten wie etwa der Kommunikationsforschung durchgeführt worden sind - vgl. u.a. Cohen (1962, 1965, 1973 b), Brewer (1972), Katzer & Sodt (1973), Chase & Tucker (1975), Kroll & Chase (1975), Chase & Baran (1976), Chase & Chase (1976), Schmidt, Hunter & Urry (1976), Treiber (1977), Treinies (1977) sowie Cascio, Valenzi & Silbey (1978, 1980). Diesen Arbeiten läßt sich übereinstimmend entnehmen, daß in der Mehrzahl der reanalysierten Untersuchungen die Teststärke $1 - \beta$ (sehr) gering war.

Als weiteres Resultat der genannten Reanalysen ist bemerkenswert, daß die in den Experimenten aufgetretenen experimentellen Effekte fast durchgängig nur „klein“ gewesen sind - ein kleiner EE entspricht etwa einem Wert von 0,01 in den Tabellen 7.1 und 7.2 (vgl. zu näheren Einzelheiten die Abschnitte 10.3.3.2

²²⁾ Einer der Gründe hierfür mag darin zu suchen sein, daß derjenige der bedeutenden Statistiker, der die heutzutage praktizierte Form der statistischen Inferenz wohl am stärksten geprägt hat, R. A. Fisher, dieses Konzept strikt ablehnte. Er schrieb dazu in seinem Werk „The design of experiments“ (1935, 1951, 17): „The notion of an error of the so-called ‚second kind‘, due to accepting the null hypothesis »when it is false« . . . has no meaning with respect to simple tests of significance, in which the only available expectations are those which flow from the null hypothesis being true.“

und 11.1.3). Wegen der erwähnten Beziehung zwischen der Teststärke und dem EE (s.o. und Abschnitt 7.4) ist dieser Befund nicht verwunderlich. Angemerkt sei noch, daß die publizierten EE fast ausnahmslos als statistisch signifikant beurteilt worden waren.

Solange die Größe eines aufgedeckten experimentellen Effektes unter ausschließlicher Verwendung des (nachträglich bestimmten) p-Wertes „beurteilt“ wird, kann die Bedeutung der einzelnen empirischen Befunde für die zu prüfende wissenschaftliche Hypothese nicht abgeschätzt werden, und vor diesem Hintergrund muß der folgenden Aussage von Guttman (1977, 92) im Prinzip zugestimmt werden: „No one has yet published a scientific law in the social sciences which was developed, sharpened, or effectively substantiated on the basis of tests of significance.“

7.4.3 Forschungs- und Publikationspraxis III: Entwicklung einer vorläufigen Zielvorstellung

Aus unseren bisherigen Ausführungen geht u.a. hervor, daß von den vier Determinanten eines jeden Signifikanztests sowohl das Signifikanzniveau α als auch die Teststärke $1 - \beta$ vom E selbst zu kontrollieren und damit vor einem Experiment festzulegen sind; m.a.W. muß der E jeweils angeben, wie groß seine maximale Bereitschaft ist, einen Fehler 1. und einen Fehler 2. Art zu begehen.

Folgt der E diesem Vorgehen, ist der Ausgang des Signifikanztests nur noch von der tatsächlichen Größe des EE und dem Stichprobenumfang N abhängig.

Nun ist der Stichprobenumfang N in der Praxis meist eine Größe, die sich nach der Anzahl der jeweils (ad hoc) zur Verfügung stehenden Untersuchungseinheiten (V_{pn}) ergibt (siehe u.a. Smart, 1966; Higbee & Wells, 1972; Oakes, 1972). Aus dieser Tatsache ergeben sich die folgenden Probleme (vgl. auch Teil 4):

- (1) Wenn die Stichprobengröße i.a. nach dem „Kriterium“ der Bereitwilligkeit und Verfügbarkeit der V_{pn} fixiert wird, ist die Größe des bei vorgegebenen Höchstwerten für α und β im Experiment entdeckbaren EE vom E nicht kontrollierbar. Es sollte jedoch deutlich geworden sein, daß es gerade die Größe des EE ist, der die besondere Aufmerksamkeit des E gelten muß, denn diese Größe stellt einen wesentlichen Bestandteil der Interpretation der experimentellen Daten im Hinblick auf die wissenschaftliche Hypothese dar (vgl. zu Einzelheiten insbesondere Teil 11).

Man kann sich diesen Zusammenhang auf einer intuitiven Basis verdeutlichen, indem man sich überlegt, daß eine WH_{ij} , die für eine spezifische Versuchsanordnung einen EE

in der Größenordnung $r^2 = 0,60$ vorhersagt²³⁾, wesentlich „attraktiver“ ist als eine konkurrierende Hypothese WH_v zur Erklärung des gleichen Phänomens, die aber nur einen EE der Größe $r^2 = 0,10$ erwarten läßt. Die WH_u ist attraktiver, weil sie größere Anteile der Datenvarianz „aufzuklären“ vermag und damit auch genauere Vorhersagen ermöglicht als die konkurrierende WH_v . Dieser Aspekt ist sowohl im Hinblick auf das Ziel einer möglichst genauen und sparsamen wissenschaftlichen Erklärung empirischer Sachverhalte von Bedeutung als auch für die Ableitung von technologischen Prognosen (siehe dazu Teil 5) und würde bei ausschließlicher Betrachtung der statistischen Signifikanz weitestgehend unbeachtet bleiben. Anzumerken ist hierbei noch, daß dem EE auch dann eine zentrale Bedeutung zukommt, wenn die psychologische Hypothese keine quantitativen Prognosen zuläßt, weil auch in diesem Fall festgestellt werden kann, wieviel Datenvarianz sie im Vergleich zu konkurrierenden Hypothesen „aufklärt“ - siehe dazu im einzelnen Teil 9.

Diese Ausführungen legen insgesamt nahe, zusätzlich zu den maximalen Fehlerwahrscheinlichkeiten α und β als weiteres Kriterium auch noch den erwarteten experimentellen Effekt als Mindesteffekt (EEM) festzulegen und unter Benutzung der Teststärkefunktionen die Stichprobengröße so zu bestimmen, daß dieser EEM auch mit der vorgegebenen Teststärke im Experiment entdeckt wird, sofern er tatsächlich vorhanden ist. Im Teil 10 werden wir Strategien vorstellen, die die Bestimmung des Stichprobenumfanges nach diesen Kriterien ermöglichen, und im Teil 11 werden wir darauf eingehen, auf welche Arten man zu Festlegungen des EEM gelangen kann.

Wenden wir uns zunächst jedoch dem zweiten Problem zu, das sich aus bestimmten üblichen Praktiken ergeben kann:

- (2) Die zur Verfügung stehenden V_{pn} liefern in der Regel keine Daten, die man als zufällige Stichprobe im oben definierten Sinne auffassen kann. Welche Rechtfertigung gibt es angesichts dieser Tatsache für den Einsatz von Signifikanztests?

Obwohl wir diese Frage erst im Abschnitt 8.2.6 behandeln werden, wollen wir im folgenden Abschnitt 7.5 einige zu ihrer Beantwortung relevante Vorinformationen zusammentragen, indem wir u.a. auf Signifikanztests eingehen, deren valider Einsatz nicht an die Voraussetzung von Zufallsstichproben gebunden ist. Des weiteren sprechen wir dort Tests an, deren Anwendung insgesamt an schwächere Voraussetzungen gebunden ist, als dies bei den bisher angesprochenen Verfahren (wie t -, χ^2 - und F -Test) der Fall ist; die Kenntnis bestimmter Eigentümlichkeiten dieser sog. „nicht-parametrischen und verteilungsfreien“ Verfahren ist für verschiedene nachfolgende Ausführungen von besonderem Interesse.

²³⁾ Quadrierte Korrelationen (r^2) sind, wie wir im Teil 9 noch sehen werden, spezielle Maße für experimentelle Effekte (vgl. auch Abschnitt 7.4).

7.5 Arten statistischer Hypothesen und ihre Prüfung

Neben den bereits eingeführten Differenzierungen zwischen einer statistischen Null- und einer Alternativhypothese (siehe Abschnitt 7.2 und 7.3) sowie zwischen einfachen und zusammengesetzten Hypothesen (Abschnitt 7.3.1) sind weitere Unterscheidungen zu beachten, auf die wir im folgenden eingehen wollen.

7.5.1 Gerichtete und ungerichtete Hypothesen und ihre Prüfung

Eine „gerichtete statistische Hypothese“ legt eine Ordnung für die zu vergleichenden Parameter fest; sie spezifiziert die „Richtung“ eines Unterschiedes etwa wie folgt:

$$H_1: \mu_1 < \mu_2 \text{ oder } H_1: \mu_1 - \mu_2 < 0.$$

Dagegen sagt eine „ungerichtete statistische Hypothese“ nur etwas über Unterschiede zwischen Parametern in beliebiger Richtung aus, etwa in dieser Form:

$$H_1: \mu_1 \neq \mu_2 \text{ oder } H_1: \mu_1 - \mu_2 \neq 0.$$

Diesen Hypothesenarten entsprechen unterschiedliche wissenschaftliche Hypothesen; auf die Implikationsbeziehungen gehen wir im Abschnitt 8.1 näher ein.

Von der Art und Weise, wie eine statistische Hypothese formuliert wird, muß unterschieden werden, wie sie aufgrund der zur Verfügung stehenden Stichprobenverteilungen geprüft wird, nämlich entweder „einseitig“ oder „zweiseitig“.

Diese Termini beziehen sich darauf, ob an beiden „Enden“ („Seiten“) einer Stichprobenverteilung je ein Rejektionsbereich definiert wird (zweiseitiger Test) oder nur an einem „Ende“ (einseitiger Test) (vgl. u.a. Edwards, 1971, 109-112; Hays, 1977, 369-374; Bortz, 1979, 150-152).

Ein einseitiger Test führt dabei unter sonst gleichen Bedingungen eher zur Ablehnung der geprüften H_0 als ein zweiseitiger Test. Allerdings ist nur eine begrenzte Anzahl von statistischen Verfahren verfügbar, die dem Experimentator die Wahl zwischen einem oder zwei Ablehnungsbereichen ermöglichen; zu diesen zählen etwa die Binomial-Tests und der t-Test. Besteht die Wahlmöglichkeit bzgl. des Rejektionsbereiches, werden gerichtete statistische Hypothesen adäquat unter Verwendung einseitiger Tests geprüft - vgl. Abschnitt 8.1.2. In manchen Fällen bestimmt (leider) das gewählte Testverfahren, ob der E einen oder zwei Rejektionsbereiche festlegen kann.

So eignet sich die F-Verteilung zwar zur Beurteilung von gerichteten Hypothesen über Varianzen (vgl. Hays, 1977, 445-447); dagegen können mit dem gebräuchlichen Varianz- resp. regressionsanalytischen F-Test nur ungerichtete Hypothesen über Mittelwerte resp. quadrierte multiple Korrelationskoeffizienten geprüft werden²⁴⁾ - zu weiteren Einzelheiten siehe etwa Wainer (1972, 1973) und Gaito (1977a).

über die verschiedenen Arten, statistische Hypothesen zu formulieren und statistisch zu prüfen, ist es im Anschluß an eine Arbeit von Marks (1951) zu einer langandauernden kontroversen Diskussion gekommen, deren seinerzeit vorläufiger Schlußpunkt von Kaiser (1960) gesetzt wurde. Diese und weitere Originalarbeiten haben Lieberman (1971) und Steger (1971) zusammengestellt; einen Überblick über die Diskussion geben auch Glass & Stanley (1970, 288f.); ferner siehe Shaffer (1972) und Gibbons & Pratt (1975).

7.5.2 *Parametrische und nicht-parametrische Hypothesen und ihre Prüfung*

Formuliert man statistische Hypothesen über die Parameter von Populationsverteilungen, spricht man von „parametrischen Hypothesen“, wobei davon ausgegangen werden muß, daß die Population durch eine endliche Menge dieser Parameter charakterisiert werden kann (vgl. Menges, 1972, 296); Beispiele für derartige Hypothesen finden sich in den Abschnitten 6, 7.2 und 7.3.

Formuliert man dagegen Hypothesen über ganze Verteilungen, ohne diese durch Parameter näher spezifizieren zu können, handelt es sich um „nicht-parametrische Hypothesen“; ein Beispiel wäre etwa: $H_0: F_1(Y) = F_2(Y)$, d.h. die abhängige Variable Y hat in den beiden in Frage stehenden Populationen die gleiche Verteilungsfunktion; zu den Einzelheiten siehe etwa Kendall & Stuart (1961, e.g. 161f.) und Menges (1972, 325-328).

Die Prüfung parametrischer Hypothesen erfolgt üblicherweise über „verteilungsgebundene“ Testverfahren, die in der Regel „parametrische“ Tests genannt werden. Das Adjektiv „verteilungsgebunden“ bezieht sich dabei auf die Tatsache, daß die Stichprobenverteilungen der entsprechenden parametrischen Teststatistiken wie t , χ^2 und F nur unter der Voraussetzung ableitbar sind, daß die Populationsverteilungen der möglichen Rohwerte eine genau spezifizierte Form aufweisen, die allerdings bei der Hypothesenformulierung in der Praxis nur selten expliziert wird. Zur Ableitung der genannten Statistiken muß die Population der Rohwerte durch eine „normal“ genannte Dichte-Funktion beschreibbar sein (zu Einzelheiten siehe etwa Hays, 1977, Kap. 8 bis 11). Sobald diese Annahme nicht aufrecht erhalten werden kann, ist - genau genommen - die Anwendung der parametrischen Tests nicht mehr valide - zur Relativierung dieser Aussage siehe jedoch unten Abschnitt 8.2.

²⁴⁾ Auf die Bedeutung dieser Aussage kommen wir im Abschnitt 8.2 zurück.

Gegen die Berechtigung dieser Annahme sind verschiedentlich starke Bedenken geäußert worden (u.a. Bradley, 1968, 1972, 1977; Menges, 1972, 248f.). Dies hat zur Entwicklung der sog. „verteilungsfreien“ oder „nicht-parametrischen“ Testverfahren geführt, deren valide Anwendung an weniger strenge oder keine Annahmen über Populationsverteilungen gebunden ist. Das Adjektiv „nicht-parametrisch“ bezieht sich dabei auf die statistischen Hypothesen über Verteilungsfunktionen, die adäquat mittels nicht-parametrischer Tests überprüft werden. Die Bezeichnung „verteilungsfrei“ dagegen soll darauf hindeuten, daß die Ableitung der Stichprobenverteilung der jeweiligen Teststatistik unter Gültigkeit der H_0 unabhängig von spezifischen Annahmen über die Population(en) erfolgt (vgl. zur Unterscheidung neben der unten im Abschnitt 7.5.3.2 genannten Literatur etwa McSweeney & Marascuilo, 1969; sowie McSweeney & Katz, 1978). Allerdings können bestimmte Hypothesen auch mit nicht-parametrischen Tests nur geprüft werden, wenn man bereit ist, Annahmen bzgl. der Populationsverteilung(en) zu treffen, etwa die, daß zwei oder mehr Verteilungen symmetrisch oder sogar von gleicher Form („homomer“) sind - siehe dazu Edgington (1965), Lienert (1973, 107), Marascuilo & McSweeney (1977, 269, 334). Die aus einer WH abgeleitete Hypothese über die Rangordnung von Mittelwerten oder Medianen bspw. kann verteilungsfrei nur geprüft werden, wenn man die Symmetrie der zugrundeliegenden Verteilungen annimmt. Ist man nicht bereit oder in der Lage, derartige Annahmen zu akzeptieren, erlaubt die Ablehnung einer verteilungsfrei geprüften H_0 meist nur die Aussage, daß die untersuchten Stichproben Populationen mit verschiedenen Verteilungsfunktionen entstammen.

Die Ableitung der Stichprobenverteilungen erfolgt bei den meisten nicht-parametrischen Testverfahren aufgrund vglw. einfacher kombinatorischer und wahrscheinlichkeitstheoretischer Überlegungen. Dies wird besonders deutlich, wenn man als Untergruppe der verteilungsfreien Verfahren zur Prüfung nicht-parametrischer Hypothesen die sog. „Randomisierungstests“ betrachtet.

Bei ihnen erhält man bspw. die Prüfverteilung der interessierenden Statistik, indem man alle möglichen Permutationen („Randomisierungen“) oder Zufallsanordnungen der empirisch erhobenen Daten herstellt und für jede dieser Anordnungen den Wert der betr. Statistik errechnet. Auf diese Weise kann eine Häufigkeitsverteilung der Werte der Statistik konstruiert werden, aus der wiederum die relativen Häufigkeiten interessierender extremer Ereignisklassen ermittelt werden. Sind diese bestimmt, verläuft die statistische Signifikanzbeurteilung nach dem im Abschnitt 7.2 skizzierten Verfahren. Bzgl. der Einzelheiten verweisen wir auf Kempthorne (1955), Fisher (1956), Scheffé (1959, 291-330), McHugh (1963), Edgington (1964b, 1969b), Ray (1966), Bradley (1968, Kap. 4) sowie Pfanzagl (1978, 142-147, 150-153); zur Kritik dieses Ansatzes siehe etwa Witte (1980, 122f.).

Ein wesentlicher Nachteil der Permutations- oder Randomisierungstests besteht in dem fast stets enormen Rechenaufwand, den es erfordert, aus allen möglichen Permutationen der Daten die überhaupt möglichen Realisierungen der Teststatistik zu berechnen

(vgl. Edgington, 1964b, 447). Als Ausweg empfiehlt es sich, statt aller möglichen Permutationen der *Rohwerte* lediglich die Permutationen der *Ränge* zu betrachten, die man den Rohwerten zuweisen kann. Auf diesem Prinzip beruhen die gebräuchlichsten verteilungsfreien Verfahren wie die Rangtests nach Wilcoxon, Mann und Whitney, Friedman sowie Kruskal und Wallis (vgl. dazu Lienert, 1973; Marascuilo & McSweeney, 1977).²⁵⁾

Weitere Vereinfachungen im Gebrauch der Randomisierungstests ergeben sich, wenn man die gebräuchlichen parametrischen Verfahren als approximative Randomisierungstests interpretiert (vgl. Edgington, 1966, 1973; Alf & Abrahams, 1972, 1973; Bredenkamp, 1972, 28-33). Diese Interpretation ist dann vertretbar, wenn der Nachweis gelingt, daß die aus allen Permutationen der Daten eines beliebigen Experiments berechenbaren Werte der gewählten Teststatistik approximativ den entsprechenden theoretischen (kontinuierlichen) Stichprobenverteilungen folgen. Dieser Nachweis ist für den t-Test und verschiedene Varianz- und kovarianzanalytische F-Tests in der Tat gelungen (vgl. zusammenfassend Scheffé, 1959, 291-330; ferner Baker & Collier, 1966, 1968; Collier & Baker, 1966; Toothaker, 1971, 1972; sowie Robinson, 1973a, b).

Da für die valide Anwendung von Randomisierungstests nur die Forderung erfüllt sein muß, daß die Untersuchungseinheiten den experimentellen Bedingungen zufällig zugeteilt worden sind, nicht jedoch, daß sie Zufallsstichproben darstellen, ergeben sich aus diesen Befunden wesentliche Konsequenzen bzgl. der Berechtigung parametrischer Tests, auf die wir im Abschnitt 8.2.6 eingehen.

7.5.3 Zur Wahl zwischen parametrischen und nicht-parametrischen Verfahren

Aufgrund des bisher Gesagten könnte der Eindruck entstehen, als läge die Entscheidung zwischen den unterschiedlichen Arten der Hypothesenformulierung und -prüfung ausschließlich beim E, zumal aus der WH fast stets parametrische wie nicht-parametrische Hypothesen ableitbar sind. Dieser Eindruck wäre jedoch falsch.

Vielmehr sind die folgenden übergeordneten Entscheidungsgesichtspunkte von zentraler Bedeutung:

- (1) Eine Voraussetzung für die sinnvolle Interpretation der empirischen Resultate besteht darin, daß sowohl die statistischen Hypothesen wie das Auswertungsverfahren dem Skalenniveau der empirischen AV angemessen sind (vgl. Abschnitt 2.4). Wir stellen in den folgenden Abschnitten 7.5.3.1 bis 7.5.3.3 einige Hinweise auf adäquate Auswertetechniken in Abhängigkeit vom Skalenniveau der AV zusammen.

²⁵⁾ Ein anderer Ausweg besteht darin, eine Zufallsstichprobe von Werten der Statistik zu ziehen, aufgrund derer man zur Bestimmung der relativen Häufigkeit bzw. Wahrscheinlichkeit der Resultatsklasse von Ereignissen gelangt, die das eine tatsächliche Resultat enthält; man spricht dann von „approximativen Randomisierungstests“ (Edgington, 1964b, 1969a, b, 152-157; Bredenkamp, 1972, 30f.).

- (2) Hat man sich für eine parametrische Hypothese entschieden, ist zu beachten, daß die entsprechenden Tests in der Regel auf (sehr) restriktiven Annahmen beruhen, von denen wir die der normalverteilten Rohwerte in den Populationen bereits angesprochen haben; genauere Einzelheiten zu dieser und den übrigen Voraussetzungen sowie zu den Konsequenzen ihrer Verletzung finden sich im Abschnitt 8.2.
- (3) Unter dem Kriterium der Strenge einer Prüfung ist es wichtig, sich mit der Effizienz der nicht-parametrischen im Vergleich zu den parametrischen Testverfahren zu befassen - dies geschieht im Abschnitt 7.5.4.

Befassen wir uns zunächst mit der Wahl von Tests in Abhängigkeit vom Skalenniveau!

7.5.3.1 Auswertung von Häufigkeitsdaten (Nominal-Niveau)

Obwohl davon auszugehen ist, daß in einem *Experiment* relativ selten nominale (qualitative, kategoriale oder Häufigkeits-)Daten erhoben werden, geben wir hier einige Hinweise auf entsprechende Auswertetechniken, weil sie in der nicht-experimentellen Forschung vglw. häufig anzutreffen sind und weil ihr Bekanntheitsgrad bei Psychologen recht gering zu sein scheint.

Die Erhebung von Häufigkeitsdaten (vgl. zur eingehenderen Charakterisierung und Differenzierung etwa Lienert, 1973, 86f.) dient in der Regel der Überprüfung von Hypothesen über die Zusammenhänge zwischen qualitativen Merkmalen. Liegen derartige Zusammenhänge vor, äußern sie sich in unterschiedlichen Häufigkeiten oder Proportionen für einzelne oder mehrere Merkmalskombinationen - man spricht dann von einer „statistischen Interaktion oder Assoziation zwischen qualitativen Merkmalen“, bei deren Vorliegen die einzelnen Merkmale nicht unabhängig voneinander sind.

In der statistischen Null-Hypothese wird allerdings von der statistischen Unabhängigkeit ausgegangen, und unter ihrer Gültigkeit lassen sich erwartete Häufigkeiten oder Proportionen berechnen. Zur Prüfung dieser Hypothese werden häufig die Binomial- oder die Multinomialverteilung sowie deren Spezialfälle herangezogen, und zwar in Abhängigkeit davon, ob ein dichotomes oder polytomes Merkmal vorliegt. Diese Wahrscheinlichkeitsverteilungen können unter bestimmten Voraussetzungen von der χ^2 -Verteilung approximiert werden.

Die Verfahren zur Auswertung nominaler Daten, die häufig in Form sog. „Kontingenztafeln“ zusammengestellt werden, haben in den letzten Jahren stark an Vielfalt zugenommen; einen zusammenfassenden Überblick über diese Entwicklung, die besonders durch Arbeiten aus dem Bereich der Soziologie begünstigt wurde, findet man bei Meredith, Frederiksen & McLaughlin (1974) sowie besonders bei Smith (1976b). Darüber hinaus sind einzelne Verfahren ausführlicher dargestellt etwa bei Grizzle, Starmer & Koch (1969), Fleiss (1973), Shaffer (1973), Goodman (1978), Upton (1978), Küchler (1979) sowie Langeheine (1980); auf die weitergehenden Literaturhinweise bei den beiden letztgenannten Autoren sei besonders verwiesen.

Als Parallelentwicklung zu dem Log-linearen Ansatz von Goodman (siehe zusammenfassend Goodman, 1978) kann im deutschen Sprachraum die „Konfigurationsfrequenz-

analyse" und ihre Spezialfälle von Krauth & Lienert (1973; vgl. auch Lienert, 1978) verstanden werden, die sich besonders zur Behandlung „klinischer Fragestellungen“ eignet; zum Vergleich der KFA mit dem Log-linearen Ansatz siehe Krauth (1980).

Ferner sei noch der informationstheoretische Ansatz zur Auswertung von Kontingenztabellen erwähnt, der von Kuhback und seinen Mitarbeitern entwickelt worden ist (siehe etwa Ku & Kuliback, 1968; Gokhale & Kuliback, 1978; ferner Adam & Enke, 1972).

Weitere Auswertungshinweise sind den Standard-Lehrbüchern der Statistik sowie insbesondere der im folgenden Abschnitt genannten Literatur zu entnehmen.

7.5.3.2 Auswertung von Rangdaten (Ordinal-Niveau)

Ordinale oder Rangdaten unterscheiden sich von nominalen oder Häufigkeitsdaten dadurch, daß sie eine Rangordnung repräsentieren (vgl. zu näheren Einzelheiten u.a. Lienert, 1973, 87-91); dies ist etwa bei Präferenzurteilen der Fall. Andererseits kann man metrische Daten in eine Rangreihe transformieren. Dies ist bspw. dann notwendig, wenn infolge der Verletzung von Annahmen eine parametrische Hypothese durch eine nicht-parametrische ersetzt werden muß, die mittels eines Rangtests voraussetzungsärmer geprüft werden kann. Wenn man bereit ist, bestimmte Annahmen etwa über die Symmetrie der Populationsverteilungen zu akzeptieren, kann eine Hypothese über Mittelwerte bspw. durch eine solche über Mediane ersetzt werden. Die entsprechende Prüfung stellt ein nicht-parametrisches Homologon der Prüfung der Mittelwertshypothese dar, wenn ein Testverfahren gewählt wird, das besonders sensitiv auf Lageunterschiede „anspricht“; es ist dann bspw. der t-Test durch den U-Test von Mann & Whitney (1947; vgl. dazu auch Berchtold, 1979) oder den Wilcoxon-Test (vgl. Lienert, 1973) zu ersetzen, während die Rangvarianzanalysen nach Kruskal & Wallis (1952) und Friedman (1937) einfachen Varianzanalysen entsprechen. Wie insbesondere D'Agostino (1972) und Silverstein (1974) ausführen, sind diese Verfahren einander jeweils „asymptotisch äquivalent“.

Nähere Einzelheiten zu den bei Vorliegen einer Rangskala angemessenen Verfahren entnehme man den Büchern von Siegel (1956, 1976), Walsh (1962, 1965, 1968), Bradley (1968), Gibbons (1971), Puri & Sen (1971), Hollander & Wolfe (1973), Lienert (1973, 1975, 1978), Lehmann (1975), Renn (1975), Marascuilo & McSweeney (1977) sowie Büning & Trenkler (1978), die auch zahlreiche Hinweise zur Auswertung nominaler Daten enthalten; eine zusammenfassende Literatur-Übersicht gibt Singer (1979).

Spezielle Abhandlungen zur Gruppe der „Trendanalysen“ genannten Verfahren finden sich etwa bei Ferguson (1965), Sarris (1968), Marascuilo & McSweeney (1967, 1977) und Bredenkamp (1971).

7.5.3.3 Auswertung von Intervalldaten (Intervall-Niveau)

Die Behandlung der Verfahren, die bei Vorliegen intervallskalierter Daten sinnvoll und am teststärksten sind, nimmt in fast jedem Standardlehrbuch den breitesten Raum ein. Wir können uns daher an dieser Stelle darauf beschränken, nur einige derjenigen (Lehr-)Bücher anzugeben, die wertvolle Hinweise zur Auswertung von Experimenten geben können. Bei der Auswahl haben wir uns auch an dem Kriterium der raschen Verfügbarkeit dieser Werke orientiert.

Die umfassendste Darstellung der verschiedensten Aspekte der Versuchsauswertung findet sich derzeit wohl in der „Verfahrensbibliothek Versuchsplannung und -auswertung“ von Rasch, Herrendörfer, Bock & Busch (1978); vgl. ergänzend dazu Bliss (1967, 1970) und Bortz (1979). Weniger umfassend im Detail, aber stärker die Verbindung zwischen Auswerteverfahren und wissenschaftlicher Hypothese betonend sind u.a. Namboodiri, Carter & Blalock (1975), Box, Hunter & Hunter (1978) sowie Henning & Muthig (1979). überwiegend mit der Auswertung der zahlreichen „klassischen“ Versuchspläne der Varianzanalyse befassen sich neben Winer (1962, 1971), Edwards (1971, 1980) und Keppel (1973) u.a. auch Lindquist (1953), Cox (1958), Cochran & Cox (1968), Mendenhall (1968), Kirk (1968), Myers (1972), Snedecor & Cochran (1972), Kempthorne (1973), Lindman (1974), Lee (1975), John & Quenouille (1977), Eimer (1978), Diehl (1979) und McGuigan (1979).

Auf die Behandlung der gleichen einfachen und komplexen Designs durch regressionsanalytische Verfahren gehen u. a. Draper & Smith (1966), Kerlinger & Pedhazur (1973), Cohen & Cohen (1975), Gaensslen & Schubö (1973, 1976) und Moosbrugger (1978) ein.

Die parametrischen Trendanalysen werden darüber hinaus gesondert von Gaito & Turner (1963), Bredenkamp (1968), Hubert (1973), Gaito (1977b) und Cohen (1980) behandelt.

Über multivariate Testverfahren informieren Cramer & Bock (1966), Morrison (1967, 1976), Bock & Haggard (1968), Rulon & Brooks (1968), Tatsuoka (1969, 1971), McCall (1970), Cooley & Lohnes (1971), Overall & Klett (1972), Gaensslen & Schubö (1973, 1976), Kerlinger & Pedhazur (1973), Finn (1974), Bock (1975), Cohen & Cohen (1975), Harris (1975), Woodward & Overall (1975), Moosbrugger (1978) sowie Bortz (1979).

Unabhängig von der individuellen Entscheidung für oder gegen eine der vorstehenden Arbeiten über parametrische Hypothesentestung halten wir eine zusätzliche Beschäftigung mit dem Konzept der „Exploratorischen Daten-Analyse“, wie es von Tukey (1977; vgl. auch Mosteller & Tukey, 1977) entwickelt wurde, für gewinnbringend.

7.5.4 Zur Frage der relativen Effizienz

Wenn man der Frage nachgeht, warum in der Praxis fast grundsätzlich parametrische Tests den nicht-parametrischen vorgezogen werden, stößt man auf zahlreiche Aspekte, unter denen man die beiden Gruppen von Verfahren vergleichen kann - entsprechende Erörterungen findet man etwa bei Siegel (1956, 1976), Gaito (1959b), Bradley (1968, Kap. 2, 1972) und bei Lienert (1973, Kap. 4); zur Kritik der Unterscheidung siehe etwa McSweeney & Marascuilo (1969).

Wir wollen uns hier auf einige wenige Punkte beschränken. Zum einen liegt die Bevorzugung daran, daß die parametrischen Verfahren vielseitiger einsetzbar sind und in einigen Fällen sogar die einzige Möglichkeit zur Auswertung darstellen; letzteres gilt insbesondere bei der Prüfung komplexer Hypothesen, etwa über (varianzanalytische) statistische Interaktionen, oder auch bei der Auswertung komplexer, insbesondere multivariater Versuchspläne - siehe hierzu im einzelnen Bradley (1968), Puri & Sen (1971, e.g. 331-337), McSweeney & Katz (1978) und Singer (1979).

Zum anderen stellen die gebräuchlichen parametrischen Tests die teststärksten und insgesamt „besten“ Tests „ihrer“ Hypothesen dar, wenn alle Voraussetzungen zu ihrer Anwendung erfüllt sind (vgl. zu den Test-Gütekriterien Menges, 1972, 333-335). Unter genau dieser Bedingung haben die parametrischen Tests eine Effizienz von Eins, während die analogen nicht-parametrischen Tests fast ausnahmslos²⁶⁾ weniger effizient, d.h. - vereinfacht ausgedrückt - weniger teststark sind (siehe etwa Büning & Trenkler, 1978, 282). Exakte Definitionen der verschiedenen Arten von „(relativer) Effizienz“ findet man bei Marascuilo & McSweeney (1977, Kap. 4) und bei Büning & Trenkler (1978, Kap. 9).

Praktisch bedeutet eine geringere Effizienz, daß unter sonst gleichen Bedingungen bei Verwendung eines nicht-parametrischen Tests mehr Vpn benötigt werden als beim analogen parametrischen Test, um eine zutreffende Alternativhypothese aufgrund der Stichprobendaten annehmen zu können.

Sind dagegen mehrere Voraussetzungen zur validen Anwendung der parametrischen Verfahren simultan verletzt, liegt deren Effizienz in der Regel beträchtlich *unter* der der homologen nicht-parametrischen Tests - vgl. zu Einzelheiten Hodges & Lehmann (1956), Wetherill (1960), Pratt (1964), Renn (1975, 43f.), Blair, Higgins & Smitley (1980) und Hager et al. (im Druck).

²⁶⁾ Eine der wesentlichen Ausnahmen stellen die sog. „Normal-Scores-Tests“ dar, die stets mindestens so effizient sind wie die entsprechenden parametrischen Tests (vgl. Bradley, 1968, Kap. 6; Lienert, 1973, 257-262; Marascuilo & McSweeney, 1977, Kap. 11 und 12).

Daneben ist für eine ganze Reihe der häufiger verwendeten nicht-parametrischen Testverfahren gezeigt worden, daß ihre „relative asymptotische Effizienz“ (siehe Pitman, 1948; Bradley, 1968, 56-62; Büning & Trenkler, 1978, 275-282) selbst dann, wenn alle Voraussetzungen für parametrische Tests erfüllt sind, nicht unter einen - häufig nahe bei Eins liegenden - Minimalwert absinkt (vgl. u.a. Hodges & Lehmann, 1956, und Berchtold, 1979).

Die Werte für die relative asymptotische Effizienz findet man insbesondere in den Büchern von Bradley (1968, 60f.), Lienert (1973, 1978), Marascuilo & McSweeney (1977, 87) sowie Büning & Trenkler (1978, 282). Diese Hinweise sind insofern wichtig, als die Teststärkenbestimmung sich bei den meisten Rangverfahren überaus schwierig gestaltet, so daß man in diesen Fällen darauf angewiesen ist, diese unter Verwendung der nicht-zentralen Verteilungen der homologen parametrischen Tests und der Maßzahl der relativen asymptotischen Effizienz ungefähr abzuschätzen. Wie wir bereits mehrfach erwähnt haben, ist die Kenntnis und Kontrolle der Teststärke eine notwendige Voraussetzung für eine strenge Prüfung einer WH.

7.6 Zusammenfassung

Im Teil 6 haben wir herausgearbeitet, daß die Prüfung einer wissenschaftlichen Hypothese i. a. nur über eine aus ihr abgeleitete statistische Hypothese erfolgen kann. Eine der Möglichkeiten, eine statistische Hypothese zu beurteilen, bietet der Signifikanztest. Der vorangegangene Teil 7 diente im wesentlichen der Darstellung einiger Probleme, die mit der üblichen Anwendung von Signifikanztests verbunden sind und die dazu führen, daß die Beurteilung der WH häufig nicht angemessen möglich ist.

Demzufolge wurden einige Modifikationen an der „Anwenderversion“ des Signifikanztests aufgegriffen, die auf die Kontrolle seiner Determinanten hinauslaufen, und es wurde eine vorläufige Zielvorstellung bzgl. der Anwendung von Signifikanztests vorgestellt, die in den Teilen 10 und 11 vertieft werden soll.

Ferner wurde aufgezeigt, daß unterschiedlichen Arten von statistischen Hypothesen auch unterschiedliche Prüfverfahren entsprechen, deren valide Anwendung teils an sehr restriktive und teils an nur schwache Voraussetzungen gebunden ist. Auf die strengen Voraussetzungen zur Anwendung parametrischer Tests gehen wir im Abschnitt 8.2 ein, dem einige Ausführungen über die Beziehungen zwischen der WH und der statistischen Hypothese vorausgehen werden.

8. Störfaktoren der statistischen Validität und ihre Ausschaltung

Um fälschliche Falsifikationen und ungerechtfertigte Bewährungen der wissenschaftlichen Hypothese zu vermeiden, müssen bei der Prüfung der aus ihr abgeleiteten statistischen Hypothesen die Fehlerwahrscheinlichkeiten erster und zweiter Art möglichst gering sein (Teil 6). Alle Faktoren, die letztlich dazu führen, daß diese Fehlerwahrscheinlichkeiten α und β erhöht werden, setzen die statistische Validität der Untersuchung herab. Wir wollen die wichtigsten dieser Störfaktoren der statistischen Validität im folgenden -wieder in Gruppen zusammengefaßt- besprechen und jeweils auch Maßnahmen zu ihrer Vermeidung erörtern. Diese Störfaktoren (StatV) umfassen die u.E. schwerstwiegenden Fehler, die bei der statistischen Auswertung empirischer Untersuchungen (insbesondere Experimente) gemacht werden können und leider auch immer wieder gemacht werden.

8.1 Falsche statistische Hypothesen und Verfahren

8.1.1 Die wichtigsten Beziehungen zwischen psychologischen und statistischen Hypothesen

Eine wissenschaftliche (Kausal-)Hypothese WH kann über die Prüfung einer statistischen Hypothese SH falsifiziert werden, wenn zwischen ihnen eine logische Implikationsbeziehung besteht, d.h. wenn gilt: $WH \rightarrow SH$ (vgl. Teil 6). Wir wollen jetzt verschiedene Arten von psychologischen Kausalhypothesen betrachten und dabei jeweils untersuchen, welche statistischen Hypothesen aus ihnen ableitbar sind und mit welchem statistischen Verfahren diese geprüft werden können. Aus diesen Darstellungen werden sich einige mögliche Fehler ergeben, die die statistische Validität herabsetzen (zu weiteren Einzelheiten siehe Hager & Westermann, im Druck, a).

(1) Aus unserer Beispielhypothese ließ sich ableiten, daß bei Vorliegen von Dissonanz die Werte auf der Einstellungsvariablen höher sind als ohne Dissonanz. Aus WH_u folgte deshalb eine statistische Hypothese wie $H_1: \mu_2 - \mu_1 > 0$. In dieser Beziehung ist WH_u typisch für Kausalhypothesen in der Psychologie: In-der Regel implizieren psychologische Kausalhypothesen *gerichtete* statistische Alternativhypothesen (Bredenkamp, 1972; Hager & Westermann, im Druck, a). Dies hat wichtige Konsequenzen für die Durchführung von Signifikanztests (a. a.o.): Damit die übergeordnete wissenschaftliche Kausalhypothese überhaupt falsifizierbar ist, muß man sich für die Falschheit der statistischen Alternativhypothese (und damit für die Richtigkeit der H_0) entscheiden können. Diese Entscheidung ist aber für den Experimentator nur dann zu verantworten, wenn er die Wahrscheinlichkeit dafür kennt, daß diese Entscheidung falsch ist und wenn diese Fehlerwahrscheinlichkeit ihm klein genug

erscheint. Zur Prüfung von wissenschaftlichen Hypothesen, die eine statistische Alternativhypothese implizieren, muß also auch β auf einen kleinen Wert festgelegt werden.

(2) Läßt sich aus einer wissenschaftlichen Hypothese nur ableiten, daß Mittelwertsunterschiede in irgendeiner Richtung bestehen, folgt aus ihr eine ungerichtete Alternativhypothese wie $H_1 : \mu_2 - \mu_1 \neq 0$. Wissenschaftlich dürften solche unspezifischen Vorhersagen nur von geringem Interesse sein.

(3) Relativ selten sind auch Kausalhypothesen, deren ausschließliche Aussage darin besteht, daß unter bestimmten Bedingungen keine Unterschiede oder Veränderungen zu erwarten sind. Aus ihnen folgen einfache statistische Nullhypothesen wie $H_0 : \mu_1 - \mu_2 = 0$.

(4) In den exakten Naturwissenschaften sind Theorien im allgemeinen so präzise formuliert, daß im Konklusionsteil („dann . . .“) von Kausalhypothesen genaue Angaben über den Wert einer bestimmten Variablen gemacht werden. Aus ihnen folgen dann statistische Nullhypothesen, nach denen ein Parameter einen ganz bestimmten Wert hat, also z.B. $H_0 : \mu = 45$. In der Psychologie sind solch präzise Folgerungen nur in den seltenen Fällen zu erwarten, in denen Theorien nicht nur verbal, sondern auch mathematisch formuliert sind (vgl. Restle & Greeno, 1970; Bredenkamp, 1972, 186-198; Coombs, Dawes & Tversky, 1975; Bredenkamp & Hager, 1979).

In den bisher besprochenen vier Fällen impliziert die wissenschaftliche Hypothese stets eine statistische Hypothese über einen oder zwei Mittelwerte. Kann man bestimmte Annahmen über die Verteilung der Werte in den Populationen machen, können die jeweiligen statistischen Null-Hypothesen über den parametrischen t-Test geprüft werden. Kann oder will man die Verteilungsannahmen nicht aufrechterhalten, sind aus der psychologischen Hypothese andere statistische Hypothesen (z.B. über Mediane) abzuleiten, die über nicht-parametrische Testverfahren geprüft werden können.

Als wichtiges Ergebnis unserer Überlegungen ist festzuhalten, daß in den meisten Fällen der psychologischen Forschungspraxis aus der wissenschaftlichen Hypothese eine gerichtete Alternativhypothese folgt, deren Prüfung über einen einseitigen statistischen Test erfolgen sollte. Zweiseitige Tests sind nur dann gerechtfertigt, wenn von der psychologischen Hypothese - ausnahmsweise - tatsächlich eine ungerichtete Alternativhypothese oder aber eine einfache Null-Hypothese impliziert wird - wir kommen darauf im Teil 11 zurück.

Wenden wir uns nun weiteren Arten von psychologischen Hypothesen zu!

(5) Schon häufiger als die in (4) besprochenen exakten numerischen Vorhersagen treten in der Psychologie Hypothesen auf, die ganz bestimmte *funktionale*

Zusammenhänge zwischen Variablen annehmen (Beispiel: die „psychophysischen Funktionen“). Die aus ihnen ableitbaren statistischen Nullhypothesen werden durch parametrische oder nicht-parametrische Trendanalysen geprüft.

(6) Wissenschaftliche Kausalhypothesen können so formuliert sein, daß sie die Gleichheit mehrerer Parameter implizieren; allerdings ist diese Art von Kausalhypothesen in der Psychologie noch seltener anzutreffen als der unter (3) dargestellte Sonderfall, aus dem die Gleichheit zweier Parameter folgte. In der Mehrzahl der Fälle handelt es sich bei den in Frage stehenden Parametern um Mittelwerte. Deshalb kann die statistische Null-Hypothese wie folgt angegeben werden :

$$H_0: \mu_1 = \mu_2 = \dots = \mu_K.$$

Es handelt sich hierbei um die Null-Hypothese der parametrischen Varianzanalyse. Die zu ihr gehörige Alternativhypothese H_1 besagt, daß zwischen mindestens zwei Mittelwerten ein Unterschied besteht; sie lautet also formal:

$$H_1: \mu_m \neq \mu_{m'}, \text{ für mindestens ein Paar von Bedingungen } X_m \text{ und } X_{m'}, \text{ wobei gilt, daß } m \neq m'.$$

Diese Art von Alternativhypothesen folgt aus psychologischen Kausalhypothesen, die unspezifisch irgendeinen Unterschied in irgendeiner Richtung vorhersagen. Gerade wegen des ausgeprägten Mangels an Spezifität der Vorhersage sind solche wissenschaftlichen Hypothesen nur von geringem Interesse.

Insgesamt gesehen folgen also aus einer wissenschaftlichen Hypothese der Psychologie nur selten die Null- oder Alternativhypothese einer *Varianzanalyse*. Das steht im Gegensatz zur Tatsache, daß bei der großen Mehrheit der veröffentlichten experimentellen Untersuchungen die Auswertung über parametrische und nicht-parametrische Varianzanalysen erfolgt und daß die Bücher zur Versuchsplanung („Experimental Design“) überwiegend varianzanalytische Auswertungstechniken behandeln - siehe dazu die Literaturangaben in den Abschnitten 7.5.3.2 und 7.5.3.3.

Man kann also davon ausgehen, daß das angewendete statistische Verfahren in vielen Fällen gar nicht die Information erbringt, die zur (strengen) Prüfung der betrachteten Kausalhypothese erforderlich ist (Hager & Westermann, im Druck, a).

(7) Psychologische Hypothesen und Theorien sind nur in den seltensten Fällen präzise genug, um die Ableitung numerischer Vorhersagen oder Aussagen über die Form funktionaler Beziehungen vornehmen zu können. Typischer für den Entwicklungsstand der meisten psychologischen Teilgebiete sind wissenschaftliche Hypothesen wie folgende Ergänzung WH_v zu unserer Kausalhypothese WH_u : „Je stärker die auftretende Dissonanz ist, desto größer ist die

Einstellungsänderung in Richtung auf die dissonanzerzeugende Kognition.“ Aus derartigen wissenschaftlichen Hypothesen folgt stets eine bestimmte *Rangordnung von Mittelwerten*, also z. B.:

$$H_1: \mu_2 < \mu_3 < \dots < \mu_K$$

mit der dazugehörigen Null-Hypothese

$$H_0: \mu_m \geq \mu_{m'} \text{ für mindestens ein Paar } m \text{ und } m', \text{ mit } m < m'.$$

Die Planung von Experimenten zur Prüfung solcher Hypothesen über einen monotonen Trend wirft einige schwierige Probleme auf, die wir im Abschnitt 8.3.2 ansprechen werden.

(8) Aus psychologischen Theorien können häufig Hypothesen abgeleitet werden, die einander ergänzen wie die oben besprochenen WH_u und WH_v . In diesem Fall können beide simultan in einem einzigen Experiment überprüft werden. Dazu müssen - um im Beispiel zu bleiben - eine experimentelle Bedingung „Keine Dissonanz“ (X_1) und mindestens zwei Treatments mit ansteigender Stärke der Dissonanz ($X_2, X_3, \dots, X_K$) hergestellt werden. Die Hypothese WH_u impliziert dann eine statistische Hypothese wie die folgende $H_{1(1)}$:

$$H_{1(1)}: \frac{1}{K-1} \left\{ (\mu_2 + \mu_3 + \dots + \mu_K) \right\} > \mu_1;$$

die Hypothese WH_v dagegen impliziert die $H_{1(2)}$:

$$H_{1(2)}: \mu_2 < \mu_3 < \dots < \mu_K.$$

Auf die Prüfung solcher (u.U. hierarchisch geordneter) Mengen von Hypothesen gehen wir im Abschnitt 8.3.1 ein.

(9) Lassen sich aus einer psychologischen Theorie mehrere statistische Hypothesen ableiten, die sich auf die Werte der gleichen AV unter den Ausprägungen verschiedener UV beziehen, können diese Hypothesen über mehrfaktorielle Designs simultan überprüft werden. Diese Designs gestatten auch die Prüfung von Hypothesen über die Wechselwirkung (Interaktion) der unabhängigen Variablen in bezug auf die abhängige Variable. Eine statistische Interaktionshypothese würde z.B. aus einer wissenschaftlichen Hypothese abzuleiten sein, die folgendes aussagt: „Kognitive Dissonanz führt zu einer Einstellungsänderung in Richtung auf die Information aus einer glaubwürdigen Quelle, aber in entgegengesetzter Richtung bei Information aus einer unglaublichen Quelle.“ Derartigen Interaktionshypothesen wird ein zunehmendes Interesse entgegengebracht. Auf die mit ihrer Prüfung verbundenen Probleme gehen wir im Abschnitt 8.5 ein.

Wir haben in den neun Punkten die wichtigsten Grundformen psychologischer Kausalhypothesen dargestellt, und wir haben erörtert, welche Arten von statistischen Hypothesen jeweils aus ihnen folgen. Ferner sind wir darauf eingegangen, mit welchen Verfahren die statistischen Hypothesen geprüft werden können. Aus dieser Diskussion ergeben sich unmittelbar drei mögliche Beeinträchtigungen der statistischen Validität (StatV) einer Untersuchung, denen die Abschnitte 8.1.2 bis 8.1.4 gewidmet sind.

8.1.2 Falsche Umsetzung der wissenschaftlichen in eine statistische Hypothese als Störfaktor (StatV)

Die Prüfung einer statistischen Hypothese SH kann nur dann der strengen Prüfung einer wissenschaftlichen Kausalhypothese WH dienen, wenn SH tatsächlich ein Implikat von WH ist. Dieser Tatbestand wird in der Forschungspraxis sehr häufig nicht beachtet; dies äußert sich bspw. darin, daß eine Varianzanalyse durchgeführt wird, obwohl aus der psychologischen Hypothese statistische Hypothesen über die Ordnung zweier oder mehrerer Mittelwerte ableitbar sind (vgl. die Punkte (1), (6) und (7) im Abschnitt 8.1.1). Diese Vorgehensweise setzt die Strenge einer Prüfung herab, wie wir am Beispiel aufzeigen wollen (vgl. auch Hager & Westermann, im Druck, a): Aus den im Punkt (7) besprochenen psychologischen Hypothesen WH_u und WH_v folgt die statistische Hypothese:

$$H_1: \mu_1 < \mu_2 < \mu_3, \text{ mit } H_0: \mu_1 \geq \mu_2 \text{ und/oder } \mu_2 \geq \mu_3.$$

Zum Vergleich: Die entsprechenden Hypothesen einer einfachen Varianzanalyse lauten

$$H'_0: \mu_1 = \mu_2 = \mu_3 \text{ und } H'_1: \mu_1 \neq \mu_2 \text{ und/oder } \mu_2 \neq \mu_3.$$

Die Alternativhypothese der Varianzanalyse umfaßt also auch „wahre“ Parameterverhältnisse, die im Widerspruch zur psychologischen Hypothese stehen (beispielsweise $\mu_1 > \mu_2 = \mu_3$). Wird also fälschlicherweise aus den zu prüfenden wissenschaftlichen Hypothesen WH_u und WH_v statt der obigen H_1 die Alternativhypothese einer Varianzanalyse abgeleitet, so ist (unter sonst gleichen Bedingungen) für den Fall der Falschheit von WH_u und WH_v die Wahrscheinlichkeit hypothesenkonträrer Ergebnisse geringer. Durch inadäquate Ableitung statistischer aus wissenschaftlichen Hypothesen wird also die Strenge des Prüfexperiments herabgesetzt (a. a. 0.).

8.1.3 Falsche Auswahl der zu prüfenden statistischen Hypothese

In der Regel können aus einer psychologischen Hypothese mehrere statistische Hypothesen abgeleitet werden. Die Prüfung der wissenschaftlichen Hypothese ist dann am strengsten, wenn sie über diejenige der aus ihr ableitbaren statistischen Hypothesen erfolgt, bei der die Wahrscheinlichkeiten für fälschliche Bewährungs- und fälschliche Falsifikationen am geringsten ist (siehe im einzelnen Hager & Westermann, im Druck, a).

Verdeutlichen wir dies an zwei Beispielen!

- (1) Folgt aus der HW eine Rangordnung von Mittelwerten, ist unter sonst gleichen Bedingungen eine empirische Prüfung um so strenger, je mehr Mittelwerte experimentell verglichen werden.
- (2) Die Prüfung psychophysischer Trendhypothesen an den Daten jeder einzelnen Person ist strenger als eine Prüfung an Mittelwerten von verschiedenen Personen - vgl. zu Einzelheiten Bredenkamp (1980, 12, 49f.).

8.1.4 Falsche statistische Analyse

Hat man aus der wissenschaftlichen Hypothese eine adäquate statistische Hypothese abgeleitet, kann die Validität einer Untersuchung noch durch die falsche Wahl der statistischen Prüfverfahren herabgesetzt werden. Dieser Fall tritt bspw. dann ein, wenn sich nach dem Experiment herausstellt, daß bestimmte Voraussetzungen bzgl. der validen Anwendung des betr. Tests aufgrund der Daten nicht aufrechterhalten werden können, und wenn dann dieser Test dennoch zur Anwendung gelangt - vgl. hierzu im einzelnen Abschnitt 8.2. Ferner werden sehr häufig statistische Tests benutzt, mit denen nur eine ungerichtete Alternativhypothese beurteilt werden kann, obwohl aus der WH eindeutig eine gerichtete Alternativhypothese folgt (siehe Abschnitt 8.1.1, Punkt (6)). Diese oft *allerdings unvermeidliche* Verfahrensweise hat zur Folge, daß der Ablehnungsbereich in der einzig interessierenden Richtung um die Hälfte zu gering gewählt wird, wodurch zwangsläufig die Wahrscheinlichkeit für einen Fehler 2. Art mehr als verdoppelt wird, wenn wir davon ausgehen, daß tatsächlich die aus der WH abgeleitete statistische Alternativhypothese zutrifft. Durch die somit insgesamt erhöhten Fehlerwahrscheinlichkeiten bei der Entscheidung über die statistische Hypothese wird auch die Wahrscheinlichkeit für eine falsche Entscheidung über die wissenschaftliche Hypothese erhöht.

Weitere in diesem Zusammenhang mögliche Fehlerquellen ergeben sich aus den Ausführungen in den Abschnitten 7.5, 8.2, 8.4.1 und 8.4.6; siehe ferner Lindquist (1953, 7f.) sowie Mosteller (1968, 122).

8.2 Verletzung der Annahmen bei statistischen Tests als Störfaktor (StatV)

Die Fehlerwahrscheinlichkeiten bei der statistischen Hypothesenprüfung und damit auch die Wahrscheinlichkeiten für falsche Entscheidungen über das Zutreffen oder Nicht-Zutreffen der wissenschaftlichen Hypothesen können auch dadurch ansteigen, daß die für die Anwendung der inferenzstatistischen Verfahren notwendigen Voraussetzungen oder Annahmen nicht erfüllt sind.

Während die valide Anwendung von nicht-parametrischen Auswertetechniken häufig nur an vglw. schwache Voraussetzungen gebunden ist, beruhen die parametrischen Tests stets auf mehreren restriktiven Annahmen resp. Voraussetzungen. Wegen der bereits wiederholt angesprochenen zentralen Bedeutung, die den parametrischen Verfahren in der Praxis zukommt, wollen wir diese Voraussetzungen im folgenden etwas ausführlicher darstellen und die Konsequenzen ihrer Verletzung erörtern.

8.2.1 Das Allgemeine Lineare Modell (ALM) und die Annahmen

Die meisten parametrischen Hypothesen, die der empirisch arbeitende Wissenschaftler einer Prüfung zu unterziehen beabsichtigt, beziehen sich auf Mittelwerte, Varianzen und Korrelations- bzw. Regressionskoeffizienten. Die zur Prüfung benutzten Test-Statistiken sind in der Regel entweder t-, χ^2 - oder F-verteilt.

Diese Testverfahren lassen sich allesamt aus einem einzigen grundlegenden Modell ableiten, dem sog. „Allgemeinen Linearen Modell“ (ALM). Dieses ist durch die folgenden Eigenschaften zu kennzeichnen (vgl. dazu im einzelnen Moosbrugger, 1978, 57; ferner Fennessey, 1968, 3; Auslitz, Hesse & Rieder, 1975, 3; Schach & Schäfer, 1978, 5):

- (1) $Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \dots + \beta_m X_{mi} + \dots + \beta_K X_{Ki} + e_i$
- (2) Der Mittelwert der Fehler ist in jeder der Populationen gleich Null.
- (3) Die Fehler e_i sind innerhalb der experimentellen Bedingungen wie auch zwischen ihnen unabhängig voneinander.
- (4) Die Varianz σ_e^2 der Fehler e_i ist in jeder der K Populationen gleich oder homogen:

$$\sigma_{e_1}^2 = \sigma_{e_2}^2 = \dots = \sigma_{e_K}^2 = \sigma_e^2$$

Die Merkmalsausprägung jeder Vp i auf der abhängigen oder Kriteriumsvariablen Y, also Y_i , wird im ALM also dargestellt als Summe der mit β_k gewichteten Werte X_k der unabhängigen oder Prädiktorvariablen („Einflußgrößen“) X und einer Zufallsgröße e_i , dem unsystematischen Fehler oder Residuum. Die Werte X_k sind (vom E) festgelegte Größen, von denen X_0 in der Regel gleich

Eins ist; die „Gewichtsgrößen“ β_k bezeichnen der Größe nach unbekannte Parameter, die aus den Daten nach dem Kriterium der „Kleinsten Quadrate“ zu bestimmen sind (vgl. dazu den Satz von Gauss-Markoff, etwa in Scheffé, 1959, 13-19; oder in Menges, 1972, 319).

Zur Prüfung von Hypothesen vor dem Hintergrund des ALM können zwei im wesentlichen formal unterschiedliche Wege beschritten werden.

Wertet man die empirischen Daten aus $K \geq 2$ Stichproben über einen Simultanvergleich der K Mittelwerte aus, spricht man traditionell von einer „Varianzanalyse“ (VA). (Vgl. Fisher, 1925, 1950; ferner Scheffé, 1959; Hays, 1963, 1977; Cochran & Bliss, 1970, sowie die in Abschnitt 7.5.3 angegebene Literatur.)

Bei der Prüfung der Mittelwerts-Hypothese wird davon ausgegangen, daß die K Modalitäten des Faktors vom E bewußt ausgewählt und festgelegt worden sind, daß m.a. W. sog. „fixierte Effekte“ vorliegen (vgl. Abschnitt 2.3). Verwendet der E dagegen in seinem Experiment eine Zufallsstichprobe der Größe K aus einer Population möglicher Abstufungen der UV, so liegen „zufällige Effekte“ vor; mit dem Experiment werden statistische Hypothesen geprüft, die sich auf die Varianz dieser Effekte beziehen. Werden zufällige und fixierte Effekte in einem Experiment simultan untersucht, spricht man von „gemischten Effekten“.

Diese Unterscheidungen sind wichtig, weil ihnen unterschiedliche varianzanalytische Modelle und F-Tests entsprechen - vgl. zu Einzelheiten insbesondere Eisenhart (1947), Wilk & Kempthorne (1955) oder Hays (1977, 377f.). Wir beziehen uns im folgenden stets auf das varianzanalytische Modell der fixierten Effekte - zur Begründung siehe die Abschnitte 2.3 und 8.2.3.

Prüft man dagegen Hypothesen über (multiple) Korrelationsquadrate $R_{Y.X_k}^2$ (resp. über Regressionskoeffizienten β_k), nennt man diese Art der Auswertung „Multiple Regressions- und Korrelationsanalyse“ (MRA); zur formalen Unterscheidung siehe im einzelnen etwa Auslitz, Hesse & Rieder (1975, 7f.) sowie Schach & Schäfer (1978, 6).

Die oben erwähnte VA mit fixierten Effekten ist als Spezialfall der allgemeineren (und im ganzen vielseitigeren) MRA darstellbar; Einzelheiten hierzu entnehme man etwa Jennings (1967), Cohen (1968), Darlington (1968), Fennessey (1968), Wottawa (1974), Auslitz, Hesse & Rieder (1975), Schach & Schäfer (1978) sowie der bereits im Abschnitt 7.5.3.3 genannten Literatur; zu bestimmten konzeptuellen Unterschieden siehe etwa Witte (1978, 1980, 172f.).

Diese Auswerteverfahren lassen sich ihrerseits unter dem allgemeinen Modell der „Kanonischen Analyse“ subsumieren, wie Knapp (1978) gezeigt hat.

Zur validen Anwendung der o. gen. Tests und Prüfverteilungen sind neben den bereits unter (1) bis (4) aufgeführten Annahmen noch die folgenden Annahmen zu treffen (vgl. dazu auch Eisenhart, 1947; Cochran, 1947; Lindquist,

1953, 72-78; Gaito, 1959b; Scheffé, 1959, Kap. 10; Moosbrugger, 1978, 69; sowie einführende Lehrbücher der Versuchsplanung und -auswertung):

- (5) Die Terme e_i sind in jeder der im Experiment untersuchten Populationen normalverteilt:
- (6) Die zu untersuchenden Rohwerte stellen Zufallsstichproben aus den interessierenden Grundgesamtheiten dar.

Bei der Darstellung der sechs Voraussetzungen ist ein univariater einfaktorieller Versuchsplan mit jeweils mehreren Vpn pro experimenteller Bedingung zugrunde gelegt worden; die Voraussetzungen beziehen sich jedoch in modifizierter und/oder erweiterter Form auch auf alle anderen Varianz- und regressionsanalytischen Designs. Auf Erweiterungen dieser Voraussetzungen gehen wir im Zusammenhang mit der Kovarianzanalyse (Abschnitt 8.4.2) und mit wiederholten Messungen (Abschnitt 8.4.6) ein, weil diesen Verfahren in der Praxis eine große Bedeutung zukommt; zu weiteren Einzelheiten verweisen wir auf einschlägige Lehrbücher - siehe dazu etwa Abschnitt 7.5.3.3.

Diese Voraussetzungen sind zur validen Hypothesenprüfung unerlässlich. Zu ihnen treten noch zwei Restriktionen, die mit der Hypothesenprüfung jedoch nicht im Zusammenhang stehen:

- (7) Die sog. „Reparametrisierungsbedingung“ ist erforderlich, um zu mathematisch eindeutigen Lösungen der sog. „Normalgleichungen“ gelangen zu können. Sie besagt für das varianzanalytische Modell der fixierten Effekte (und damit auch für die MRA), daß - in varianzanalytischer Terminologie - die Summe der (einfachen) Abweichungen der Treatment-Mittelwerte μ_k von ihrem Gesamtmittelwert μ für jeden Faktor gleich Null sein muß.

Für den Praktiker ist diese Restriktion meist ohne Belang, weswegen wir sie nicht erörtern; Einzelheiten sind etwa Mendenhall (1968, Kap. 6), Searle (1971 a, 209-220), Glass, Peckham & Sanders (1972, 241) und Rasch et al. (1978, 65-67) zu entnehmen.

- (8) Die empirische AV soll mindestens intervallskaliert sein, damit die Ergebnisse der parametrischen Auswertung über eine VA oder MRA sinnvoll interpretierbar sind - siehe dazu Abschnitt 2.4.

Im Zusammenhang mit den sechs Voraussetzungen wollen wir im folgenden jeweils drei Fragen nachgehen:

Wie können diese Voraussetzungen geprüft werden?

Ist eine derartige Prüfung notwendig und/oder sinnvoll?

Welche Konsequenzen ergeben sich bei Verletzung der jeweiligen Voraussetzung?

8.2.2 Additivität

Im Rahmen der auf dem ALM basierenden Auswertetechniken und insbesondere der „klassischen“ VA wird der Terminus „Additivität“ in zwei Bedeutungen benutzt (vgl. u.a. Gaito, 1959b; Lee, 1961; Glass, Peckham & Sanders, 1972; Henning, 1978).

- (1) Zum einen bezieht er sich auf die additive Verknüpfung der Komponenten, die jeden einzelnen Rohwert konstituieren; etwa:

$$(8.1) \quad \begin{aligned} Y_{ik} &= \mu_k + \beta_k + e_{ik} \text{ (einfaktorielle VA)} \\ &= \beta_0 + \beta_1 X_1 + e_{ik} \text{ (regressionsanalytische Darstellung);} \end{aligned}$$

oder für den Fall einer univariaten bifaktoriellen Versuchsanlage mit zwei UV A und B sowie einer AV Y:

$$(8.2) \quad \begin{aligned} Y_{ijk} &= \mu + \beta_k + \alpha_j + (\alpha\beta)_{jk} + e_{ijk} \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e_{ijk} \end{aligned}$$

(In der varianzanalytischen Darstellung, die in den jeweils oberen Zeilen der Gleichungen (8.1) und (8.2) gewählt wurde, kennzeichnen die β_k den Populationseffekt zu Lasten der Treatment-Modalität B_k , also $\beta_k = \mu_k - \mu$; in den jeweils unteren Zeilen derselben Formeln bezeichnen die β_k die Populations-Regressions- bzw. Gewichtungskoeffizienten (vgl. detaillierter etwa Fennessey, 1968; Searle, 1971 a; Wottawa, 1974; Gaensslen & Schubö, 1976; Moosbrugger, 1978; Henning, 1978; Henning & Muthig, 1979).)

Die Regressionskoeffizienten β_k treten in den drei angegebenen grundlegenden Gleichungen stets nur in der 1. Potenz auf, weswegen man vom „Allgemeinen Linearen Modell“ spricht. Die Größen X_k können dabei in jeder beliebigen Potenz oder multiplikativen Verknüpfung auftreten; hierdurch wird die Form der sog. „Response Surface“ (auch: „Response Curve“) bestimmt - siehe dazu Box (1968), Bliss (1970, 427-433), John (1971, Kap. 10), Snedecor & Cochran (1972, 346-358) und John & Quenouille (1977, Kap. 9).

Die (wichtige) Unterscheidung zwischen der Modellform, die die Linearität der Y_{ik} als Funktion der unbekannten Regressionsgewichte β_k postuliert (Mendenhall, 1968, 55), und der Bedeutung der & für die Response Curve resp. Surface wird nicht immer beachtet, wie der Artikel von Brown (1975) zeigt; vgl. zu dieser Unterscheidung im einzelnen Mendenhall (1968, 51-55) und Henning (1978).

Die gebräuchlichen additiven Modellgleichungen repräsentieren dabei das einfachste Prinzip, nach dem man sich empirische Scores Y_{ik} zusammengesetzt denken kann. Es muß sich dabei nicht immer auch gleichzeitig um das

beste oder auch nur um ein adäquates Prinzip handeln, wie bereits Fisher & MacKenzie (1923, 315) ausgeführt haben. über andere, allerdings kaum gebräuchliche Möglichkeiten informieren etwa Searle (1971 a, 75 f.), Glass, Peckham & Sanders (1972, 240f.) sowie Namboodiri, Carter & Blalock (1975, 279f.).

Die Angemessenheit dieser Art von Modellgleichung ist im konkreten Einzelfall ebenso wenig empirisch prüfbar wie die entsprechende Modellannahme im Rahmen der klassischen Testtheorie (vgl. dazu Lord & Novick, 1968; Fischer, 1974; Kranz, 1979; Wottawa, 1980).

- (2) In mehrfaktoriellen Plänen spricht man daneben dann häufig von „Additivität“, wenn der Term $(\alpha\beta)_{jk}$ oder $\beta_3 X_1 X_2$ aus Formel (8.2) sowie alle Interaktionsterme höherer Ordnung gleich Null sind (vgl. Elston, 1961; Namboodiri, Carter & Blalock, 1975, 279-285). In bestimmten Versuchsanordnungen ist die Annahme, daß alle Interaktionsterme in den Populationen gleich Null sind, notwendig, um die interessierenden experimentellen Effekte auf statistische Signifikanz prüfen zu können (vgl. dazu neben Abschnitt 8.4.3 auch Wilk & Kempthorne, 1957; Scheffé, 1959, Kap. 5; Winer, 1971, 398f., 696f.; darüber hinaus weitere Lehrbücher der Versuchsplanung und -auswertung).

Diese Art der Additivität ist häufig einer statistischen Prüfung zugänglich. In den Fällen, in denen eine von der Interaktion unabhängige Fehler- und Prüfvarianz bestimmt werden kann (bspw. im Zufallsgruppenversuchsplan; vgl. dazu Abschnitt 3.3.2), erfolgt die Prüfung der Interaktion (oder Additivität) auf statistische Signifikanz unter Verwendung eben dieser Fehlervarianz. Für einige der Versuchspläne, in denen die die Interaktion enthaltende Residualvarianz die Prüfgröße darstellt, sind spezielle Testverfahren zur Prüfung der Additivitätsannahme entwickelt worden (Tukey, 1949; Johnson & Graybill, 1972). Die Leistungsfähigkeit dieser Tests ist u.a. von Hegemann & Johnson (1976) untersucht worden; Hinweise auf die Anwendungsbereiche finden sich etwa in den Arbeiten von Scheffé (1959, 130-134), Bliss (1967, 324-330, 451-465), Winer (1971, 394-397, 473-478), Bortz (1979, 398-400).

Ergibt die Anwendung eines dieser Testverfahren ein signifikantes Resultat, das für das Vorliegen von Nicht-Additivität oder das Vorhandensein einer Interaktion spricht, wird eine spezifische Voraussetzung für die Signifikanzprüfung im Rahmen der o. gen. Pläne über die F-Verteilung verletzt.

Als Ausweg wird in diesem Fall von zahlreichen (Lehrbuch-)Autoren empfohlen, die numerischen Werte der nicht-linear derart zu transformieren, daß die Wechselwirkung (Interaktion oder Nicht-Additivität) eliminiert wird - siehe z.B. Edwards (1971, 191f.) und Winer (1971, 398).

Dieser Empfehlung können wir uns nicht anschließen - zur Begründung siehe Abschnitt 8.2.4.2 -, sondern halten in der genannten Situation alternative Auswerteverfahren für eher angemessen - vgl. dazu Abschnitt 7.5.3.

8.2.3 Normalverteilung der Modellresiduen (Fehler)

In der Varianz- und Regressionsanalyse wird unter „Fehler“ derjenige Anteil am Score einer Vp verstanden, der in beliebiger Richtung vom Mittelwert der Gruppe abweicht, oder: der übrigbleibt, wenn den Daten eines Treatments ein bedingungspezifischer Wert („fit“) angepaßt worden ist. Formal kann man für den einfachsten Fall etwa schreiben:

(8.3) Fehler = Residuum = beobachteter Wert - angepaßter Wert.

Inhaltlich wird unter diese Fehler alles das subsumiert, was nicht auf den Einfluß der UV(n) zurückgeführt werden kann, also insbesondere individuelle Unterschiede (Persönlichkeitsmerkmale, -variablen; vgl. Herrmann, 1973, 1976) und ggf. Meßfehler (vgl. dazu Cochran, 1968b, 1970; Lord & Novick, 1968); es sei ergänzend nur angemerkt, daß gerade diese sog. Fehler unter anderen Fragestellungen der Gegenstand des Forschungsinteresses sind (siehe Herrmann, 1976). Wir gehen auf die eher inhaltlichen Aspekte der Fehlerterme nicht weiter ein, sondern verweisen auf die speziellen Arbeiten von Underwood (1957, Kap. 4 und 5), Cox (1961), Lord & Novick (1968), Cochran (1968b), Elashoff (1968), Mosteller (1968), Bredenkamp (1969a), Namboodiri, Carter & Blalock (1975, Kap. 12 und 13) sowie Henning & Muthig (1979, e.g. 35-37, 105-107).

Die Normalverteilung der individuellen Fehlerterme e_i in den einzelnen Populationen wird im Grunde vorausgesetzt, weil die Klasse von „normal“ genannten mathematischen Funktionen Eigenschaften aufweist, die zu Ableitungen von „bestechender Eleganz“ (Menges, 1972, 248) und Einfachheit führen; zu weiteren Gründen siehe Menges (a. a.o.). Bspw. sind Mittelwert μ und Varianz σ^2 dieser Verteilung stochastisch unabhängig voneinander (Geary, 1936), weswegen sich die Ableitung der Stichprobenverteilung des Mittelwerts M und der Varianz S^2 vglw. problemlos gestaltet (vgl. Hays, 1977, Kap. 8 und 11).

Die bekannteren der zur Prüfung der Normalitätsannahme entwickelten Tests lassen sich grob in zwei Gruppen einteilen (vgl. dazu Rasch, Enderlein & Herrendörfer, 1973, Kap. 6), nämlich in Tests für die Anpassungsgüte (dazu u.a. Hays, 1977, Kap. 17; Bortz, 1979, 191-194) und in Tests für die Kennwerte „Schiefe“ („skewness“) und „Exzeß“ („kurtosis“) (dazu u.a. Gebhardt, 1966; Schmidtke & Jäger, 1976).

Die angesprochenen und verschiedene weitere, in der Literatur empfohlene Testverfahren sind in Simulationsstudien etwa von Shapiro, Wilk & Chen (1968) sowie von Saniga & Miles (1979) untersucht worden - auf diese Arbeiten und die dort enthaltenen Literaturangaben sei verwiesen.

Wir gehen auf diese Verfahren zur Prüfung der Normalverteilungsannahme nicht im einzelnen ein, da ihre routinemäßige Anwendung aus (mindestens) zwei Gründen kontraindiziert scheint.

Zum einen sind die entsprechenden Tests meist selbst abhängig von bestimmten Voraussetzungen oder Annahmen (vgl. etwa Box, 1953; Scheffé, 1959, 362; Gaensslen & Schubö, 1976, 58f.).

Zum anderen muß in fast jedem Fall mit einem signifikanten Resultat gerechnet werden, das für bedeutsame Abweichungen der Modellresiduen von der Normalität spricht, weil es nur wenige Gründe zu der Vermutung gibt, daß die Fehler sich *exakt* nach der „normal“ genannten Funktion verteilen: „Normality is a myth: there never was, and never will be, a normal distribution.“ (Geary, 1947, 241)

Wie Menges (1972, 249) ausführt, trifft diese Feststellung nicht notwendigerweise den „Kern der Dinge“: „... denn nicht nach der Existenz der Normalverteilung ist sinnvoll zu fragen, sondern nach dem Typ von empirischer Situation, für die das Modell, das sie repräsentiert, eine gute Approximation liefert.“ Demnach ist es zweckmäßig, die übliche Frage „Are normal theory ANOVA assumptions met?“ durch eine andere zu ersetzen, nämlich „How important are the inevitable violations of normal theory ANOVA assumptions?“ (Glass, Peckham & Sanders, 1972, 237).

Die auf eine „Ja-Nein-Antwort“ abzielende Frage wird also durch eine ersetzt, die nach der (qualitativ stufbaren) „Robustheit“ der gebräuchlichen parametrischen Testverfahren fragt. Der Fachausdruck „Robustheit“ geht auf Box (1953, 318) zurück und bezeichnet die „Unempfindlichkeit (eines Testverfahrens) gegenüber Abweichungen von den postulierten Modellannahmen“ (vgl. Büning & Trenkler, 1978, 296).

In Abschnitt 7.5.2 hatten wir angesprochen, daß die Normalverteilung der Fehler eine der Voraussetzungen zur Ableitung der Stichprobenverteilungen von t , χ^2 und F darstellt. Ist diese Voraussetzung nicht erfüllt, resultieren Verteilungen dieser Teststatistiken, die von den tabellierten zentralen und nicht-zentralen Verteilungen abweichen. Dies bedeutet, daß die tatsächlichen Wahrscheinlichkeiten für bestimmte Resultatsklassen nicht mehr mit den in den Tabellen angegebenen übereinstimmen, m.a. W. weichen die tatsächlichen Werte für α und β von den festgesetzten nominellen Werten ab. Ein Test ist demnach robust, wenn unter Verletzung einer oder mehrerer Voraussetzungen

die tatsächlichen Werte für α und β „nicht wesentlich“ von den nominellen abweichen - vgl. zum Versuch der quantitativen Definition von Robustheit die instruktive Arbeit von Bradley (1978); zum Hinweis auf einige Probleme beim ausschließlich qualitativen Robustheitskonzept siehe u.a. Büning & Trenkler (1978, 296f.) und Hager, Lübbecke & Hübner (im Druck).

Zur Beantwortung der Frage nach der Robustheit der klassischen parametrischen Teststatistiken t , χ^2 und F unter Abweichung der Verteilung der Modellresiduen von der Normalität sind zahlreiche Untersuchungen durchgeführt worden (vgl. bereits Pearson, 1931; zusammenfassend Cochran, 1947; Lindquist, 1953, 78-90; Gaito, 1959b; Scheffé, 1959, Kap. 10; Bradley, 1968, 24-43; Glass, Peckham & Sanders, 1972; darüber hinaus im einzelnen u.a. Tikau, 1971; Kemp & Conover, 1973; Bevan, Benton & Myers, 1974; Feir-Walsh & Toothaker, 1974; Havlicek & Peterson, 1974; Lee, Desu & Gehan, 1975; Pearson & Please, 1975; Bowman, Beauchamp & Shenton, 1977; Posten, 1978; Trachtman, Giambalvo & Dippner, 1978; Bradley, 1980a, b, c; Blair, Higgins & Smitley, 1980).

Diese Studien an einfachen univariaten Versuchsplänen enthalten eine Fülle von Hinweisen darauf, daß die o.a. parametrischen Teststatistiken über einen weiten Bereich der Abweichungen von der Normalität der Fehlerverteilungen in den untersuchten Populationen robust sind, sofern sie zur Prüfung von Hypothesen über Mittelwerte oder quadrierte Korrelationsquadrate herangezogen werden (Modell der fixierten Effekte). Bei diesen Tests kann in aller Regel davon ausgegangen werden, daß die tatsächlichen Raten für Fehler 1. und 2. Art *nicht nennenswert* von den a priori spezifizierten nominellen Werten abweichen. Allerdings ist eine Quantifikation dieser Aussage etwa sensu Bradley (1978) u.E. derzeit noch nicht möglich, da die Befunde im Detail noch kein einheitliches Bild ergeben (siehe auch Büning & Trenkler, 1978, 296).

Im einzelnen ist mit *Überschreitungen* der nominellen Fehlerwahrscheinlichkeiten bei „breitgipfligen“ („platykurtischen“) Verteilungen zu rechnen, bei denen sich die Daten in den beiden Extrembereichen häufen; sehr ausgeprägt sind diese Abweichungen bei Stichproben aus L-förmigen Populationen, wie sie anscheinend Reaktionszeiten oft zugrundeliegen - siehe hierzu ausführlich Bradley (1968, 1977, 1978, 1980a, b, c).

Mit *Unterschreitungen* der nominellen Fehlerwahrscheinlichkeiten muß häufig bei „spitzgipfligen“ („leptokurtischen“) Verteilungen gerechnet werden. Abweichungen von der Symmetrie der Verteilungen wirken sich i. a. vornehmlich bei einseitigen Signifikanztests aus.

Komplexe Versuchspläne sind bislang nicht in der gleichen Ausführlichkeit und Systematik untersucht worden wie einfache Designs, weswegen die Robustheit von F und t für diese Fälle noch nicht beurteilbar ist.

Im Rahmen des varianzanalytischen Modells der zufälligen Effekte wird die Annahme normal-verteilter Treatment-Effekte gemacht, und der F-Test prüft eine Null-Hypothese bzgl. der Varianz dieser Effekte. Bei diesem Test haben insbesondere Abweichungen von der die Treatment-Effekte betreffenden Normalitätsannahme schwerwiegende Konsequenzen bzgl. der tatsächlichen Fehlerwahrscheinlichkeiten - vgl. hierzu etwa Lehmann (1968, 45), Glass & Stanley (1970, 462) sowie Hays (1977, 540f.).

Üblicherweise wird dann, wenn vermutet wird oder bekannt ist, daß die Fehler e_i in den Populationen nicht normal-verteilt sind, alternativ eine der beiden folgenden Empfehlungen gegeben:

1. Anwendung von nicht-parametrischen Verfahren.

Im Anschluß an unsere Erörterungen zu diesen Verfahren im Abschnitt 7.5.3.2 schließen wir uns dieser Empfehlung unter der Voraussetzung an, daß die Abweichungen von der Normalität sehr ausgeprägt sind oder aber daß außer der Normalitätsannahme noch mindestens eine weitere Voraussetzung nicht erfüllt ist.

2. Anwendung von Transformationen, die die Stichprobendaten „normalisieren“. Dieser Empfehlung können wir uns nicht anschließen - zur Begründung siehe Abschnitt 8.2.4.2.

8.2.4 Homogenität der Fehlervarianzen in den Populationen

Eine weitere Voraussetzung für die valide Anwendung von F-Tests besteht in der Annahme der Varianzhomogenität, die besagt, daß die Modellresiduen in allen untersuchten Populationen gleiche (homogene) Varianzen aufweisen müssen.

Die Auswirkungen von Verletzungen dieser Voraussetzung auf das Verhalten der parametrischen Teststatistiken ist ebenfalls sehr ausführlich untersucht worden; entsprechende Zusammenfassungen der theoretischen und empirischen Resultate geben Cochran (1947), Scheffé (1959, Kap. 10), Lindquist (1953, 78-90), Box (1954a, b), Glass, Peckham & Sanders (1972); neuere Untersuchungen sind durchgeführt worden von Brown & Forsythe (1974a), Kohr & Garnes (1974), Feir-Walsh & Toothaker (1974), Ekbohm (1976), Keselman, Rogan & Feir-Walsh (1977), Havlicek & Peterson (1974), Howell & Garnes (1973), Rogan & Keselman (1977) und Hager et al. (im Druck).

Diesen Studien läßt sich entnehmen, daß in den meisten Fällen Varianzheterogenität tolerabel ist hinsichtlich der interessierenden tatsächlichen Werte für die Fehlerwahrscheinlichkeiten 1. und 2. Art, sofern die Stichprobengrößen für alle Bedingungen gleich sind. Nennenswerte Ausnahmen von dieser allgemeinen Aussage sind nur unter sehr extremen Bedingungen zu erwarten (vgl. Box, 1954a; Scheffé, 1959, 340; Rogan & Keselman, 1977).

Sind dagegen die Fehlervarianzen wie auch die Stichprobenumfänge ungleich, ist generell mit stärkeren Abweichungen der realen von den nominellen Fehlerwahrscheinlichkeiten zu rechnen (vgl. die o. a. Literatur).

Liegt dieser Fall vor, empfiehlt sich die Verwendung alternativer Testverfahren, etwa nicht-parametrischer (Abschnitt 7.5.3) oder spezieller parametrischer, auf die wir anschließend kurz eingehen werden. Zunächst wollen wir jedoch die Frage zu beantworten suchen, wie Varianzheterogenität festgestellt werden kann und ob dies sinnvoll ist.

Zur Prüfung einer Hypothese bzgl. der Gleichheit von Varianzen sind spezielle Testverfahren u.a. von Bartlett (1937), Hartley (1950) und Cochran (1951) entwickelt worden. Diese Verfahren sind jedoch ihrerseits empfindlich gegenüber Verletzungen der Normalverteilungsannahme - siehe dazu Pearson (1931), Box (1953), Scheffé (1959, 83-87), Levene (1960), Overall & Woodward (1974), Levy (1975a, 1978a) sowie den vorangegangenen Abschnitt.

Diese Tatsache hat zu einer Fülle von Versuchen geführt, Homogenitätsprüfungen für Varianzen zu entwickeln, die robust sind gegenüber Abweichungen von der Normalität (e.g. Box, 1953; Scheffe, 1959, 83; Levene, 1960; Glass, 1966; Leslie & Brown, 1966; Miller, 1968; Gartside, 1972; Layard, 1973; Brown & Forsythe, 1974c; Talwar & Gentle, 1977; O'Brien, 1978, 1979).

Die verschiedentlich angestellten Simulationsuntersuchungen zu den Testverfahren (vgl. Pearson, 1966; Garnes, Winkler & Probert, 1972; Hall, 1972; Layard, 1972; Brown & Forsythe, 1974c; Levy, 1975c, d, 1978a; Martin, 1976; Church & Wike, 1976; Martin & Garnes, 1977; Talwar & Gentle, 1977; O'Brien, 1978; Samiuddin, Hanif & Asad, 1978; Garnes, Keselman & Clinch, 1979a, b) haben zu recht unterschiedlichen Empfehlungen geführt, welches der Verfahren im konkreten Fall anzuwenden sei, und zwar in Abhängigkeit von der untersuchten Verteilung der Rohwerte in den Populationen. Da diese dem E üblicherweise nicht bekannt ist, ist uns eine explizite Empfehlung für einen bestimmten Varianzhomogenitätstest an dieser Stelle nicht möglich.

Es kann lediglich festgestellt werden, daß die *routinemäßige* Verwendung des in vielen Lehrbüchern der Versuchsplanung und -auswertung empfohlenen Tests nach Bartlett (1937) sicherlich *nicht* indiziert ist. Im Grunde scheint uns die Empfehlung, überhaupt auf routinemäßige Tests zur Prüfung der Varianzhomogenität zu verzichten, sofern die Stichprobenumfänge gleich groß sind, am ehesten der Mehrzahl der Fälle angemessen zu sein - zur Begründung dieser Aussage siehe die Erörterungen und Befunde von Horsnell, 1953; Box, 1953; Scheffé, 1959, 340f.; Young & Veldman, 1963; Glass, Peckham & Sanders, 1972, 242-246; Howell & Garnes, 1973; Kohr & Garnes, 1974; Havlicek & Peterson, 1974.

Für den Fall, daß die Stichprobenumfänge ungleich sind und eine Prüfung der Varianzhomogenität notwendig scheint, mag man eines der Verfahren auswählen, die in den

größeren der genannten Studien wiedergegeben und untersucht worden sind, also etwa bei Gartside (1972), Garnes, Winkler & Probert (1972), Hall (1972), Brown & Forsythe (1974c), Church & Wike (1976), Martin & Garnes (1977), O'Brien (1978). Zur Verwendung dieser Tests im Falle eines mehrfaktoriellen (varianzanalytischen) Versuchsplanes findet man einige wesentliche Hinweise etwa bei Overall & Woodward (1974), O'Brien (1978, 1979) und Garnes, Keselman & Clinch (1979b); zur Frage der Anschlußtests nach einem signifikanten generellen oder „Overall“-Test und der damit verbundenen Probleme orientiere man sich bei Levy (1975a, b), Garnes (1978a, b) sowie bei Garnes, Keselman & Clinch (1979a).

Ergeben diese Prüfungen, daß Varianzheterogenität vorliegt, ist bei gleichzeitigem Vorliegen von ungleichen Stichprobenumfängen von der Verwendung der Teststatistiken F und t abzuraten.

8.2.4.1 Zur Frage des Prüfverfahrens bei Varianzheterogenität

Das Problem, welches Verfahren im Falle ungleicher Varianzen bei der Prüfung von Hypothesen über Mittelwerte aus zwei normalverteilten Populationen angemessen ist, wurde erstmals von Behrens (1929) und wenig später von Fisher (1935) zu lösen versucht; man findet daher oft in diesem Zusammenhang die Bezeichnung „Behrens-Fisher-Problem“. Die erste den Praktiker befriedigende Lösung dieses Problems wurde von Welch (1947, 1949, 1951), Aspin (1948, 1949) sowie James (1951) vorgeschlagen; weitere Lösungsansätze findet man bei Scheffé (1970) sowie Mehta & Srinivasan (1970) zusammengestellt und diskutiert. Das Welch-James-Verfahren kann deshalb als eine „befriedigende Lösung für den Praktiker“ apostrophiert werden, weil die von Aspin (1948, 1949), Trickett & Welch (1954), Trickett, Welch & James (1956) (vgl. auch Pearson & Hartley, 1962, Tab. 11) erarbeiteten Tabellen sehr genau über die t -Verteilung approximiert werden können, deren Freiheitsgrade „adjustiert“ worden sind - vgl. hierzu insbesondere Wang (1971); auch die Teststärke dieses Verfahrens kann im Vergleich zum F -Test als durchgängig gut bezeichnet werden, wie Golhar (1972) und Levy (1978 b) sowie Hager, Lübbecke & Hübner (im Druck) gezeigt haben.

Empirische Vergleiche des von Welch, James und Aspin empfohlenen Tests mit anderen zur Lösung des Behrens-Fisher-Problems vorgeschlagenen Verfahren mittels Simulationsstudien sind u.a. von Brown & Forsythe (1974b), Kohr & Garnes (1974), Ekbohm (1976), Levy (1978c, d), Keselman, Garnes & Rogan (1979a) angestellt worden. In diesen Studien konnte sich der hier empfohlene Test durchaus bewähren.

Um so unverständlicher ist es daher, daß er nicht wesentlich stärker in Standardlehrbüchern berücksichtigt wird; die für Psychologen wohl besten Darstellungen des Testverfahrens nach Welch, James und Aspin finden sich in Li (1964, 435-438), Winer (1971, 41-44), Rasch, Enderlein & Herrendörfer (1973, 80), Clauß & Ebner (1978, 213f.) sowie Pfanzagl (1978, 216f.).

Wichtig ist die abschließende Bemerkung, daß das empfohlene Verfahren auch für die Prüfung von mehr als zwei Mittelwerten, also auch für varianzanalytische Hypothesen, Verwendung finden kann (James, 1951; Welch, 1951; Brown & Forsythe, 1974a).

Neben dem parametrischen Test nach Welch et al. kann in den Fällen, in denen Varianzheterogenität gepaart mit ungleichen Stichprobengrößen auftritt, selbstverständlich ein nicht-parametrisches Verfahren zur Anwendung kommen. Darüber hinaus ist auf diese Gruppe von Verfahren zurückzugreifen, wenn mehr als eine der Voraussetzungen verletzt ist, weil in diesen Fällen starke Abweichungen von den nominellen Fehlerwahrscheinlichkeiten resultieren. Zum Vergleich der parametrischen Teststatistiken mit ihren nicht-parametrischen Analoga orientierte man sich im einzelnen bei Boneau (1962), Neave & Granger (1968), Kemp & Conover (1973), Feir-Walsh & Toothaker (1974), Lee, Desu & Gehan (1975), Keselman, Rogan & Feir-Walsh (1977), Blair, Higgins & Smitley (1980) und bei Hager et al. (im Druck).

8.2.4.2 Zur Bedeutung von Transformationen

Zum Herstellen der Normalverteilung, zur Elimination von Nicht-Additivität (statistischen Interaktionen) und zur „Stabilisierung“ der Fehlervarianzen, d.h. Herbeiführung der Varianzhomogenität, werden nicht-lineare Transformationen der die Stichprobendaten generierenden Zufallsvariablen bereits seit geraumer Zeit vorgeschlagen, wie die Zusammenfassung von Bartlett (1947) ausweist - siehe ferner die einschlägigen Lehrbücher der Versuchsplanung und -auswertung.

Eine Systematisierung der verschiedenen möglichen Transformationen hat Lienert (1962) geleistet; weitere zusammenfassende Darstellungen finden sich bei Tukey (1949, 1957), Box & Cox (1964), Draper & Hunter (1969), Box & Tiao (1973, Kap. 10), Hoyle (1973), Schlesselman (1973), Smith (1976a) sowie Henning (1978).

Nur wenige Autoren nehmen eine kritische Haltung gegenüber Transformationen ein. Garnes & Lucas (1966) bspw. haben in Simulationsstudien gefunden, daß Transformationen zur Herstellung von normal verteilten Stichprobendaten nur selten zum gewünschten Ziel, nämlich validen Aussagen über Wahrscheinlichkeiten, führen (siehe auch Glass, Peckham & Sanders, 1972) und stellen fest: „In general, the use of a clearly interpretable scale of measurement certainly should be the dominant consideration“ (Garnes & Lucas, 1966, 326). Es ist insbesondere die Schwierigkeit der inhaltlichen Interpretation der (nicht-linear) transformierten Werte, die auch andere Autoren (e. g. Scheffé, 1959, 365-367; Digman, 1966, 475; Lindman, 1974, 35; Namboodiri, Carter & Blalock, 1975, 286) zu ähnlichen Empfehlungen gelangen lassen. Für uns ist ein anderer Grund von mindestens gleichrangiger Bedeutung, wenn wir dazu raten, auf nicht-lineare Transformationen zu verzichten: Ist die theoretische AV ein metrischer Begriff und muß für die empirische AV deshalb Intervallskalenniveau angestrebt werden, sind ausschließlich lineare Transformationen zulässig. Hat die empirische AV nur Ordinalskalenniveau, sind die hier betrachteten parametrischen Testverfahren nicht sinnvoll anwendbar, weil sich die Ergebnisse dieser Tests unter erlaubten und willkürlich möglichen monotonen Transformationen der empirischen AV ändern können (vgl. hierzu Abschnitt 2.4).

Eine ganz ähnliche Argumentation findet sich auch bei Henning (1978), der eine der wenigen Arbeiten verfaßt hat, die die Transformationen sowohl unter mathematischen als auch unter meßtheoretischen Aspekten untersuchen; seine zusammenfassende und

vertiefende Erörterung der bei Datentransformationen möglichen Probleme sei dem Interessierten empfohlen. Die Darstellung der wesentlichen Gedanken dieser Arbeit findet sich auch in Henning & Muthig (1979, 205-213).

8.2.5 Unabhängigkeit der Fehlerterme

Die statistischen Fehler oder Residuen als nicht auf die systematische Bedingungsvariation zurückführbare Komponenten von empirischen Scores können auf verschiedene Arten voneinander und von anderen Komponenten abhängig sein.

Bei den folgenden Darstellungen bezeichnen wir mit e_s und $e_{s'}$ ($s \neq s'$) zwei unterschiedliche Fehlerkomponenten und mit m und m' ($m \neq m'$) zwei verschiedene Treatmentbedingungen.

- (1) Innerhalb eines Treatments m soll die Korrelation R zwischen allen möglichen Paaren von Fehlertermen gleich Null sein: $R(e_{sm} e_{s'm}) = 0$; zur Berechnung siehe etwa Snedecor & Cochran (1972, 294f.).

Korrelierte Fehler *innerhalb* einer experimentellen Gruppe können etwa entstehen, wenn bei der Zuweisung der V_{pn} zu den Treatments wissentlich oder unwissentlich nach einer bestimmten Systematik verfahren wird, so daß sich die V_{pn} innerhalb einer Experimentalgruppe sehr ähnlich verhalten.

- (2) Zwischen je zwei experimentellen Bedingungen soll die (Inter-)Korrelation zwischen allen möglichen Paaren von Fehlern gleich Null sein:
 $R(e_{sm} . e_{s'm'}) = 0$.

In der Regel führen (unkontrollierte oder Stör-)Bedingungen, die gemeinsam mit einer bestimmten Realisation der UV auftreten, zu einer Konfundierung von Variablen, die auch die Unabhängigkeit der Fehler zwischen zwei beliebigen Treatments beeinträchtigt. Da hierbei der Einfluß der experimentell untersuchten UV nicht mehr von den Störbedingungen getrennt werden kann, liegt auch eine Verletzung der internen Validität vor.

Wird bei Vorliegen von korrelierten Fehlern der unter (1) und (2) beschriebenen Art ein F-Test durchgeführt, so sind im Falle des Vorliegens von positiven Korrelationen zu viele fälschlich signifikante Resultate zu erwarten, weil der Erwartungswert der Teststatistik F unter Gültigkeit der H_0 größer ist als $df_N / (df_N - 2)$. Im seltener vorliegenden Fall von negativen Korrelationen ist dagegen mit einer ungerechtfertigt erhöhten Anzahl von nicht-signifikanten Ergebnissen in F-Tests zu rechnen, weil der Erwartungswert von F unter H_0 kleiner als 1 werden kann. Nähere Einzelheiten hierzu finden sich u.a. in Cochran (1947), Glass & Stanley (1970), Snedecor & Cochran (1972), Keppel (1973), Lissitz & Chardos (1975) sowie Hays (1977).

Insgesamt ist festzuhalten, daß der F-Test nicht robust ist, wenn die Annahme der unabhängigen Fehlerterme nicht erfüllt ist. Daher ist es unabdingbar, die beiden unter (1) und (2) angesprochenen Abhängigkeiten zwischen Fehlertermen zu vermeiden. Dies ist zu erreichen, indem man die Vpn zufällig auf die Untersuchungsbedingungen verteilt - siehe Box (1954b), Bliss (1967, 340), Bredenkamp (1969a, 338), Öchran & Bliss (1970, 45-47). Ferner muß der E dafür Sorge tragen, daß die Werte der AV innerhalb jeder und zwischen allen experimentellen Gruppen unabhängig voneinander erhoben werden (können). Hieraus folgt nicht, daß keine Gruppenexperimente mehr durchgeführt werden sollten; die Forderung nach Unabhängigkeit bezieht sich auf die Werte der AV, die der statistischen Analyse unterzogen werden - vgl. zur weiteren Erörterung des Unterschiedes zwischen „experimentellen Einheiten“ und „Einheiten der statistischen Analyse“ etwa Glass & Stanley (1970, 501-508) sowie Abschnitt 3.3.2.

Die in diesem Abschnitt erfolgte sehr kurze Darstellung einiger der wesentlichsten möglichen Abhängigkeitsbeziehungen zwischen den verschiedenen Komponenten eines empirischen Scores, die meist nur dann erfaßbar sind, wenn man den Daten (a priori) ein entsprechendes statistisches Modell anpaßt, kann durch die Lektüre insbesondere von Hays (1977, 467, 481-483, 502, 528f., 535f., 540-543, 553, 568-574) sowie darüber hinaus etwa von Scheffé (1959, 333-339), Keppel (1973, 76, 199f., 462-467) und Bortz (1979, 344-347) wesentlich vertieft werden.

8.2.5.1 Zur Residuenanalyse; Ausreißerwerte

Bei genauerer Betrachtung wird offensichtlich, daß die Mehrheit der für die valide Anwendung der parametrischen Signifikanztests notwendigen Voraussetzungen sich auf die (theoretische) Verteilung der Fehler oder Residuen e_{im} beziehen. Diese stellen den individuellen Anteil an den Daten dar, der nicht auf systematische Bedingungseinflüsse zurückführbar ist.

Wegen der daraus sich ergebenden Bedeutung, die den Residuen im Rahmen des Konzepts der statistischen Validität zukommt, ist eine Inspektion dieser Fehler nach der Datenerhebung durchaus zu empfehlen; zur weiteren Begründung dieser Forderung vgl. etwa das Beispiel von Broekman (1973) und das Zahlenmaterial in Hampel (1980, 9). Spezielle Techniken zur Analyse der Fehlerterme sind von Anscombe & Tukey (1963), Draper & Smith (1966, Kap. 3), Wooding (1969), Behnken & Draper (1972) und Tukey (1977) vorgestellt worden. Unabhängig davon, ob man sich dieser teilweise sehr elaborierten Verfahren bedienen will, ist eine graphische Veranschaulichung der Rohwerte oder aber der Fehler als Bestandteil der erhobenen Rohwerte in der Regel unerläßlich - einen Überblick über graphische Darstellungsformen und Techniken findet man bei Wainer & Thissen (1981).

Durch Inspektion dieser Graphen etc. erhält man einen ersten Eindruck über die Verteilung der Fehler (in der Stichprobe). Insbesondere kann man hierbei sog. „Ausreißer-“ oder „Extremwerte“, die „Outliers“, entdecken, also einzelne Roh- oder Fehlerwerte, die sich durch einen ziemlichen Abstand von der Masse der Daten auszeichnen, d.h. die im Vergleich zu den übrigen Werten entweder ungewöhnlich niedrig oder hoch sind.

Für die Entstehung derartiger Outliers sind nun mindestens drei Gründe denkbar (vgl. etwa Cochran & Bliss, 1970, 47; Hampel, 1980):

- (1) Es wurden Fehler bei der Datenerhebung gemacht, etwa falsches Registrieren, Ablesen oder Übertragen von Werten.

Sofern der richtige Wert zu ermitteln ist, sollte dieser den falschen Wert ersetzen. Im anderen Fall empfiehlt es sich, den falschen Wert ersatzlos zu streichen und als „missing value“ zu behandeln - ggf. ist dann in den restlichen experimentellen Bedingungen ebenfalls je ein Wert (zufällig) zu eliminieren (siehe dazu Abschnitt 10.4.2).

- (2) Der oder die Extremwert(e) deutet (deuten) darauf hin, daß die Populationsverteilung der Fehler von einer Normalverteilung stark abweicht.
- (3) Die Fehlerwerte folgen zwar einer Normalverteilung, aber die festgestellte Ausreißerwert repräsentiert das sog. „seltene Ereignis“, von dem der E stets annimmt, daß es ihm nicht widerfährt (vgl. Fisher, 1951, 14; Neyman, 1952, 43; Haagen & Seifert, 1979, 185).

Derartige Ausreißerwerte können die numerische Größe der gebräuchlichen parametrischen Statistiken „Mittelwert M “ und „Varianz S^2 “ sehr stark beeinflussen bzw. verzerren; diese Statistiken sind nicht „resistent“ (oft auch: „robust“) gegenüber Ausreißern - vgl. zu Beispielen und weiteren Einzelheiten etwa Mosteller & Tukey (1977, 203-212).

Die Tatsache der unzulänglichen Resistenz der „klassischen“ parametrischen Statistiken gegenüber Ausreißern hat zu unterschiedlichen Empfehlungen geführt, wie man sich angesichts von Extremwerten verhalten sollte.

- (1) Man verzichtet auf die parametrischen Statistiken M , S^2 und r und ersetzt sie durch resistente analoge Kenngrößen wie etwa den Median M_d und Rangkorrelationskoeffizienten oder aber durch spezielle, besonders in jüngster Zeit vorgeschlagene Statistiken, über die im einzelnen etwa Andrews et al. (1972), Wainer (1976), Wainer & Thissen (1976), Tukey (1977) sowie Mosteller & Tukey (1977) informieren. über diese analogen Statistiken können häufig die üblichen Signifikanztests, teilweise leicht modifiziert, durchgeführt werden (vgl. u.a. Schrader & McKean, 1977).
- (2) Man verwendet zwar die üblichen Statistiken, modifiziert oder eliminiert jedoch die Ausreißerwerte.

Mittels spezieller Testverfahren kann bspw. beurteilt werden, ob die Ausreißerwerte einer Normalverteilung entstammen und daher nur ein zufällig extremes Resultat darstellen - eine zusammenfassende Darstellung dieser Verfahren findet sich bei Barnett & Lewis (1978); neuere Tests wurden von Tiku (1975) und Hawkins (1979) vorgestellt. Die Tests entdecken erst vergleichsweise extreme Ausreißerwerte.

Bei Vorliegen von zufälligen, einer Normalverteilung entstammenden extremen Residuen werden unterschiedliche Strategien empfohlen, die darauf abzielen, den Einfluß dieser Werte auf die Größe der zu berechnenden Statistiken zu minimieren.

Die einfachste Lösung stellt das sog. „Stutzen“ oder „Trimming“ dar. Hierbei werden die extremen Werte bei der Analyse unberücksichtigt gelassen. Eine andere Lösung, die sog. „Winsorization“, ersetzt die Outliers nach einem bestimmten Modus durch weniger extreme Daten. Einzelheiten zu beiden Ansätzen sind den Publikationen von Tukey & McLaughlin (1963), Dixon & Tukey (1968), Andrews et al. (1972), Wainer (1976), Mosteller & Tukey (1977) und Hampel (1980) zu entnehmen. All die vorgeschlagenen Maßnahmen, deren grundsätzliche Rechtfertigung und Validität durch eine großangelegte Computer-Studie belegt wird (Andrews et al., 1972; ergänzend Wegman & Carroll, 1977), entheben den E jedoch nicht der Notwendigkeit, das Zustandekommen der Outliers besonders zu diskutieren: „Sometimes the outlier is providing information which other data points cannot due to the fact that it arises from a unusual combination of circumstances which may be of vital interest and requires further investigation rather than rejection.“ (Draper & Smith, 1966, 95)

Weitere Informationen zum Komplex der Residuenanalyse und der Resistenz von statistischen Kennwerten gegenüber Extremwerten können den zusammenfassenden Arbeiten etwa von Huber (1972), Hogg (1974, 1977, 1979), Bickel (1976), Wainer (1976) sowie Hampel (1980) entnommen werden. über eine spezielle Technik im Zusammenhang mit Varianzen, das „Jackknifing“, informieren zudem u.a. die Arbeiten von Miller (1968, 1974), Mosteller & Tukey (1968, 1977) sowie Gray & Schucany (1972).²⁷⁾

U. E. sollte eine Sonderbehandlung von Extremwerten ausschließlich dann in Erwägung gezogen werden, wenn dadurch die Validität des Experiments zur Prüfung der interessierenden WH erhöht wird. Dies ist beispielsweise dann der Fall, wenn die Extremwerte auf fehlerhafte Datenerhebung beruhen - nach Hampel (1980, 9) eine der häufigsten Ursachen für Ausreißer - oder wenn vermutet werden muß, daß sie von Vpn stammen, die die Instruktionen nicht angemessen befolgt haben.

²⁷⁾ „Jackknifing“ wird üblicherweise auf Varianzen angewendet, ist jedoch nicht auf diese Anwendung beschränkt (siehe Gray & Schucany, 1972).

8.2.6 Problem der Zufallsstichproben

Eine der Voraussetzungen für parametrische Testverfahren besteht darin, daß die zu untersuchenden Daten eine (oder mehrere) Zufallsstichprobe(n) aus einer (oder mehreren) (in der Regel normalverteilten) Populationen darstellen. Nur dann lassen sich die Wahrscheinlichkeiten für bestimmte Realisationen von Statistiken angeben. Diese Angaben wiederum sind notwendig, um überhaupt statistische Tests im oben beschriebenen Sinne (vgl. Teil 7) durchführen zu können, bei denen die Wahrscheinlichkeiten für Fehlentscheidungen - und sei es auch nur innerhalb gewisser Grenzen - ange- und kontrollierbar sind.

Mit dem Begriff „Zufallsstichprobe“ wird eine bestimmte mathematische Modellvorstellung bezeichnet (s. Fisz, 1970, 394; Menges, 1972), die angemessen ist, wenn aus einer endlichen Grundgesamtheit eine Teilmenge so ausgewählt wird, daß jedes Element der Grundgesamtheit mit der gleichen Wahrscheinlichkeit „gezogen“ wird. In der Psychologie werden Untersuchungen aber in aller Regel mit den experimentellen Einheiten durchgeführt, die gerade „verfügbar“ sind, die in diesem Sinne also keine Zufallsstichprobe aus einer bestimmten existierenden Population sind. Man kann parametrische Tests aber auch unter diesen Umständen anwenden:

Viele parametrische Tests können als rechnerische Vereinfachungen entsprechender Randomisierungstests interpretiert werden (s. Abschn. 7.5.2). Bei Randomisierungstests wird die Verteilung der Teststatistik unter Gültigkeit der Nullhypothese aber abgeleitet, ohne daß Annahmen über Eigenschaften von Populationen gemacht werden. Deshalb können diese Tests unabhängig davon durchgeführt werden, aus welchen Populationen die beobachteten Daten stammen. Es ist dann aber auch möglich, jeden Datensatz als Zufallsstichprobe aus einer *hypothetischen* Population anzusehen (Alf & Abrahams, 1973), deren Verteilungsfunktion genau jene Eigenschaften hat, bei denen die Wahrscheinlichkeit maximiert wird, daß die empirischen *Daten* eine Zufallsstichprobe aus dieser Population sind. Die von Randomisierungstests geprüften Nullhypothesen sind Aussagen über diese hypothetischen Populationen, z.B. über die Gleichheit der Verteilungsfunktionen der Populationen, die den verschiedenen Gruppen von Daten (also etwa den experimentellen Bedingungen) entsprechen, oder über die Unabhängigkeit von Variablen in diesen Populationen (siehe Bradley, 1968, 68-73).

Insgesamt ist es zur Durchführung und Interpretation parametrischer Testverfahren also nicht notwendig, daß die beobachteten *experimentellen Einheiten* eine Zufallsstichprobe aus einer real existierenden oder inhaltlich definierbaren Population darstellen. Diese Frage der statistischen Validität ist zu unterscheiden von der Frage, an welchen experimentellen Einheiten eine Kausalhypothese

se zu prüfen ist, damit die entsprechende Untersuchung eine möglichst hohe Populationsvalidität hat. Nach den Ergebnissen des Abschnitts 4.1 sind aber auch dazu keine Zufallsstichproben notwendig.

8.3 Kumulierung der Wahrscheinlichkeiten für Fehler erster und zweiter Art

Führen wir *einen* Signifikanztest durch, ist die Wahrscheinlichkeit für eine fälschliche Ablehnung von H_0 höchstens gleich α . Führen wir *zwei* unabhängige Signifikanztests durch, und ist in beiden Fällen H_0 richtig, so ist die Wahrscheinlichkeit, daß in beiden Fällen auch richtigerweise ein für H_0 sprechendes Stichprobenergebnis eintritt, gleich $(1 - \alpha)^2$.

Die Komplementärwahrscheinlichkeit, daß einer oder beide Tests fälschlicherweise für H_1 sprechen, ist damit gleich $1 - (1 - \alpha)^2$. Allgemein berechnet sich daher die Wahrscheinlichkeit, bei T unabhängigen Signifikanztests, bei denen jeweils die Nullhypothese zutrifft, *mindestens* ein für H_1 sprechendes Stichprobenergebnis zu erhalten, zu:

$$(8.4) \quad P_\alpha = 1 - (1 - \alpha)^T$$

(Ryan, 1959, 1962; Hays, 1977, 611). Bei ungleichen α -Niveaus für die einzelnen Tests ergibt sich die Gesamtwahrscheinlichkeit entsprechend aus

$$(8.5) \quad P_\alpha = 1 - (1 - \alpha_1) \cdot (1 - \alpha_2) \cdot \dots \cdot (1 - \alpha_T) = 1 - \prod_{t=1}^T (1 - \alpha_t)$$

(Miller, 1966, 8).

Sind die statistischen Tests nicht unabhängig, stellt P_α zugleich eine gute Annäherung und eine *obere Schranke* für die gesuchte Wahrscheinlichkeit dar (Kimball, 1951; Miller, 1966, 101-102, 1977; Petrinovich & Hardyck, 1969).

Nach (8.4) steigt bei Durchführung mehrerer Signifikanztests die Wahrscheinlichkeit von mindestens einem zufällig signifikanten Ergebnis schnell an, bei einem $\alpha = 0,05$ und den 7 Tests einer dreifachen Varianzanalyse kann beispielsweise P_α schon 0,30 betragen, bei den 15 möglichen Tests einer vierfachen Varianzanalyse schon 0,54 (vgl. die Tabellen bei Jacobs, 1976). Deshalb sollten komplexe Versuchspläne, die mehrere Signifikanztests erforderlich machen, nur verwendet werden, wenn es zur Prüfung der betrachteten wissenschaftlichen Hypothese unumgänglich ist. Insbesondere sollten in Varianz- und Regressionsanalysen nur solche Faktoren berücksichtigt und nur solche (Haupt- und Interaktions-)Effekte geprüft werden, die tatsächlich von theore-

tischer oder praktischer Wichtigkeit sind (vgl. Rule, 1976). Hat man dennoch mehrere Signifikanztests durchzuführen, kann man den Kumulationseffekt dadurch ausgleichen, daß man bei T Tests die Irrtumswahrscheinlichkeit α jedes Einzeltests so wählt, daß P_α höchstens gleich einem festzusetzenden Grenzwert α_G ist. Mit $P_\alpha = \alpha_G$ ergibt sich aus (8.4) nämlich:

$$(8.6) \quad \alpha = 1 - \sqrt[T]{1 - \alpha_G}$$

(vgl. Garnes, 1971b). Der Wert α_0 gibt dabei die festgesetzte, auf das gesamte Experiment bezogene Wahrscheinlichkeit für mindestens einen Fehler 1. Art an. Der Einfachheit halber kann man auch die aus der *Bonferroni-Ungleichung* abgeleitete Regel

$$(8.7) \quad \alpha = \alpha_G / T$$

(Dunn, 1961; Dunnett, 1970; Garnes, 1971b) verwenden, die nur zu ganz geringfügig niedrigeren Werten führt (vgl. Lunney, 1969).²⁸⁾

Ein Nachteil dieser Vorgehensweisen ist darin zu sehen, daß durch das Herabsetzen der Fehlerwahrscheinlichkeit α für die einzelnen Tests bei gleichbleibender Zahl der Beobachtungen die Teststärke $1 - \beta$ der Tests sinkt, bzw. daß mehr Beobachtungen notwendig sind, um eine gleichbleibende Power zu erzielen. Diese verminderte Effizienz stellt auch den entscheidenden Nachteil der Verfahren für multiple Mittelwertsvergleiche dar, die von vornherein von einem adjustierten Signifikanzniveau ausgehen, wie es etwa bei den Tests nach Tukey und Scheffé der Fall ist (vgl. Abschnitt 8.3.1).

Die dargestellten traditionellen Überlegungen zur Kumulierung von Fehlern 1. Art sind noch ergänzungsbedürftig, wenn wir wieder stärker die strenge Prüfung wissenschaftlicher Hypothesen als übergeordnetes Ziel aller unserer (experimentellen) Bemühungen berücksichtigen. Werden bei der empirischen Überprüfung einer Hypothese oder Theorie nicht nur eine, sondern mehrere statistische Hypothesen aus ihr abgeleitet, handelt es sich meist um Alternativhypothesen $H_{1(1)}, H_{1(2)}, \dots, H_{1(T)}$ (vgl. die Punkte (6), (7) und (8) in Abschnitt 8.1.1). Wird die wissenschaftliche Hypothese nur dann als bewährt betrachtet, wenn alle implizierten Alternativhypothesen angenommen werden können, ist die Wahrscheinlichkeit, daß man diese Entscheidung fälschlicherweise trifft, stets höchstens gleich α . Deshalb muß ein konventionell beispielsweise auf 0,05 festgesetzter Wert für α nicht adjustiert werden (Westermann, im Druck, b; Hager & Westermann, im Druck, b).

²⁸⁾ Zur Wahl von α , wenn die Wahrscheinlichkeit für mindestens r ($r = 2, 3, \dots$) Fehler erster Art einen festen Wert α_G nicht überschreiten soll, siehe Hsu (1978), Hurlburt & Spiegel (1976) und die Tabellen bei Feild & Armenakis (1974).

Auf der anderen Seite könnte man sich für eine Falsifikation der wissenschaftlichen Hypothese entscheiden, wenn mindestens eine der implizierten statistischen Hypothesen nicht zutrifft. Durch Übertragung der Ergebnisse aus den Überlegungen zur Kumulierung der Fehler erster Art wissen wir, daß bei Gültigkeit der statistischen Alternativhypothesen $H_{1(1)}, H_{1(2)}, \dots, H_{1(T)}$ die Wahrscheinlichkeit für mindestens ein (fälschlich) für eine Null-Hypothese $H_{0(t)}$ sprechendes empirisches Ergebnis bis auf

$$(8.8) \quad P_\beta = 1 - \prod_{t=1}^T (1 - \beta_t),$$

ansteigen kann.

Impliziert eine wissenschaftliche Hypothese mehrere statistische Alternativhypothesen, muß man daher zur Vermeidung ungerechtfertigter Falsifikationen die β -Wahrscheinlichkeit für jeden einzelnen Test möglichst klein wählen (Westermann, im Druck, a). Idealerweise sollte sie analog zu (8.7) durch

$$(8.9) \quad \beta = \beta_G/T$$

bestimmt werden.

Beispiel: Soll bei Durchführung von 3 Signifikanztests die Wahrscheinlichkeit von mindestens einer fälschlichen Annahme der Nullhypothese höchstens gleich 0,20 sein, muß die Zahl N der Beobachtungen so gewählt werden, daß jeder Signifikanztest eine Teststärke von $1 - \beta = 1 - \beta_G/T = 1 - 0,20/3 = 0,93$ aufweist.

Diese Überlegungen hinsichtlich einer Herabsetzung von α und β zum Ausgleich von Kumulierungseffekten können auf alle Entscheidungen zwischen zwei Alternativen A und nicht-A verallgemeinert werden, die aufgrund der Ergebnisse mehrerer Signifikanztest getroffen werden: Wird eine Entscheidung (z.B. für A) getroffen, wenn alle statistischen Hypothesen einer bestimmten Menge angenommen werden können, muß die Wahrscheinlichkeit für die fälschliche Annahme einer dieser Hypothesen nicht adjustiert werden; fällt die Entscheidung (für A) aber schon, wenn (*mindestens*) eine der statistischen Hypothesen angenommen wird, sollte die Wahrscheinlichkeit für eine Fehlentscheidung auf die Gesamtmenge der statistischen Tests bezogen werden, für den einzelnen Test also herabgesetzt werden (vgl. Rule, 1976).

Diskutiert wird häufig die Frage, wie die „Familie“ von Tests zu spezifizieren ist, für die die Irrtumswahrscheinlichkeit insgesamt festgelegt wird (Miller, 1966, 31-35; Ryan, 1962; Kirk, 1968, 77f.; Dunnett, 1970, 100). Uns erscheint es ungerechtfertigt, generell alle Signifikanztests eines Experiments zusammenzufassen („experimentwise error rate“), da es z.B. recht willkürlich sein kann, ob eine Hypothese nun allein oder (in einem multifaktoriellen

Design) zusammen mit anderen überprüft wird. Die obigen Erörterungen bieten u.E. bessere Leitlinien: Für mehrere Signifikanztests ist genau dann insgesamt eine maximale Fehlerwahrscheinlichkeit festzulegen, wenn mit der Annahme einer einzigen der statistischen Hypothesen die Voraussetzung für eine bestimmte Entscheidung (z.B. über die Falsifikation einer wissenschaftlichen Hypothese oder die Einleitung einer praktischen Maßnahme) erfüllt ist.

Im folgenden wollen wir die Behandlung des Problems der kumulierten Fehlerwahrscheinlichkeiten für zwei besonders wichtige Hypothesenarten noch konkreter betrachten.

8.3.1 Multiple Mittelwertsvergleiche

Von multiplen Mittelwertsvergleichen wollen wir sprechen, wenn zugleich mehrere statistische Hypothesen über eine Menge von Mittelwerten geprüft werden. Dies kann in folgenden Situationen der Fall sein:

- (1) Die wissenschaftliche Hypothese impliziert die Null- oder Alternativhypothese einer Varianzanalyse (siehe Punkt (6) in Abschnitt 8.1.1). Die Annahme der statistischen H_1 ist dann gleichbedeutend mit der Entscheidung, daß mindestens zwei der betrachteten Mittelwerte $\mu_1, \dots, \mu_K$ ungleich sind. In diesem Fall kann es von Interesse sein zu prüfen, welche der Mittelwerte dies sind und/oder ob ganz bestimmte Mittelwerte ungleich sind.
- (2) Die psychologische Hypothese impliziert verschiedene statistische Hypothesen über eine Menge von Mittelwerten (vgl. Punkt (8) im Abschnitt 8.1.1), also z.B. die folgenden vier Null- bzw. Alternativhypothesen:

$$H_{0(1)}: \mu_1 = \mu_2$$

$$H_{0(2)}: \mu_3 = \mu_4$$

$$H_{0(3)}: \mu_4 = \mu_5$$

$$H_{1(4)}: (\mu_1, \mu_2) > (\mu_3, \mu_4, \mu_5)$$

Dabei bezeichnet z.B. (μ_3, μ_4, μ_5) den Mittelwert $\mu_{3,4,5}$ der aus den Populationen 3, 4 und 5 gebildeten Gesamtpopulation, der sich zu $(1/3)(\mu_3 + \mu_4 + \mu_5)$ ergibt. Geprüft werden derartige Hypothesen über entsprechend konstruierte Linearkombinationen von Stichprobenmittelwerten („Methode der (orthogonalen) Kontraste“; siehe dazu u.a. Winer, 1971, 170-177; Myers, 1972, 352-362; Hays, 1977, 581-605).

Jeder Mittelwertshypothese entspricht eine Hypothese über einen Kontrast, d.h. eine Linearkombination von Mittelwerten, die unter den üblichen Annahmen des ALM (vgl. Abschnitt 8.2.1) über t-verteilte Teststatistiken geprüft

werden kann.²⁹⁾ Nach dem zu Beginn dieses Abschnitts Gesagten muß je nach Fragestellung α und/oder β für die Einzeltests so gewählt (adjustiert) werden, daß die Entscheidung für oder gegen die wissenschaftliche Hypothese nicht allein aufgrund der erhöhten Wahrscheinlichkeit für mindestens einen Fehler 1. oder 2. Art getroffen wird. Wir wollen dann von adjustierten multiplen t-Tests sprechen.

Die Bestimmung kritischer t-Werte ist auch für „ungewöhnliche“ Signifikanzniveaus α nicht schwierig. Sie erfolgt über die ausführlichen Tabellen der t-Verteilung bei Pearson & Hartley (1962, 132-134) oder über die speziell für diesen Zweck eingerichteten Tabellen von Dunn (1961; auch in Kirk, 1968, 551; und Marascuilo & McSweeney, 1977, 483) oder von Dayton & Schafer (1973).

Auch wenn aus der wissenschaftlichen Hypothese eine Menge von T Mittelwerts-hypothesen folgt, kann es sinnvoll sein, vor Durchführung der T Mittelwertsvergleiche als globalen oder „Overall-Test“ eine Varianzanalyse oder eine ihrer nicht-parametrischen Entsprechungen über alle Mittelwerte durchzuführen (vgl. Fisher, 1951, 196-204; Miller, 1966, 90; Dunnett, 1970; Carmer & Swanson, 1971, 1973; Cohen & Cohen, 1975, 162-165; Gaito, 1978; Keselman, Garnes & Rogan, 1979a, 1980; Swaminathan & deFriesse, 1979; Ryan, 1980). Dabei muß die Zahl der Beobachtungen so gewählt werden, daß α und β auf kleine Werte fixiert werden können. Wird dann der Test der Null-Hypothese, daß alle Mittelwerte gleich sind, nicht signifikant, können auch die Null-Hypothesen aller interessierenden Mittelwertskontraste angenommen werden.

Impliziert also die wissenschaftliche Hypothese die Gültigkeit der Alternativ-hypothesen dieser Kontraste, kann sie damit bereits aufgrund des globalen Tests falsifiziert werden (vgl. im einzelnen Teil 11), ohne daß es zur Kumulierung von Fehlerwahrscheinlichkeiten kommen kann. Wird dagegen die H_1 des Overall-Tests angenommen, kann mit der relativ klein angesetzten Irrtumswahrscheinlichkeit α geschlossen werden, daß *mindestens* zwei Mittelwerte unterschiedlich sind. Ob dies allerdings auch die von der wissenschaftlichen Hypothese implizierten sind, muß anschließend über die multiplen Vergleiche geprüft werden. Analoges gilt, wenn aus der psychologischen Hypothese die Null-Hypothese einer Menge von Mittelwertsvergleichen folgt.

²⁹⁾ Zur nicht-parametrischen Prüfung von Kontrasten siehe vor allem Marascuilo & McSweeney (1967, 1977, S. 306-312, 362-371); ferner Keselman & Rogan (1977), Wike & Church (1977, 1978), Silverstein (1978), Church & Wike (1979, 1980), Levy (1979a, b) sowie Shuster & Boyert (1979).

Sind aus der wissenschaftlichen Hypothese nicht Hypothesen über Mittelwerte, sondern über andere Parameter abgeleitet worden (etwa über Varianzen, Korrelationen oder Porportionen), dann lassen sich diese ebenfalls über die hier beschriebene Technik der multiplen Vergleiche prüfen (siehe Garnes, 1978a, b); das gleiche gilt für Hypothesen über Trends (siehe z.B. Winer, 1971, 177-185; Hays, 1977, 687-694).

Eine Alternative zu den adjustierten multiplen t-Tests stellen multiple Vergleichstechniken dar, bei denen die kritischen Werte für die Einzeltests von vornherein so gewählt werden, daß die Wahrscheinlichkeit für mindestens ein fälschlich signifikantes Ergebnis in einer bestimmten „Familie von Tests“ einen festzulegenden Wert α_G nicht überschreitet. Die bekanntesten dieser Techniken stammen von Scheffé (1959; vgl. Gabriel, 1964, 1969, 1978; Boik, 1979) und Tukey (1953; zit. n. Ryan, 1959; vgl. Gabriel, 1969; Williams, 1974, Keselman & Rogan, 1977, 1978; Keselman, Garnes & Rogan, 1979b). Beim Scheffé-Test wird der Rejektionsbereich so bestimmt, daß bei Testung aller möglichen Kontraste die Wahrscheinlichkeit mindestens eines Fehlers 1. Art höchstens gleich α_G ist, bei den Tukey-Tests bezieht sich dieser Wert auf alle möglichen Kombinationen je zweier Mittelwerte.

Wir können auf diese und die anderen vorliegenden multiplen Vergleichsmethoden hier nicht näher eingehen. Leicht verständliche Darstellungen der gebräuchlichsten Verfahren geben Kirk (1968, S. 65-98), Dunnett (1970), Keppel (1973, Kap. 8) und Diehl (1979, Kap. 3). Des weiteren verweisen wir auf die Übersichten von Miller (1966, 1977), Garnes (1971 b, 1978a, b), O'Neill & Wetherill (1971), Lippman & Taylor (1972), Thomas (1973), Hopkins & Anderson (1973) sowie von Spjøtvoll (1974); zur Anwendung der Verfahren bei ungleichen Stichprobengrößen siehe u.a. Garnes (1971a), Smith (1971), Shaffer (1974a), Keselman, Toothaker & Shooter (1975), Keselman, Murray & Rogan (1976), Hochberg (1976) und Tamhane (1977). Vergleichende Untersuchungen zu einzelnen Verfahren auch hinsichtlich ihrer Robustheit gegenüber Verletzungen der Annahmen findet man etwa bei Ury (1971), Ury & Wiggins (1971, 1975), Ramseyer & Tcheng (1973), Keselman & Toothaker (1974), Howell & Garnes (1974), Garnes & Howell (1976), Keselman (1976), Kohr & Garnes (1977), Rogan, Keselman & Breen (1977), Keselman & Rogan (1977, 1978), Tamhane (1979); zur Teststärke dieser Verfahren siehe ferner Harter (1957), Steffens (1970), David, Lachenbruch & Brandis (1972), Einot & Gabriel (1975), Ramsey (1978).

Als Anhaltspunkte für die Wahl eines angemessenen Analyseverfahrens seien die folgenden Aspekte hervorgehoben:

- (1) Multiple t-Tests und der Scheffé-Test können zur Prüfung jedes *beliebigen* Mittelwertskontrasts herangezogen werden. Dabei sind multiple t-Tests effizienter, sofern die Zahl der zu prüfenden Vergleiche nicht wesentlich höher ist als die Zahl der Mittelwerte (vgl. u.a. Miller, 1966, 69; Garnes, 1978a).
- (2) Die Verfahren von Tukey sind für *paarweise* Mittelwertsvergleiche effizienter als multiple t-Tests, es sei denn, es werden nur vglw. wenige der insgesamt möglichen Vergleiche durchgeführt (Dunn, 1961; Miller, 1966, 69; Garnes, 1978a).
- (3) Multiple t-Tests ermöglichen als einziges der genannten Verfahren stets auch die (teststärkere) Prüfung gerichteter Hypothesen.
- (4) Multiple t-Tests sind einfacher durchzuführen als die anderen Verfahren; ihre Teststärke ist problemlos bestimmbar; sie sind robust gegenüber Ver-

letzungen der zu ihrer validen Anwendung notwendigen Voraussetzungen, und daneben liegen vergleichbar einfache alternative Verfahren (siehe Kohr & Garnes, 1977; Garnes, 1978b) sowie relativ effiziente nicht-parametrische Analoga vor, deren Einsatz etwa bei mangelndem Skalenniveau indiziert ist.

Besonders in Anbetracht dessen, daß in hypothesen- oder theoriengeleiteten Untersuchungen nur relativ wenige Vergleiche von Interesse sein dürften (relativ zur Zahl möglicher (paarweiser) Kontraste), sollten in der Regel multiple t-Tests mit adjustierten α - und β -Wahrscheinlichkeiten durchgeführt werden. Dies gilt unabhängig davon, ob die Mittelwertsvergleiche vor Datenerhebung geplant sind (s.o. Fall 2) oder nicht (Fall 1) und auch unabhängig davon, ob die Vergleiche statistisch unabhängig (orthogonal) sind oder nicht (siehe dazu Ryan, 1959; Davis, 1969; Ury & Wiggins, 1971, 1974; Myers, 1972, 362; Rodger, 1973). Erforderlich ist lediglich, daß vor der Durchführung der Mittelwertsvergleiche deren Anzahl festgelegt wird, um α und β für jeden Einzeltest entsprechend adjustieren zu können.

8.3.2 *Monotone Trendhypothesen*

Wir betrachten nun den in der Psychologie sehr häufigen Fall, daß sich aus der wissenschaftlichen Hypothese eine statistische Hypothese über die Ordnung von Mittelwerten ableiten läßt (vgl. Abschn. 8.1.1): $H_1: \mu_1 < \mu_2 < \dots < \mu_K$. Diese Hypothese kann falsifiziert werden, wenn die Nullhypothese einer einfachen Varianzanalyse angenommen wird. Die Varianzanalyse kann also als Overall-Test zur Prüfung der hier betrachteten Hypothese benutzt werden. Wird die Alternativhypothese der Varianzanalyse angenommen, müssen zur Prüfung von H_1 aber noch $K-1$ paarweise Mittelwertsvergleiche mit den Alternativhypothesen $H_{1(1)}: \mu_1 < \mu_2$ bis $H_{1(K-1)}: \mu_{K-1} < \mu_K$ durchgeführt werden.

Als Entscheidungsregel kann man vereinbaren, daß die Gesamt-Alternativhypothese H_1 angenommen werden soll, wenn alle Einzel-Alternativhypothesen $H_{1(k)}$ (mit $k = 1, \dots, K-1$) angenommen worden sind, wobei selbstverständlich auch abgeschwächte Kriterien denkbar sind. Die Wahrscheinlichkeit α für die einzelnen Tests kann dann durchaus auch größer als $\alpha = 0,05$ angesetzt werden. Sollen bspw. fünf Paar-Vergleiche durchgeführt werden, könnte man α auf 0,20 festlegen, denn dann ist die Wahrscheinlichkeit einer Annahme der Gesamt- H_1 , obwohl alle Einzel- $H_{1(k)}$ falsch sind, mit $0,20^5 = 0,0003$ extrem gering. Aber auch für den Fall, daß nur eine der $H_{1(k)}$ falsch sein sollte, scheint eine Irrtumswahrscheinlichkeit von 0,20 für diese der wissenschaftlichen Hypothese doch noch recht gut entsprechenden Situation tolerabel zu sein.

Dagegen sollte die Wahrscheinlichkeit eines Fehlers 2. Art für jeden Einzeltest sehr klein sein, da es sonst leicht zu mindestens einer fälschlichen Annahme

einer $H_{0(k)}$ und damit zu einer fälschlichen Ablehnung der Gesamt- H , kommt. Daraus folgt die Notwendigkeit, hier möglichst effiziente statistische Tests zu verwenden. Nach dem im Abschnitt 8.3.1 Gesagten sind zur Prüfung monotoner Trendhypothesen also multiple t-Tests (oder entsprechende nicht-parametrische Verfahren) geeigneter als spezielle multiple Vergleichstechniken wie die Methoden von Scheffé und Tukey.

Als Overall-Test können anstelle der einfachen Varianzanalyse Verfahren angewendet werden, die von einer gerichteten Alternativhypothese ausgehen, d.h. nur dann zu signifikanten Resultaten führen, wenn aufgrund des Stichprobenergebnisses mindestens ein Mittelwertsunterschied in der spezifizierten Richtung besteht.

Ein entsprechendes parametrisches Verfahren stammt von Bartholomew (1959a, b, 1961a, b; s.a. Barlow et al., 1972); nicht-parametrische Verfahren sind für unabhängige Stichproben der Si-Test von Jonckheere (1954a, b; vgl. zur Darstellung und Beurteilung etwa Pm-i, 1965; May & Konkin, 1970; Odeh, 1972; Lienert, 1973, 279f.; Nelson & Toothaker, 1975 und Berenson, 1978) und für Designs mit einer Block- oder Kontrollvariablen der L-Test von Page (1963; vgl. zur Darstellung etwa Lienert, 1973, 357-360).

Diese Verfahren sind bei der hier betrachteten Fragestellung in der Regel effizienter als der varianzanalytische F-Test bzw. die entsprechenden Rangvarianzanalysen (Bartholomew, 1961a, b; Boersma, deJonge & Steilwagen, 1964; Puri, 1965; Garnes, 1966; Nelson & Toothaker, 1975). Da beim Overall-Test β zur Vermeidung ungerechtfertigter Falsifikationen sehr klein gewählt werden muß, ist die Verwendung dieser Tests in der hier betrachteten Situation von Vorteil.

Eine aus der psychologischen Hypothese abgeleitete statistische Hypothese über die Rangordnung von Mittelwerten kann selbstverständlich auch durch die ausschließliche Anwendung eines dieser Trendtests oder aber auch mit einem (Rang-)Korrelationstest geprüft werden. Da aber der Test jedes einzelnen Unterschiedes zwischen „benachbarten“ Mittelwerten zu einer insgesamt strengeren Prüfung der W führt, ist diese verfahrensweise i.a. vorzuziehen (vgl. Hager & Westermann, im Druck, a).

8.4 Mangelnde Präzision

Beim Signifikanztest erfolgt die Entscheidung über die Gültigkeit statistischer Hypothesen, indem der empirische Wert der Zufallsvariablen „Teststatistik“ in Beziehung zur Prüfverteilung gesetzt wird, d.h. zur Verteilung dieser Zufallsvariablen bei Gültigkeit der Nullhypothese (vgl. Abschn. 7.2). Die Wurzel aus der Varianz dieser Verteilung wird „Standardfehler“ der Teststatistik genannt (s. insbesondere Guilford & Fruchter, 1973). Wir können jetzt allgemein definieren: *Die Präzision der Prüfung einer statistischen Hypothese (oder einfacher: Die Präzision eines Experiments) ist um so höher, je kleiner der Standardfehler der entsprechenden Teststatistik ist.*

Bei festen Größen für die Wahrscheinlichkeit eines Fehlers 1. Art und für den in absoluten Einheiten ausgedrückten experimentellen Effekt (z.B. $\alpha = 0,05$, $EE = \mu_1 - \mu_2 = 3$) besteht folgende Beziehung: *Je geringer der Standardfehler der Teststatistik ist, desto geringer ist die Wahrscheinlichkeit β für einen Fehler 2. Art.* Deshalb ist (unter sonst gleichen Bedingungen) die statistische Validität eines Experimentes um so höher, je größer seine Präzision ist.

Welche Faktoren bestimmen die Größe des Standardfehlers von Teststatistiken und damit die Präzision von Experimenten?

- (1) Der Standardfehler einer Teststatistik ist (unter sonst gleichen Bedingungen) stets um so geringer, je größer die Zahl N der erhobenen Werte der abhängigen Variablen ist. (Zu den Folgerungen aus dieser Beziehung im Hinblick auf die Wahl der Stichprobengröße siehe Abschnitt 10.1.)
- (2) Bei parametrischen Testverfahren ist der Standardfehler einer Teststatistik (wiederum unter sonst gleichen Bedingungen) um so geringer, je kleiner die Varianz σ_e^2 der sog. „unsystematischen Fehler“ ist, d.h. der Abweichungen der einzelnen Werte auf der abhängigen Variablen (in der Population) von ihrem Mittelwert (vgl. z.B. Lindquist, 1953, S. 2-5; Bredenkamp, 1969a, S. 339 und Abschn. 8.2.3). Deshalb wird die Präzision eines Experiments erhöht, wenn die Varianz σ_e^2 der abhängigen Variablen in den Behandlungspopulationen verringert wird. Im folgenden werden wir eine Reihe von Techniken besprechen, die auf diese Weise zur Präzisionserhöhung und damit auch zu einer Vergrößerung der statistischen Validität beitragen können.

8.4.1 Parallelisierung als Kontrolltechnik (StatV)

Besteht die Vermutung, daß die Varianz der AV durch eine ganz bestimmte individuelle Variable C beträchtlich erhöht wird und lassen sich Informationen über die Ausprägungen aller Probanden auf dieser Variablen beschaffen, so kann die Präzision des Experiments erhöht werden, indem diese Variable C als Kontrollfaktor ins Design eingeführt wird. Man spricht dann von einer „Parallelisierung“ (matching) der Probanden hinsichtlich dieser Variablen. Die Vorgehensweise sei am Beispiel eines Designs mit Ausprägungen einer UV erläutert.

Soll etwa hinsichtlich der Variablen „Intelligenzquotient“ parallelisiert werden, sind zunächst alle Probanden gemäß ihres IQ in eine Rangordnung zu bringen. Dann werden jeweils K Probanden der Reihe nach zu einem sog. „Block“ zusammengefaßt; die Zahl der Probanden muß also ein ganzzahliges Vielfaches von K sein. Innerhalb der $Q=N/K$ Blöcke wird jeder experimentellen Bedingung dann genau ein Proband zufällig zugeordnet; daher rührt die Bezeichnung „Plan der Zufallsblöcke“ für diesen Versuchsplan 10.

(Vpl. 10)

		R-UV B				
		B ₁	...	B _m	...	B _K
(Kontroll- (Block-) variable) B-UV C	C ₁					
	⋮					
	C _r			Y _{rm}		
	⋮					
	C _Q					

Zur Auswertung derartiger Pläne orientiere man sich in der in den Abschnitten 7.5.3.2, 7.5.3.3 und 8.4.6 angegebenen Literatur.

Eine Parallelisierung führt allerdings nur dann zu einer Erhöhung der Präzision des Experiments (d.h. zu einer Verringerung der Varianz der Prüfverteilung), wenn die Kontrollvariable eine zumindest mäßige statistische Assoziation mit der abhängigen Variablen aufweist - siehe dazu etwa Feldt (1958) sowie zur eingehenderen Begründung z.B. Myers (1972, 153-156).

Empfehlenswert ist es, mehrere benachbarte Blöcke zusammenzufassen, so daß in jeder Zelle des Versuchsplans 10 mehrere Werte der AV Y stehen. Dann kann nämlich auch die Interaktion zwischen der unabhängigen Variablen B und dem Kontrollfaktor C geprüft werden (vgl. Bredenkamp, 1969a). Feldt (1958) hat die Anzahl von Modalitäten des Kontrollfaktors bestimmt, die zu einer möglichst großen Präzisionserhöhung führen. Diese optimale Anzahl ist um so größer, je höher die Korrelation zwischen der Kontroll- und der abhängigen Variable, je größer die Gesamtzahl N der Beobachtungen und je kleiner die Zahl K der experimentellen Bedingungen ist.

Eine Parallelisierung mit mehreren experimentellen Einheiten pro Zelle des Designs ist auch durchführbar, wenn hinsichtlich der Kontrollvariablen C die Probanden nur grob in Klassen eingeteilt werden können (z.B. hohe, mittlere, niedrige Intelligenz) oder wenn C gar nur eine Klasseneinteilung ohne Definition einer Rangordnung darstellt (z.B. männlich/weiblich). Eine Parallelisierung ist nicht auf Experimente mit einer einzigen experimentellen unabhängigen Variablen beschränkt, vielmehr können die Vpn innerhalb eines Blocks natürlich auch zufällig auf die möglichen Kombinationen der Ausprägungen mehrerer Behandlungsvariablen aufgeteilt werden. Generell ist bei der Inter-

pretation und Analyse von Versuchsplänen mit Parallelisierung aber zu beachten, daß die Block- oder Kontrollvariable einer organismischen Variablen (im weitesten Sinn) entspricht, deren Ausprägung für jeden Probanden „vorgegeben“ ist. Wegen der fehlenden Zufallsordnung ist es nicht möglich, den statistischen Zusammenhang zwischen solchen Kontrollvariablen und der abhängigen Variablen kausal zu interpretieren (vgl. Abschn. 3.3).

8.4.2 Kovarianzanalyse als Kontrolltechnik (StatV)

Bei der Parallelisierung wird die Fehlervarianz dadurch verringert, daß die mutmaßlich mit der AV verbundene Kontrollvariable als zusätzlicher Faktor in ein varianzanalytisches Design eingeführt wird. Bei der Kovarianzanalyse wird das gleiche Ziel zu erreichen versucht, indem man die Kontrollvariable C als unabhängige Variable im Sinne der Regressionsrechnung betrachtet. Die Kovarianzanalyse ist praktisch nichts anderes als eine Varianzanalyse, bei der aus den individuellen Unterschieden der Vpn auf der AV derjenige Teil herauspartialisiert wird, der mit Hilfe der linearen Regression auf die unterschiedlichen Werte auf der Kontrollvariablen zurückgeführt werden kann. Einzelheiten hierzu finden sich in der im Abschnitt 7.5.3.3 genannten Literatur sowie in den gesonderten Darstellungen der Kovarianzanalyse von Federer (1955, 483-522), Cochran (1957), Röhr (1975), Wildt & Ahtola (1976), Burnett & Barr (1977), Huitema (1980).

Gegenüber einer Varianzanalyse müssen bei Anwendung der Kovarianzanalyse mehrere zusätzliche *Voraussetzungen* erfüllt sein: Die Kontrollvariable muß quantitativ erfaßt sein, und innerhalb aller Behandlungsgruppen müssen die Regressionskoeffizienten sowie die Varianzen der Restvariablen Y^* gleich sein. (Zur Prüfung der statistischen Voraussetzungen siehe u.a. Kirk, 1968, 469-471 und Winer, 1971, 772-775; zu den Konsequenzen der Verletzung der Annahmen siehe Elashoff, 1969; Glass, Peckham & Sanders, 1972; Hamilton, 1976, 1977; Hollingsworth, 1980.)

Die statistischen Voraussetzungen sind stets dann nicht erfüllt, wenn man die Werte auf der Kontrollvariablen nach der experimentellen Behandlung ermittelt und wenn die Behandlungen sich unterschiedlich auf diese Werte auswirken (vgl. Evans & Anastasio, 1968; Sprott, 1970). Aus diesem Grund können mit Hilfe der Kovarianzanalyse auf keinen Fall Störfaktoren der internen Validität ausgeschaltet werden, d.h. ihre Anwendung kann einzig und allein der Erhöhung der Präzision dienen. (Zur Fehlanwendung der Kovarianzanalyse siehe weiter Lord, 1967; Harris, Bisbee & Evans, 1971; Overall & Woodward, 1977a, b.)

Da selbst bei Erfüllung der genannten Voraussetzungen die Kovarianzanalyse nur unter selten gegebenen Umständen zu einer wesentlich stärkeren Präzi-

sionserhöhung führt als eine Parallelisierung nach der entsprechenden Variablen (Cochran, 1957; Cox, 1957; Feldt, 1958), ist der Parallelisierung in der Regel der Vorzug zu geben. Mitunter kann aber deshalb nur eine Kovarianzanalyse durchgeführt werden, weil die Zeitspanne zwischen Erhebung der Kontrollvariablen und experimenteller Behandlung zu kurz ist, um parallele Gruppen zu bilden.

8.4.3 Homogenisierung als Kontrolltechnik (StatV)

Wenn die Varianz σ_c^2 zum Teil durch Unterschiede zwischen den Untersuchungseinheiten bedingt ist, kann sie verringert werden, indem man nur Vpn verwendet, die hinsichtlich einer oder mehrerer mit der AV zusammenhängenden Merkmalen homogen sind. Beispiel: Durch Prüfung unserer Hypothese WH_u ausschließlich an Personen mit (nahezu) gleichen Ausprägungen auf den Variablen „politisches Interesse“, „Alter“, „Art der Erfahrung mit Gastarbeitern“ und „Autoritätsgläubigkeit“ wird die Populationsvarianz in den Behandlungsgruppen wahrscheinlich verringert und die Präzision erhöht. Allerdings wird durch Beschränkung auf eine homogene Teilpopulation die Populationsvalidität stets vermindert. Die beste Lösung dieses Problems stellt die *systematische Variation* der vordem konstantgehaltenen Variablen dar. So könnte z.B. eine Untersuchung an mehreren, möglichst verschiedenartigen, in sich aber homogenen Probandengruppen durchgeführt werden, wodurch sowohl eine hohe Präzision als auch eine gute Populationsvalidität zu erreichen ist.

8.4.4 Konstanthaltung und Elimination als Kontrolltechniken (StatV)

Im Abschnitt 3.3 haben wir Konstanthaltung und Elimination als Techniken zur Sicherung der internen Validität vorgestellt. Wir werden jetzt sehen, daß sie auch der Erhöhung der Präzision dienen können. Die Fehlervarianz σ_c^2 entsteht auch durch Unterschiede hinsichtlich der genauen Bedingungen, unter denen die Werte der AV für jede experimentelle Einheit zustandekommen. Je stärker sich Bedingungen wie Darbietungszeit, Versuchsleiterverhalten, Lärmpegel, Gruppengröße, Tageszeit usw. für die einzelnen Vpn unterscheiden, desto größer wird σ_c^2 . Die Präzision eines Experiments läßt sich also dadurch erhöhen, daß man die Unterschiede in solchen situationalen Variablen minimiert oder daß man diese Variablen idealerweise sogar konstant hält (z.B. gleiche Darbietungszeiten) bzw. ganz eliminiert (z.B. Ausschaltung einer Lärmquelle).

Diese Konstanthaltung ist im Laboratorium leichter möglich als in „natürlichen“ Situationen. Deshalb sind Experimente in der Regel präziser als Feldexperimente; anders ausgedrückt: Zur Entdeckung eines in absoluten Einhei-

ten gemessenen experimentellen Effektes sind im Labor weniger Beobachtungen notwendig als „im Felde“. Auf der anderen Seite schränkt jede Konstanthaltung von (situationalen) Bedingungen die Situationsvalidität ein (vgl. Abschnitt 4.2). Im Interesse einer möglichst strengen Prüfung der betrachteten Hypothese ist es daher oft ratsamer, diese Bedingungen zufällig variieren zu lassen und die Zahl der Beobachtungen dementsprechend zu erhöhen.

8.4.5 Eingenistete Faktoren als Kontrolltechnik (StatV)

Eine weitere Möglichkeit zur Erhöhung der Präzision liegt in der Verwendung von hierarchischen Versuchsplänen, d.h. von Designs, bei denen die abhängige Variable nicht unter allen (theoretisch) möglichen Kombinationen von Modalitäten der unabhängigen Variablen beobachtet wird. Ein einfaches Beispiel möge dies verdeutlichen (vgl. Vpl. 11). Es sollen Hypothesen über die Wirksamkeit (AV Y) zweier Therapiemethoden A_1, A_2 in Kombination mit drei Arten flankierender Maßnahmen B_1, B_2, B_3 geprüft werden. Nehmen wir an, es gäbe drei Therapeuten C_1, C_2, C_3 , die die Methode A_1 durchführen können, und drei *andere* Therapeuten C_4, C_5, C_6 , die Methode A_2 beherrschen. Da die Modalitäten des Faktors C („Therapeuten“) nicht mit allen Modalitäten des Faktors A kombiniert werden können, bezeichnet man C als „in A eingenistet“ oder „geschachtelt“. Zur Prüfung der interessierenden Hypothesen wäre nur ein zweifaktorielles Design (vgl. Vpl. 6 in Abschn. 3.3.2) notwendig. Die Berücksichtigung des Faktors C hat aber (auch wenn er „eingenistet“ werden muß) den Vorteil, daß dadurch die Fehlervarianz verringert wird, die Präzision der Untersuchung also erhöht werden kann.

(Vpl. 11)

		R - U V B		
		B_1	B_2	B_3
R - U V A	A_1 N-UV C	C_1		
		C_2		
		C_3		
	A_2 N-UV C	C_4	Y_{i422}	
		C_5		
		C_6		

„N“ bedeutet „eingenistet“

Große Vorsicht ist bei der Verwendung hierarchischer Designs aber geboten, falls man auch über den eingestuteten Faktor Hypothesen prüfen will. Dies ist nur unter der Annahme möglich, daß bestimmte Interaktionen (in Vpl. 11 A x C und A x B x C) gleich Null sind (siehe dazu und zur Auswertung Winer, 1971, 359-366 und Bortz, 1979, 493-505).

8.4.6 Wiederholte Messungen als Kontrolltechnik (StatV)

Wir wollen unter der gewählten Überschrift, die den häufigsten Verwendungszweck von Meßwiederholungen enthält, einige wesentliche Aspekte derartiger Pläne im Zusammenhang behandeln und insbesondere auf Auswertungsprobleme eingehen.

Meßwiederholungs- oder „Within-Subjects“-Designs werden in der einschlägigen Literatur unter verschiedenen Schwerpunktsetzungen und Überschriften behandelt, etwa als „Cross-over“- , „Change-over“- , „(multiple) Zeitreihen“- oder auch „Vor-Nachtest-Pläne“ (vgl. Hedayat & Afsarineyad, 1975) sowie „Split-Plot“-Pläne (siehe Kirk, 1968, Kap. 8; Winer, 1971, Kap. 7) oder „gemischte Pläne“ (Lindquist, 1953, Kap. 18).

Gemeinsam ist all diesen Versuchsanordnungen, daß von einer Beobachtungseinheit (Vp) mehr als ein Beobachtungsdatum erhoben wird. Hierbei sind in Abhängigkeit von der WH zwei Situationen zu unterscheiden:

(1) Jede Vp „liefert“ mehrere Scores auf mehreren AVn Y_q („multiple criteria“); die Vpn müssen nicht notwendigerweise mehreren Treatments unterzogen werden. Derartige Versuchspläne sind ihrer Natur nach multivariat und dienen sehr oft der intraexperimentellen Replikation eines theoretischen Konzeptes (vgl. Abschnitt 2.2). Beziehen sich die wissenschaftlichen Hypothesen auf die Relation zwischen der (den) UV(n) und den AVn und ferner auf die Wechselwirkungen zwischen den letzteren, sollte eine multivariate Auswertung vorgenommen werden, da nur durch diese die Interaktionen zwischen den AVn erfaßt werden können - ein instruktives Beispiel findet sich etwa bei Gabriel & Glavin (1978).

Interessieren dagegen nur die Wirkungen der UVn auf die einzelnen AVn und/oder interagieren diese nicht oder nicht nennenswert, sollte die Auswertung i.a. über mehrere univariate Tests erfolgen, denen zur Kontrolle der Fehlerwahrscheinlichkeiten ein globaler multivariater Test vorangestellt werden kann. Bei dieser Vorgehensweise werden jedoch auch andere statistische Hypothesen geprüft als mit den multivariaten Verfahren - siehe Gabriel & Hopkins (1974) und Bredenkamp (1980, 87).

(2) Jede Vp wird nur hinsichtlich einer AV beobachtet und gemessen, aber sukzessiv unter mehreren Modalitäten der UV, so daß pro Vp ebenfalls mehr

als ein Wert resultiert (vgl. Abschnitt 3.1). Außer in den Fällen, in denen sich derartige Pläne aus der zu prüfenden wissenschaftlichen Hypothese ergeben, werden sie häufig angewandt, um die Anzahl der bei interindividueller Bedingungsvariation benötigten V_{pn} zu reduzieren und gleichzeitig die Präzision zu erhöhen.

Letzteres geschieht, indem durch den Versuchsplan (und ein entsprechend adaptiertes statistisches Ausgangsmodell) die Trennung der Varianz zu Lasten der individuellen Merkmale von der zu Lasten der Fehleranteile erfolgt, die ja beide bei interindividueller Bedingungsvariation zur Fehlervarianz σ_e^2 zusammengefaßt werden. Es resultiert eine für den statistischen Test der Treatmenteffekte um diesen individuellen Anteil reduzierte Prüfvarianz, sofern die übrigen Bedingungen unverändert bleiben - siehe dazu unten.

Wiederholte Messung als Technik zur Erhöhung der Präzision der Untersuchung kann als Spezialfall einer Parallelisierung betrachtet werden, bei der die Modalitäten des Kontrollfaktors den verschiedenen experimentellen Einheiten entsprechen (Vpl. 12).

(Vpl. 12)		W-UV B				
		B_1	...	B_m	...	B_K
Vpn-UV V (Untersuchungseinheiten- (Versuchspersonen- Variable)	V_1					
	$\vdots$					
	V_s			V_{sm}		
	$\vdots$					
	V_n					

„W“ bedeutet UV mit wiederholter Messung

Eine Erweiterung des dargestellten Versuchsplanes ergibt sich, wenn man Meßwiederholungen über mehr als eine UV durchführt. In diesem Fall wird jede experimentelle Einheit unter allen möglichen Bedingungskombinationen beobachtet.

Von „gemischten Plänen“ spricht man dann, wenn in einem Experiment zugleich die inter- und die intraexperimentelle Bedingungsvariation realisiert wird. Im einfachsten Fall werden etwa jeder der J Modalitäten der einen experimentellen UV, etwa A , n Untersuchungseinheiten zugeordnet, und jede die-

ser $J \cdot n$ Einheiten wird unter allen K Modalitäten der anderen UV, etwa B , beobachtet (Vpl. 13).

Im Vergleich mit vollständig randomisierten Plänen liegt der Vorteil von Meßwiederholungsplänen darin begründet, daß - wie bereits erwähnt - *unter sonst gleichen Bedingungen* eine höhere Präzision bei einer geringeren Anzahl von Vpn erzielt werden kann.

Allerdings ist hierbei zu berücksichtigen, daß sich die Vpn bei Mehrfachmessungen in einer anderen Situation befinden, sich anders verhalten als solche Vpn, die einem Treatment nur einmal unterzogen und beobachtet werden - vgl. Greenwald (1976, 212). Dies führt u.a. zu dem meist unvermeidbaren Auftreten der in den Abschnitten 3.2 und 3.4 aufgeführten unerwünschten Sequenzeffekte wie Übertragung, Übung, Ermüdung und Sensibilisierung, durch die die interne Validität bei Mehrfachmessungen an den gleichen Vpn schwieriger zu gewährleisten ist als bei interindividueller Bedingungsvariation (vgl. dazu auch Winer, 1971, 516-518).

(Vpl. 13)

			W-UV B				
			B_1	...	B_m	...	B_K
	Vpn-UV $V_{(A_1)}$ A_1	$V_{(A_1)1}$					
		$\vdots$					
		$V_{(A_1)s}$					
		$\vdots$					
		$V_{(A_1)n}$					
R-UV A	Vpn-UV $V_{(A_1)}$ A_1	$\vdots$					
		$V_{(A_1)1}$					
		$\vdots$					
		$V_{(A_1)s'}$			$Y_{s'lm}$		
		$\vdots$					
	Vpn-UV $V_{(A_j)}$ A_j	$V_{(A_1)n'}$					
		$\vdots$					
		$V_{(A_j)1}$					
		$\vdots$					
		$V_{(A_j)s''}$					
		$\vdots$					
		$V_{(A_j)n''}$					

Erlebacher (1977, 1978) hat in diesem Zusammenhang einen Versuchsplan und eine Auswertetechnik vorgeschlagen, mit denen festgestellt werden kann, inwieweit die Art der Bedingungsvariation (inter- vs. intraindividuell) die Beziehung zwischen UV und AV beeinflusst.

Insgesamt ist aus den aufgeführten Gründen von der routinemäßigen Erhebung von Meßwiederholungen mit den *ausschließlichen Zielen* der Präzisionssteigerung und/oder Versuchspersonenökonomie abzuraten.

Kann jedoch die wissenschaftliche Hypothese strenger mittels eines Meßwiederholungsplanes geprüft werden, sollte dieser auch realisiert werden. Je nach Art der zu prüfenden Hypothese(n) und den Voraussetzungen für bestimmte statistische Tests ist dann zwischen unterschiedlichen Analysestrategien zu wählen.

8.4.6.1 Analyse von Zeitreihen und Veränderungsmessungen

(1) Die WH bezieht sich auf Unterschiede oder Veränderungen in Abhängigkeit vom Zeitfaktor, dessen Einfluß daher bei der Analyse explizit zu berücksichtigen ist.

In diesem Fall liegt ein „Zeitreihenplan“ vor, dessen Auswertung in der Regel mittels der speziellen, unter dem Namen „Zeitreihenanalysen“ zusammengefaßten Verfahren vorgenommen werden sollte, die auch bei multifaktoriellen und multivariaten Plänen zur Anwendung kommen können. Kurze einführende Darstellungen in diese Techniken findet man etwa bei Rasch, Enderlein & Herrendörfer (1973), Dierkes (1977), Petermann (1978), Cook & Campbell (1979) und bei Revenstorf (1979); zur ausführlicheren, teilweise aber auch wesentlich schwierigeren Darstellung informiere man sich u.a. bei Wetzell (1970), Kendall (1973, 1976), Brillinger (1975), Chatfield (1975), Glass, Willson & Gottman (1975) sowie bei Box & Jenkins (1976); nicht-parametrische Zeitreihenanalysen stellt Lienert (1978) dar.

(2) Die WH beziehe sich auf Veränderungen im Sinne von Zuwachsraten etc., wobei der Zeitfaktor nicht explizit berücksichtigt wird.

Analysiert man die Daten als Differenz- oder Zuwachsrate („gain scores“, „change scores“), sind zahlreiche Probleme zu beachten, über die im einzelnen die Reader von Harris (1963) und de Gruijter & van der Kamp (1975, Teil 11) sowie die Arbeiten von Cronbach & Furby (1970), Linn & Slinde (1977), Rennert (1977), Petermann (1978) sowie Petermann & Hehl (1979) informieren. Diesen Problemen kann man teilweise ausweichen, indem man eine re-

gressionsanalytische Auswertung vornimmt, wie sie etwa Cohen & Cohen (1975, 377-393) skizziert haben.

8.4.6.2 Univariate und multivariate Analysen

Die WH beziehe sich auf Unterschiede im Verhalten der gleichen Vpn unter verschiedenen experimentellen Bedingungen, wobei oft der Einfluß des Zeitfaktors auf die AV etwa durch randomisierte Reihenfolgen der Behandlungen ausgeglichen oder eliminiert werden kann - siehe Abschnitt 3.4.

In Abhängigkeit von den Voraussetzungen, die im konkreten Fall erfüllt sind, kann man sich grundsätzlich zwischen den folgenden Strategien der Auswertung entscheiden:

- (1) Man berechnet exakte univariate F-Tests, die auf vglw. restriktiven Voraussetzungen beruhen - siehe Abschnitt 8.4.6.2.1.
- (2) Man berechnet approximative univariate F-Tests, die bzgl. der Verletzung einer wesentlichen Voraussetzung adjustiert sind - siehe Abschnitt 8.4.6.2.1.
- (3) Man führt multivariate Varianz- bzw. Regressionsanalysen durch, die auf weniger restriktiven Bedingungen beruhen - vgl. Abschnitt 8.4.6.2.2.
- (4) Man nimmt eine nicht-parametrische Auswertung der Daten vor, die an relativ schwächere Voraussetzungen gebunden ist - vgl. Abschnitt 8.4.6.2.3.

Wir gehen im folgenden kurz auf die verschiedenen Auswertungstechniken ein.

8.4.6.2.1 Exakte und approximative univariate Tests

Die im Abschnitt 8.2 angesprochenen Voraussetzungen zur Durchführung von Varianz- bzw. regressionsanalytischen F-Tests gelten uneingeschränkt auch für Versuchspläne mit wiederholten Messungen. Diese sind vornehmlich dadurch gekennzeichnet, daß die Beobachtungen unter den verschiedenen Treatments nicht unabhängig voneinander sind, weil sie jeweils an den gleichen Vpn erhoben wurden (vgl. Versuchsplan 12 in Abschnitt 8.4.6).

Zur Prüfung von Hypothesen über den wiederholten Faktor B mit K Modalitäten (s.u.) wird vorausgesetzt, daß die $n \cdot K$ Beobachtungen einer K-dimensionalen Normalverteilung mit dem Mittelwertsvektor μ und der (Varianz-)Kovarianz-Matrix Σ entstammen. Die Populations-Kovarianz-Matrix Σ über alle Paare von (potentiellen) Beobachtungswerten unter Faktor B hat folgende allgemeine Form (vgl. etwa Kirk, 1968; Tatsuoaka, 1971; Moosbrugger, 1978):

$$(8.10) \quad \Sigma = \begin{bmatrix} B_1 & B_2 & \dots & B_m & \dots & B_K \\ \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1m} & \dots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2m} & \dots & \sigma_{2K} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_m^2 & \dots & \sigma_{mK} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{K1} & \sigma_{K2} & \dots & \sigma_{Km} & \dots & \sigma_K^2 \end{bmatrix} \begin{matrix} B_1 \\ B_2 \\ \dots \\ B_m \\ \dots \\ B_K \end{matrix}$$

Hypothesen über den Mittelwertsvektor μ werden i.a. mittels des folgenden univariaten F-Tests geprüft (vgl. Winer, 1971, 281):

$$(8.11) \quad F = \frac{s_{\text{Treatment B}}^2}{s_{\text{p}n \times \text{Treatment B}}^2}; \quad df = (K - 1); \quad (n - 1)(K - 1)$$

Dieser Test ist bspw. dann valide, d.h. F ist unter Gültigkeit der Nullhypothese zentral F-verteilt, wenn die Kovarianz-Matrix folgende spezifische Struktur aufweist:

$$(8.12) \quad \Sigma = \begin{bmatrix} \sigma_e^2 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sigma_e^2 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_e^2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & \sigma_e^2 \end{bmatrix} = \sigma_e^2 \cdot I_K$$

I_K bezeichnet die $K \times K$ Einheitsmatrix; ferner ist $\sigma_k^2 = \sigma_e^2$ und $\sigma_{mm'} = 0$ für alle $m \neq m'$. Dies bedeutet, daß alle Beobachtungen unabhängig voneinander und daß alle Varianzen homogen sind (vgl. Abschnitt 8.4.2).

Die Werte des F-Bruches sind allerdings auch dann F-verteilt, wenn die Varianzen und die Kovarianzen homogen sind, wenn also gilt:

$$(8.13) \quad \Sigma = \begin{bmatrix} \sigma_e^2 & R \cdot \sigma_e^2 & \dots & R \cdot \sigma_e^2 & \dots & R \cdot \sigma_e^2 \\ R \cdot \sigma_e^2 & \sigma_e^2 & \dots & R \cdot \sigma_e^2 & \dots & R \cdot \sigma_e^2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ R \cdot \sigma_e^2 & R \cdot \sigma_e^2 & \dots & R \cdot \sigma_e^2 & \dots & \sigma_e^2 \end{bmatrix}$$

R bezeichnet die Populationskorrelation zwischen Beobachtungen an zwei beliebigen Meßzeitpunkten.

Weisen Matrizen die vorstehende Form auf, spricht man von „Uniformität“ („compound symmetry“) (vgl. Winer, 1971, 277; Davidson, 1972). Uniformität der Kovarianz-Matrix ist jedoch trotz anderslautender Behauptungen (e.g. Davidson, 1972; Poor, 1973; Keppel, 1973) keine notwendige, sondern lediglich eine hinreichende Bedingung für die Validität des F-Bruches (8.11).

Um die notwendigen und hinreichenden Bedingungen für die Validität des obigen F-Bruches zur Auswertung eines einfachen Meßwiederholungsplanes wie Versuchsplan 12 (s.o.) kurz skizzieren zu können, gehen wir davon aus, daß die $n \cdot K$ Beobachtungen in Form einer $n \times K$ Matrix angeordnet sind, deren n Zeilen die Beobachtungen jeder V_p auf den K Variablen (Modalitäten von B) enthalten (siehe Versuchsplan 12 in Abschnitt 8.4.6). Die folgenden Definitionen beziehen sich auf diese Rahmenbedingungen und sind zum Verständnis der Voraussetzungen unerlässlich (vgl. Rouanet & Lepine, 1970).

Ein *Kontrast* zwischen den Treatmentbedingungen des wiederholten Faktors B bezeichnet einen Vektor c , dessen K Komponenten sich zu Null aufaddieren. Die Norm dieses Kontrastes ist definiert als $(c' \cdot c)^{1/2}$, und ein *standardisierter* Kontrast hat die Norm 1. Die $K \times f$ Matrix C wird dann als *Kontrast-Matrix* bezeichnet, wenn ihre f Spaltenvektoren Kontraste darstellen. Sind die f Spaltenvektoren paarweise orthogonal, heißt C spaltenweise oder *f-orthogonal*, und sind zudem alle Spalten standardisiert, wird C *spaltenweise* oder *f-orthonormal* genannt.

Eine *f-orthonormale* $K \times f$ Matrix kann daher bspw. die folgende Form aufweisen (vgl. Mendoza, Toothaker & Crain, 1976; Rogan, Keselman & Mendoza, 1979; Huynh & Mandeville, 1979):

$$(8.14) \quad C = \begin{bmatrix} 1/c_1 & 1/c_2 & \dots & 1/c_f \\ -1/c_1 & 1/c_2 & \dots & 1/c_f \\ 0 & -2/c_2 & \dots & 1/c_f \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -f/c_f \end{bmatrix}$$

mit $c_g = g(g+1)^{1/2}$, $g = 1, \dots, f$.

Für eine Matrix der Form (8.14) gilt:

$$(8.15) \quad C'C = I_f, \text{ mit } I_f : f \times f \text{ Einheitsmatrix.}$$

Ein *Vergleich* zwischen den K Stufen des wiederholten Faktors B wird durch die Kontrast-Matrix C repräsentiert, und die in ihr enthaltenen f Kontrastvektoren können mit der Einschränkung beliebig gewählt werden, daß sie orthonormiert sein müssen (s. o.). Dabei bezeichnet f die Freiheitsgrade dieses Vergleichs (siehe Rouanet & Lepine, 1970; Rogan, Keselman & Mendoza, 1979, 275).

Rouanet & Lepine (1970, 152) unterscheiden nun u.a. die folgenden Arten von Vergleichen: Ein Vergleich mit $df = 1$ wird *eindimensional* genannt; er kann durch einen einzigen Kontrast repräsentiert werden. Ein Vergleich mit $df = f$ heißt *partial*, wenn $f < K - 1$, und *global* („Overall“), wenn $f = K - 1$.

Interessiert nun der *globale* Vergleich, ist die Kontrast-Matrix als $K \times (K - 1)$ Matrix mit $K - 1$ orthonormalen Spaltenvektoren zu konstruieren, und der entsprechende F-Test (8.11) ist genau dann valide, wenn für die Kovarianz-Matrix Σ gilt (vgl. Rouanet & Lepine, 1970; Huynh & Feldt, 1970):

$$(8.16) \quad C'\Sigma C = \sigma_e^2 I_{K-1}$$

M. a. W. : Können die K ursprünglichen Variablen, für die jeweils n wiederholte Messungen vorliegen, unter Verwendung einer geeigneten orthonormalen Kontrast-Matrix (s.o.) in $K - 1$ orthonormale Variablen transformiert werden, die unabhängig voneinander und homogen variant sind, kann die statistische Signifikanz des wiederholten Faktors mittels F-Test valide beurteilt werden (Huynh & Mandeville, 1979, 965; Rogan, Keselman & Mendoza, 1979). In diesem Fall besitzt die Kovarianz-Matrix die Eigenschaft der „Zirkularität“ („circularity“, „sphericity“) *bzgl. des gewählten Vergleichs*, die die notwendige und hinreichende Bedingung für den o.a. F-Test darstellt (vgl. zum Beweis Rouanet & Lepine, 1970 sowie Huynh & Feldt, 1970).

Liegt der zweifaktorielle Versuchsplan 13 (s. o.) mit wiederholten Messungen auf einem Faktor (B) vor, ist mit jeder der J Stufen von A eine eigene Kovarianz-Matrix Σ_j verbunden. Die Tests für B und die Interaktion $A \times B$ sind nur dann valide, wenn gilt, daß die J Kovarianz-Matrizen $C'\Sigma_j C$ auf allen Stufen von A gleich sind und wenn die gemeinsame Matrix $C'\Sigma C$ zirkulär ist („multisample circularity“; vgl. Huynh & Mandeville, 1979; Rogan, Keselman & Mendoza, 1979). In einem zweifaktoriellen Plan mit zwei wiederholten Faktoren B mit K Stufen und D mit M Stufen erfolgt die Testung von B gegen $s_{V_{pn} \times B}^2$, von D gegen $s_{V_{pn} \times D}^2$ und von $B \times D$ gegen $s_{V_{PN} \times B \times D}^2$ (Eimer, 1978, 118ff.). Hier ist die Kovarianz-Matrix Σ vom Typ $KM \times KM$, und mit jedem Test ist eine eigene Kontrast-Matrix C verbunden. Diese ist für den Faktor B vom Typ $KM \times (K - 1)$, für den Faktor D vom Typ $KM \times (M - 1)$ und für die Interaktion vom Typ $KM \times (K - 1)(M - 1)$. Mit jedem der drei durch seine zugehörige Kontrast-Matrix repräsentierten globalen Vergleiche ist eine eigene Zirkularitätsbedingung der Form (8.16) verknüpft; die drei Skalare σ_B^2 , σ_D^2 und σ_{BD}^2 müssen nicht gleich sein (siehe zu den Einzelheiten Mendoza, Toothaker & Crain, 1976; Huynh & Mandeville, 1978; vgl. auch Scheifley & Schmidt, 1978). In dieser Literatur finden sich auch Hinweise auf die Validitätsbedingungen in weiteren Versuchsplänen mit wiederholten Messungen auf einigen oder allen Faktoren.

Die Überprüfung der Zirkularität der Matrizen kann über die entsprechenden Tests nach Mauchly (1940) und Box (1950) erfolgen, die u.a. in Kirk (1968), Winer (1971), Morrison (1976) und Bortz (1979) dargestellt sind; eine Erweiterung des Tests von Mauchly (1940) hat Mendoza (1980) vorgeschlagen. Die genannten Verfahren sind jedoch nicht robust gegenüber Verletzungen der Annahme der multinormalen Verteilung der Fehlerterme und darüber hinaus

sehr teststark, weswegen i. a. die Hypothese der Zirkularität zurückgewiesen werden dürfte (vgl. Davidson, 1972; Rogan, Keselman & Mendoza, 1979). Die Anwendung dieser Verfahren zur Überprüfung der Validitätsbedingungen der F-Tests ist daher in der Regel nicht zu empfehlen, wie Keselman et al. (1980) ausführen und begründen.

Führt man ungeachtet der Struktur der Kovarianz-Matrizen univariate F-Tests durch,³⁰⁾ ist je nach Ausmaß der Abweichung von der oben spezifizierten Struktur der Kovarianz-Matrizen teilweise mit starkem Anwachsen der tatsächlichen Wahrscheinlichkeiten für Fehler 1. Art zu rechnen (Box, 1954 b; Wilson, K., 1975). Abweichungen von dieser Struktur scheinen darüber hinaus in der Empirie die Regel zu sein (Davidson, 1972; Keppel, 1973, 464; Wilson, R. S., 1975; Greenwald, 1976; Rogan, Keselman & Mendoza, 1979; vgl. in diesem Zusammenhang auch Wallenstein & Fleiss, 1979).

Die verbreitete Auswertung von Meßwiederholungsplänen über konventionelle F-Tests ist daher insgesamt wenig empfehlenswert.

Die Abweichung einer Kovarianz-Matrix von der Zirkularität hinsichtlich des globalen Vergleichs kann mittels des von Box (1954 b) abgeleiteten Index' ϵ erfaßt werden, für den gilt:

$$(8.17) \quad \epsilon = \frac{(\text{tr } \mathbf{C}'\mathbf{\Sigma} \mathbf{C})^2}{(K-1) \text{tr}(\mathbf{C}'\mathbf{\Sigma} \mathbf{C})^2}, \text{ mit tr: Spur der genannten Matrix.}$$

(vgl. Rouanet & Lepine, 1970, 156; siehe ferner Geisser & Greenhouse, 1958; Greenhouse & Geisser, 1959; Keselman & Mendoza, 1979). Der Wert des Maßes ϵ kann aus den empirischen Daten geschätzt werden - hierüber informieren Huynh & Feldt (1976) und Huynh (1978). Box (1954b) hat zeigen können, daß unter Verwendung von ϵ die Freiheitsgrade des univariaten F-Tests so adjustiert werden können, daß ein approximativer, aber hinreichend valider Test selbst dann resultiert, wenn die Abweichungen von der spezifizierten Struktur der Matrizen sehr ausgeprägt sind (vgl. Collier et al., 1967; Davidson, 1972; Gaito, 1973; Mendoza, Toothaker & Nicewander, 1974; Wilson, K., 1975; Wilson, R. S., 1975; Rogan, Keselman & Mendoza, 1979).

Diese approximativen Tests bieten sich daher zur Prüfung univariater Hypothesen an; siehe im einzelnen Huynh (1978).

³⁰⁾ über die entsprechenden Tests informieren etwa Winer (1971, 261-273, 514-594), Myers (1972, 168-186), Keppel (1973, 401-408), Cohen & Cohen (1975, 403-412) und Eimer (1978); zur Kodierung der Versuchspersonen siehe Pedhazur (1977).

8.4.6.2.2 Multivariate Tests

Die multivariaten Varianz- und Regressionsanalysen sowie verwandte Verfahren (vgl. Knapp, 1978) ermöglichen die Prüfung von Hypothesen über die unspezifische Wirkung von einer oder mehreren UVn auf mehrere AVn Y_q und die Wechselbeziehungen zwischen letzteren (vgl. Moosbrugger, 1978, 104). Ihre valide Anwendung ist u. a. mit der Annahme multinormal verteilter Modellresiduen verknüpft (siehe zu Einzelheiten etwa Moosbrugger, 1978, 121), wobei moderate Abweichungen tolerabel sind (e.g. Olson, 1974). Ferner müssen die Kovarianz-Matrizen für alle Bedingungen des nicht-wiederholten Faktors homogen sein, wobei allerdings keine spezifischen Annahmen bzgl. ihrer Struktur erforderlich sind (vgl. Olson, 1974; Bredenkamp, 1980, 84; Stevens, 1980; Rogan, Keselman & Mendoza, 1979). Insofern ist die valide Anwendung entsprechender multivariater Tests zur Auswertung von Meßwiederholungsplänen an insgesamt etwas schwächere Bedingungen geknüpft als die univariater Tests, wobei erstere allerdings auch bei gleichen Stichprobenumfängen anfälliger gegenüber Heterogenität der Kovarianz-Matrizen sind als die adjustierten univariaten Tests (Olson, 1974; Huynh, 1978).

Generell stellt sich bei multivariaten Analysen die Frage nach dem unter Robustheitsaspekten „optimalen“ Testkriterium; hierauf gehen auch die bekannteren Lehrbücher kaum ein (vgl. die Angaben im Abschnitt 7.5.3.3). Von den verschiedenen Simulationsstudien, mit denen eine Beantwortung dieser Frage versucht wurde, verdienen insbesondere die von Olson (1974, 1976, 1979) und Stevens (1979, 1980) Erwähnung; spezielle Robustheitsuntersuchungen haben in jüngerer Zeit Everitt (1979) sowie Hakstian, Roed & Lind (1979) durchgeführt. Aus Olsons und Stevens' Befunden geht hervor, daß i.a. die Verwendung des von Bartlett (1937, 1939) und Pillai (1955) entwickelten V-Kriteriums zur geringsten Anzahl falscher Entscheidungen führen dürfte, wenn man davon ausgeht, daß dem E die „wahre“ Populationssituation (e. g. die Art der Nicht-Zentralitäts-Struktur) unbekannt ist.

Zur Frage der Anschlußtests nach einem signifikanten globalen Test siehe insbesondere Morrison (1976) sowie Stevens (1972b), Spector (1977) und Ramsey (1980).

Insgesamt stellen die multivariaten Ansätze eine Alternative zu den im übrigen nicht weniger rechenaufwendigen approximativen univariaten F-Tests dar. Die Entscheidung zugunsten einer der beiden Auswertetechniken sollte daher im Einzelfall vor allem davon abhängig gemacht werden, welche der jeweils geprüften statistischen Hypothesen am ehesten zu einer möglichst strengen Prüfung der wissenschaftlichen Hypothese führt.

Bzgl. weiterer Einzelheiten siehe neben der bereits aufgeführten Literatur u.a. Hummel & Sligo (1971), Poor (1973) sowie Romaniuk, Herbert & Levin (1977).

8.4.6.2.3 Nicht-parametrische Tests

Die adäquaten nicht-parametrischen Auswerteverfahren sind an wesentlich schwächere Voraussetzungen gebunden (vgl. Abschnitt 7.5.3). Allerdings stehen diese Techniken nur für vglw. einfache Meßwiederholungspläne zur Verfügung, so daß sie insgesamt kaum eine Alternative zu den parametrischen Tests darstellen dürften (vgl. McCall & Appelbaum, 1973), es sei denn, ein zu geringes Skalenniveau erzwingt ihren Einsatz. In diesem Fall können dann komplexe Pläne oft nicht analog ausgewertet werden. Bzgl. der Einzelheiten zu den Verfahren sei auf Lienert (1973, 345-369; 1978, 1045-1054), Siegel (1976, Kap. 7) und Marascuilo & McSweeney (1977, 354-388) verwiesen sowie ferner auf die im Abschnitt 7.5.3.1 genannte Literatur zur Auswertung nominaler Daten.

Abschließend sei nochmals hervorgehoben, daß die interne und möglicherweise auch die Variablen-Validität bei Meßwiederholungen in den meisten Fällen nicht hinreichend gesichert werden kann (vgl. auch die Abschnitte 3.2 und 3.4).

8.4.7 *Zur Beziehung zwischen der Präzision und den anderen Aspekten der experimentellen Validität*

Wir sind in diesem Abschnitt 8.4 mehrfach darauf gestoßen, daß durch Techniken zur Erhöhung der Präzision andere Aspekte der experimentellen Validität beeinträchtigt werden können: Bei Konstanthaltung von Situationsvariablen wird die Situationsvalidität herabgesetzt, mit homogenen Versuchspersonengruppen hat man eine geringere Populationsvalidität, bei Meßwiederholung ist die interne Validität schwieriger zu gewährleisten, die Einführung von Kontroll- oder Kovariablen kann zu Erwartungseffekten führen, wenn diese Variablen in Vortests erhoben werden, usf.

In Lehrbüchern der Versuchsplanung wird häufig auch empfohlen, die Präzision zu erhöhen, indem ein komplexeres Design verwendet wird: Durch Einführung weiterer systematischer Faktoren kann den Daten ein jeweils komplexeres Modell angepaßt werden. Hierdurch wird es möglich, aus der ursprünglichen Fehlervarianz solche Anteile herauszulösen, die durch die zusätzlichen Faktoren „erklärt“ werden können - Ähnliches geschieht bei der Parallelisierung, Kovarianzanalyse und der Meßwiederholung. Gegen ein unreflektiertes Befolgen dieser Empfehlung sprechen jedoch mehrere Gründe:

1. Die Beziehung zwischen der Zahl der Faktoren und der Präzision des Experiments gilt „unter sonst gleichen Bedingungen“, d.h. u.a. bei gleichbleibender Stichprobengröße pro Zelle (Bedingungskombination). Je mehr Faktoren untersucht werden, desto höher wird der Gesamtversuchspersonenbedarf, wenn man das n pro Zelle konstant halten will.
2. Je mehr Faktoren ein Versuchsplan umfaßt, desto stärker kumulieren bei dem gebräuchlichen Auswertungsvorgehen (Durchführung aller Tests) die Fehlerwahrscheinlichkeiten α und β . Darüber hinaus erfolgt in derartigen Plänen die Prüfung der statistischen Hypothesen i.a. gegen nur eine oder aber nur wenige Prüfvarianzen. Dies hat zur Folge, daß die einzelnen F-Brüche nicht mehr unabhängig voneinander sind (Kimball, 1951; Hurlburt & Spiegel, 1976; Hays, 1977, 516f., 589f.). Beide Probleme lassen sich u.E. nur dann weitgehend vermeiden, wenn man die Anzahl der Faktoren ausschließlich nach dem Kriterium der strengen Prüfung der WH festlegt und sich bei der Auswertung auf die statistischen Hypothesen beschränkt, die im Hinblick auf die WH relevant sind (Hager & Westermann, im Druck, a, b).

Aus diesen Einwänden geht hervor, daß es keineswegs stets dem Ziel möglichst strenger Prüfungen von Kausalhypothesen dienen muß, wenn man durch die Anwendung von Kontrolltechniken das Experiment so präzise wie möglich macht. Daher lehnen wir auch die starre sog. „MaxMinKon-Strategie“ (vgl. Wormser, 1974, Kap. 5; Kerlinger, 1978, 450-459) ab, nach der ein Experiment *stets* u.a. so zu planen ist, daß durch die möglichst umfassende Kontrolle von unsystematischen Einflüssen die Fehlervarianz minimiert wird.

8.5 Falsche Analyse und Interpretation statistischer Interaktionen

Wir haben bereits erwähnt, daß sich Hypothesen über das Vorliegen statistischer Interaktionen zwischen (jeweils) zwei oder mehr Variablen unmittelbar aus der betrachteten wissenschaftlichen Hypothese oder Theorie ergeben können (siehe Abschnitt 8.1.1, Punkt (9)) und daß sie von Interesse sind, wenn untersucht wird, inwieweit ein vorhergesagter Unterschied zwischen Behandlungsbedingungen von der Ausprägung einer (oder mehrerer) (störender) Populations- oder Situationsvariablen abhängig ist (vgl. Abschnitt 4.3).

Über die Bedeutung, Prüfung und Interpretation statistischer Interaktionen herrscht nach unserem Eindruck einige Unklarheit, zumal man in nur ganz wenigen Lehrbüchern eine genügend differenzierte Darstellung findet (e. g. in Lindquist, 1953; Diehl, 1979; Henning & Muthig, 1979). Diese Unklarheit kann zu falschen Entscheidungen über die Gültigkeit wissenschaftlicher Hypothesen führen, weswegen wir im folgenden einige grundsätzliche Erläute-

rungen zu diesem Thema geben wollen. Hierbei beschränken wir uns auf die sog. „varianzanalytischen Interaktionen“ und gehen auf diejenigen, die als „Zusammenhang zwischen kategorialen Merkmalen“ definiert sind, nicht ein - nähere Einzelheiten hierzu entnehme man der im Abschnitt 7.5.3.1 angegebenen Literatur zur Auswertung nominaler Daten.

Ob man eine Interaktion erfassen kann, ist von der Art des statistischen Modells abhängig, das man den Daten anpassen will. Nicht immer sind Interaktionen empirisch sinnvoll interpretierbar oder aber von theoretischem Interesse. In einem solchen Fall kann der E ein Modell anpassen, das keine oder aber eine reduzierte Anzahl von Interaktionsparametern enthält. Dies hat oft den Vorteil, daß interessierende Hypothesen etwa mittels hierarchischer Pläne (vgl. Abschnitt 8.4.5) ökonomischer geprüft werden können; zur Frage der prüfbaren Hypothesen und der alternativen Modelle vgl. neben den Standardlehrbüchern der Versuchsplanung und -auswertung etwa Elston & Bush (1964), Marascuilo & Levin (1970, 1976), Levin & Marascuilo (1972, 1973), Garnes (1973, 1978c) sowie Betz & Gabriel (1978). Allerdings ist bei der Elimination von Interaktionsparametern bei der Modellanpassung stets die Bedeutung zu berücksichtigen, die Interaktionen für die Populations- und Situationsvalidität zukommen kann (vgl. dazu Abschnitt 4.3); wir greifen diesen Gedanken im folgenden auf.

Hat die AV Rangniveau, empfiehlt sich die nicht-parametrische Auswertung des entsprechenden Versuchsplanes über die Rangtests nach Kruskal & Wallis (1952) sowie Friedman (1937) (vgl. auch Lienert, 1973), mit denen allerdings keine Interaktionen erfaßt werden können. Um dies zu ermöglichen, hat Brendenkamp (1974, 1980, 73f.) bestimmte Modifikationen dieser Verfahren vorgeschlagen, auf die wir hier nicht eingehen - siehe zu Einzelheiten auch Gebert (1977), Krüger (1977) und Engelhardt (1979).

Die folgenden Ausführungen zur Interaktion bei intervallskalierter AV gelten im wesentlichen auch bei rangskalierter AV.

8.5.1 Das Konzept der statistischen Interaktion

In einem Versuchsplan, mit dem die simultane Wirkung von mindestens zwei (quantitativen oder qualitativen) UVn auf die quantitative(n) AV(n) untersucht wird, können unabhängig vom Skalenniveau der AV stets statistische Interaktionen auftreten. Man bezeichnet häufig die Wechselwirkung zwischen (jeweils) zwei UVn als „Interaktion 1. Ordnung“, während die „Interaktionen höherer Ordnung“ durch das simultane Wirken von jeweils mehr als zwei UVn auf die AV(n) charakterisiert sind.

Die folgenden Darstellungen beschränken sich auf die Interaktionen 1. Ordnung, und wir gehen der Einfachheit halber von einem vollständig gekreuzten zweifaktoriellen varianzanalytischen Versuchsplan mit zwei UVn A und B mit J resp. K fixierten Modalitäten aus. Jeder der K-J Bedingungskombinationen AB wurden zufällig n Vpn zugewiesen. Wir nehmen an, daß die Voraussetzungen zur Durchführung einer Varianzanalyse via F-Tests gegeben sind.

In einem derartigen Experiment werden in jeder Zeile A_i K Mittelwerte M_{ik} und in jeder Spalte B_k J Mittelwerte M_{jk} auftreten, die als Stichprobenäquivalente der Modellparameter μ_{jk} angesehen werden.

Man bezeichnet nun die Unterschiede zwischen allen Mittelwerten in einer Zeile A_i als den „einfachen Haupteffekt des Faktors B in der Modalität A_i des Faktors A“ oder kurz als „einfachen (Haupt-)Effekt (von) B in A_i “; entsprechend heißen die Unterschiede zwischen allen Mittelwerten der Spalte B, „einfacher (Haupt-)Effekt (von) A in B“.

Der Durchschnitt der einfachen Haupteffekte eines Faktors X über alle Stufen des jeweils anderen Faktors, m.a. W. die Unterschiede zwischen den Rand- oder Marginalmittelwerten, nennt man den „Haupteffekt des Faktors X“, wobei „X“ entweder gleich A oder aber gleich B ist. Tab. 8.1 möge dies etwas veranschaulichen.

Tabelle 8.1: Einfache Effekte und Haupteffekte.

	Faktor B								Zeilen- mittelwerte
	B_1	B_2	...	...	B_m	.	.	.	B_K
Faktor A	A_1	einfacher Effekt B in A_1							
	A_2	einfacher Effekt B in A_2							
	...	...							
	A_i	einfacher Effekt B in A_i							
	...	...							
	A_j	einfacher Effekt B in A_j							
					einfacher Effekt A in B_m				einfacher Effekt A in B_K
									Haupteffekt A
									(über alle Spalten B_k)
Spalten- mittel- werte									Gesamt- mittel- wert
									Haupteffekt B (über alle Zeilen A_j)

Unabhängig von der Gleichheit oder Ungleichheit der Randmittelwerte zu Lasten eines Haupteffektes können Unterschiede zwischen den Zellenmittelwerten μ_{jk} bestehen, die *nicht* auf die Haupteffekte zurückführbar sind. Bestehen zwischen den Populationszellenmittelwerten μ_{jk} derartige Unterschiede oder Differenzen, spricht man von „statistischer Interaktion“ oder „Wechselwirkung“. Lindquist (1953, 126) definiert einen Interaktionseffekt „als einen Unterschied zwischen einem einfachen Effekt und dem korrespondierenden Haupteffekt“. Eine Wechselwirkung liegt m.a. W. dann vor, wenn die einfachen Effekte A nicht auf allen Stufen des Faktors B gleich sind und/oder umgekehrt (vgl. dazu im einzelnen Lindquist, 1953; Lee, 1961; Lubin, 1961; Digman, 1966; Bracht & Glass, 1968; Bracht, 1970; Plomp, 1974; Bredenkamp, 1975).

Im Fall der hier vorausgesetzten Versuchsanordnung (s.o.) kann den Daten das folgende additive Modell angepaßt werden (vgl. dazu auch Abschnitt 8.2):

$$(8.18) \quad Y_{ijk} = \mu + (\mu_j - \mu) + (\mu_k - \mu) + (\mu_{jk} - \mu_j - \mu_k + \mu) + e_{ijk} \\ = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + e_{ijk}$$

Hierbei gilt:

$$(8.19) \quad \alpha_j = (\mu_j - \mu): \text{Haupteffekt A}$$

$$(8.20) \quad \beta_k = (\mu_k - \mu): \text{Haupteffekt B}$$

$$(8.21) \quad (\alpha\beta)_{jk} = (\mu_{jk} - \mu) - (\mu_j - \mu) - (\mu_k - \mu) = \\ = (\mu_{jk} - \mu_j - \mu_k + \mu): \text{Interaktionseffekt AB}$$

Man ersieht aus Gleichung (8.21), daß die Interaktion AB denjenigen Variationsanteil darstellt, der in den einzelnen Zellen des Designs verbleibt, nachdem von dem Abweichungsausdruck $(\mu_{jk} - \mu)$ die Variation zu Lasten der beiden Haupteffekte A und B subtrahiert worden ist.

Der entsprechende varianzanalytische F-Test prüft die H_0 , daß diese „Reste“, quadriert und aufsummiert, in der Population gleich Null sind, daß also keine Interaktion vorliegt. Fiktive Populationsmittelwerte, die einer Null-Interaktion entsprechen, sind in Tab. 8.2 enthalten.

Tabelle 8.2: Null-Interaktion: Mittelwerte.

	B ₁	B ₂	B ₃
A ₁	2	8	5
A ₂	4	10	7

Wir wollen dann auf eine Null-Interaktion in der Population erkennen, wenn der erwähnte F-Test bei genügender Teststärke zu einem insignifikanten oder aber praktisch nicht bedeutsamen Ergebnis führt. Ergibt der Test jedoch ein statistisch und praktisch bedeutsames Resultat, gilt es, verschiedene Typen der Interaktion zu unterscheiden.

8.5.2 Definition verschiedener Typen der Interaktion

Im Abschnitt 4.3 haben wir ausgeführt, daß eine Variable genau dann einen Störfaktor der Populations- oder auch der Situationsvalidität darstellt, wenn sie mit dem Treatmentfaktor „disordinal interagiert“. Zur Veranschaulichung haben wir dort mit der folgenden Mittelwertstabelle gearbeitet, die wir hier nochmals aufnehmen, ohne die Treatments inhaltlich zu spezifizieren:

Tabelle 8.3: Mittelwerte bei disordinaler Interaktion für Faktor A.

	B ₁	B ₂	B ₃
A ₁	0	5	10
A ₂	2	4	6

Den Daten, die wir als Populationsmittelwerte auffassen wollen, kann entnommen werden, daß das Treatment A₂ nur unter der Modalität B₁ zu höheren Werten auf der AV Y führt; unter den beiden übrigen Modalitäten des Faktors B wirkt A₁ (zahlenmäßig) stärker auf die AV. Die geprüfte wissenschaftliche Hypothese WH, (vgl. Abschnitt 4.3) gilt also nur für die mit B₁ symbolisierte Personengruppe oder Situation oder - allgemeiner ausgedrückt - Faktormodalität; die Populations- oder die Situationsvalidität ist daher eingeschränkt.

Zu einer anderen Interpretation würde man gelangen, wenn die *gleichen* Mittelwerte wie in Tab. 8.3 eine *andere Anordnung* aufweisen würden, etwa die folgende :

Tabelle 8.4: Mittelwerte bei nicht-disordinaler Interaktion bzgl. A.

	B ₁	B ₂	B ₃
A ₁	0	4	6
A ₂	2	5	10

In diesem Beispielfall liegen unter der Bedingung A_2 durchgängig höhere Werte auf der AV Y vor als unter A_1 ; dabei sind allerdings die Unterschiede innerhalb der drei Bedingungen B_1 , B_2 und B_3 nicht gleich. Die geprüfte WH_u gilt demnach für alle unter B_k repräsentierte Personen, Situationen oder Faktormodalitäten, und es liegt daher keine Beeinträchtigung der Populations- oder der Situationsvalidität vor.

Wie man aus diesen Beispielen deutlich erkennen kann, bedeutet die Tatsache, daß eine statistisch signifikante Interaktion vorliegt-hiervon sind wir ausgegangen -, nicht notwendigerweise eine Beeinträchtigung der Populations- oder auch der Situationsvalidität. Diese ist nur dann gegeben, wenn ein ganz bestimmter Typ von Interaktion vorliegt: die bereits mehrfach angesprochene Disordinalität (bezüglich eines Faktors x).

Die Unterscheidung zwischen zwei Typen der Interaktion geht u.W. auf Lindquist (1953, 141 f.) zurück: und Lubin (1961) benutzte die Bezeichnungen „disordinal“ für den in Tab. 8.3 repräsentierten Interaktionstyp und „ordinal“ für die in der Tab. 8.4 enthaltene spezielle Anordnung oder Struktur von Mittelwerten. Die Bedeutung dieser Unterscheidung wurde jedoch nur in wenigen Teilbereichen der Psychologie beachtet, etwa im Rahmen des Konzepts der „Aptitude-Treatment-Interaction“ (ATI) - vgl. Cronbach (1957, 1975), Verreck (1974), Schwarzer & Steinhagen (1975), Treiber & Petermann (1976) sowie Garten (1980); siehe ferner Borich & Godbout (1974), Loftus (1978) und Henning (1978) sowie die oben genannte Literatur.

8.5.2.1 Disordinale Interaktion

Um zu einer formalen Definition der beiden Interaktionstypen gelangen zu können, wollen wir uns etwas näher mit den obigen Tabellen befassen, und zwar zunächst mit der Tabelle 8.3.

In dieser Tabelle drückt sich die Disordinalität der Interaktion darin aus, daß die Differenz der Mittelwerte für A in B_1 ein anderes Vorzeichen („minus“) aufweist als die entsprechende Differenz in B_2 („plus“) und B_3 (ebenfalls „plus“). Im Vergleich zu den Mittelwerten in den beiden letztgenannten Spalten sind die Mittelwerte in Spalte B_1 daher „invertiert“.

Wir bezeichnen im folgenden jeweils diejenigen Mittelwerte als „invertiert“, deren Vorzeichen bei Differenzbildung unterschiedlich sind. Dabei ist es wichtig, daß die Differenzbildung zwischen jeweils „korrespondierenden Mittelwerten“ vorgenommen wird. Unter „korrespondierenden Mittelwerten“ verstehen wir die Mittelwerte innerhalb eines einfachen Haupteffektes in Relation zu den Mittelwerten eines beliebigen anderen einfachen Haupteffektes für den gleichen Faktor. Es korrespondieren in diesem Sinne also allgemein bzgl.

Faktor A die Mittelwerte $\mu(A_i B_k)$ mit den Mittelwerten $\mu(A_i', B_k)$, und für unser Beispiel in Tab. 8.3 ergeben sich damit die folgenden Differenzen zwischen bzgl. A korrespondierenden Mittelwerten:

$$\mu(A_1 B_1) - \mu(A_2 B_1) = 0 - 2 = -2 \text{ (einfacher Haupteffekt A in } B_1)$$

$$\mu(A_1 B_2) - \mu(A_2 B_2) = 5 - 4 = +1 \text{ (einfacher Haupteffekt A in } B_2)$$

$$\mu(A_1 B_3) - \mu(A_2 B_3) = 10 - 6 = +4 \text{ (einfacher Haupteffekt A in } B_3)$$

Überträgt man die (Populations-)Mittelwerte in einen Graphen, wobei auf der Ordinate die Werte der AV und auf der Abszisse derjenige Haupteffekt abgetragen wird, für den der Interaktionstyp *nicht* festgestellt werden soll, so entspricht jeder Inversion von Mittelwerten eine Überkreuzung von Regressionsgeraden als Verbindungslinien zwischen je zwei Mittelwerten. Für das Beispiel aus Tab. 8.3 ergibt sich daher genau eine derartige Überschneidung - vgl. Abb. 8.1.

Will man diese Ausführungen in eine Definition der Disordinalität umsetzen, muß man berücksichtigen, daß wir stets von Populationsmittelwerten ausgegangen sind. In der Stichprobe kann leicht eine Inversion von Zellenmittelwerten M_{jk} auftreten, die statistisch nicht bedeutsam oder signifikant ist, der also in der Population gleiche Zellenmittelwerte entsprechen.

Nach dieser Überlegung wollen wir in Anlehnung an Bracht & Glass (1968), Bracht (1970) und Plomp (1974) eine disordinale Interaktion in der folgenden Weise definieren:

Die Hypothese, daß eine Interaktion in der Population disordinal bzgl. eines Haupteffektes X ist, soll angenommen werden, wenn für mindestens zwei Differenzen zwischen korrespondierenden Zellenmittelwerten in der Stichprobe gilt, daß

- (1) *ihre algebraischen Vorzeichen („plus“ oder „minus“) entgegengesetzt sind, und daß*
- (2) *diese Differenzen statistisch signifikant sind, und daß*
- (3) *die experimentellen Effekte, die mit diesen statistischen Signifikanzen verknüpft sind, „praktisch bedeutsam“ sind, d.h. einen bestimmten Mindestwert nicht unterschreiten (vgl. zum Konzept der „praktischen Bedeutsamkeit (oder Signifikanz)“ Abschnitt 9.2 und Teil 11).*

Wenden wir uns nun der ordinalen Interaktion zu!

8.5.2.2 Ordinale Interaktion

Eine ordinale Interaktion bzgl. Faktor ist bei den Mittelwerten in Tab. 8.4 gegeben, die wir wieder als Populationsmittelwerte auffassen wollen. Um fest-

stellen zu können, in welcher Weise sich dieser Interaktionstyp von einer disordinalen Interaktion unterscheidet, bilden wir zunächst die Differenzen korrespondierender Mittelwerte:

$$\mu(A_1B_1) - \mu(A_2B_1) = 0 - 2 = -2$$

$$\mu(A_1B_2) - \mu(A_2B_2) = 4 - 5 = -1$$

$$\mu(A_1B_3) - \mu(A_2B_3) = 6 - 10 = -4$$

Da alle Differenzen das gleiche Vorzeichen aufweisen, können keine Inversionen korrespondierender Mittelwerte vorliegen. Werden also Mittelwerte graphisch veranschaulicht, die einer bzgl. Faktor x ordinalen Interaktion entsprechen, treten keine Überschneidungen von Regressionsgeraden auf, wenn man die Populationsmittelwerte betrachtet - vgl. Abb. 8.3.

Nach diesen Ausführungen definieren wir in Anlehnung an Bracht & Glass (1968), Bracht (1970) und Plomp (1974) eine ordinale Interaktion wie folgt:

Die Hypothese, daß eine Interaktion in der Population bzgl. eines Haupteffektes X ordinal ist, soll angenommen werden, wenn für alle Differenzen zwischen jeweils korrespondierenden Zellenmittelwerten in der Stichprobe gilt, daß

- (1) *sie alle das gleiche Vorzeichen aufweisen oder aber höchstens teilweise vom Betrage Null sind, oder daß*
- (2) *diejenigen Differenzen, für die (1) nicht gilt, statistisch nicht signifikant, oder praktisch nicht bedeutsam sind.*

Um also eine statistische Interaktion als „ordinal“ beurteilen zu können, dürfen also keine statistisch abgesicherten Mittelwertsinversionen auftreten.

Wenden wir nun diese Definition an, um zu beurteilen, von welchem Typ die Interaktion bzgl. des bislang nicht betrachteten Faktors B in der Tab. 8.3 ist! Es lassen sich die folgenden drei Gruppen von je zwei korrespondierenden Mittelwertsdifferenzen bilden:

$$\begin{array}{l} \text{I:} \quad \mu(A_1B_1) - \mu(A_1B_2) = 0 - 5 = -5 \\ \quad \mu(A_2B_1) - \mu(A_2B_2) = 2 - 4 = -2 \end{array}$$

$$\begin{array}{l} \text{II:} \quad \mu(A_1B_1) - \mu(A_1B_3) = 0 - 10 = -10 \\ \quad \mu(A_2B_1) - \mu(A_2B_3) = 2 - 6 = -4 \end{array}$$

$$\begin{array}{l} \text{III:} \quad \mu(A_1B_2) - \mu(A_1B_3) = 5 - 10 = -5 \\ \quad \mu(A_2B_2) - \mu(A_2B_3) = 4 - 6 = -2 \end{array}$$

Da in den drei Paaren jeweils korrespondierender Mittelwertsdifferenzen nur jeweils gleiche Vorzeichen auftreten, ist die betr. Interaktion ordinal bzgl. Faktor B.

Die zugehörige graphische Darstellung findet sich in Abb. 8.2, und die Abb. 8.4 enthält die Mittelwerte aus Tab. 8.4, in der die Interaktion bzgl. Faktor B ebenfalls ordinal ist, wie man sich leicht selbst überzeugen kann.

Bredenkamp (1975, 1980) geht von einer anderen Definition der Ordinalität der Interaktion aus, die darauf hinausläuft, daß Null-Differenzen (s.o.) im Sinne der Disordinalität interpretiert werden (vgl. Bredenkamp, 1975, 793, 798). Stanley (1973) unterscheidet die beiden Interaktionstypen nach dem Kriterium ihrer Transformierbarkeit: ordinale Interaktionen sind durch geeignet gewählte Transformationen der Werte der AV eliminierbar, disordinale dagegen nicht - vgl. dazu auch Smith (1976 a) und Busemeyer (1980).

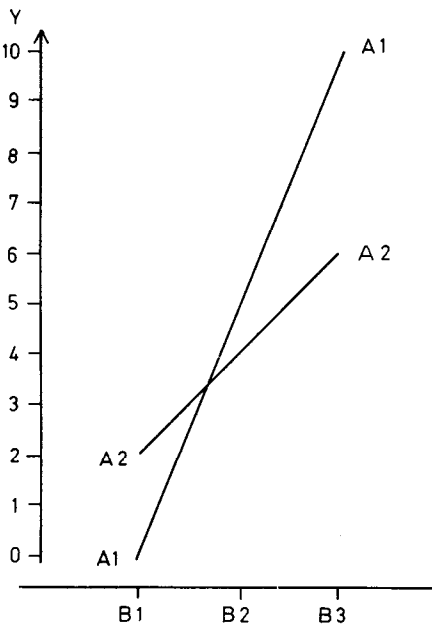


Abb. 8.1: Disordinale Interaktion für Faktor A; Daten aus Tab. 8.3.

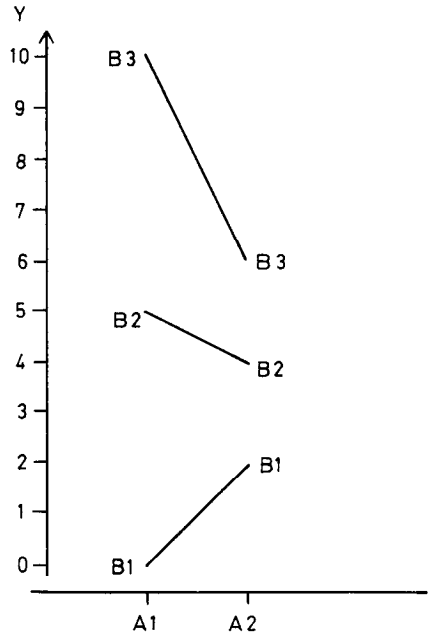


Abb. 8.2: Ordinale Interaktion für Faktor B; Daten aus Tab. 8.3.

8.5.2.3 Zur graphischen Darstellung von Interaktionen

Die Definition verschiedener Interaktionstypen abschließend, ist noch eine wichtige Anmerkung angebracht. Häufig wird in der Literatur nicht differenziert, für welchen Faktor eine Interaktion ordinal oder disordinal ist, sondern lediglich allgemein festgestellt, eine Interaktion sei ordinal (oder disordinal). Diese Beurteilung erfolgt oft ausschließlich aufgrund der graphischen Darstellung der Stichprobenzellenmittelwerte. üblicherweise wird dabei jedoch nur einer der beiden möglichen Graphen konstruiert. Tritt dann der Fall auf, daß

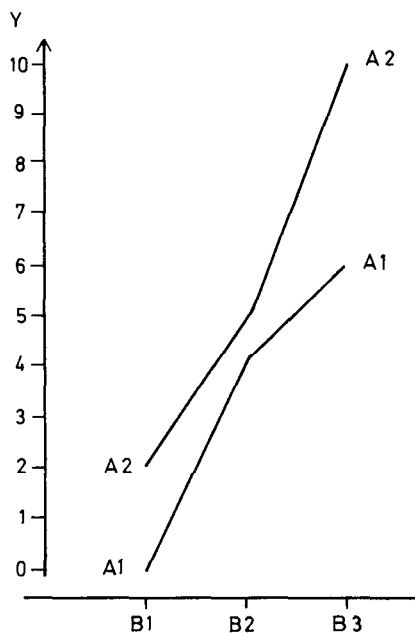


Abb. 8.3: Ordinale Interaktion für Faktor A; Daten aus Tab. 8.4.

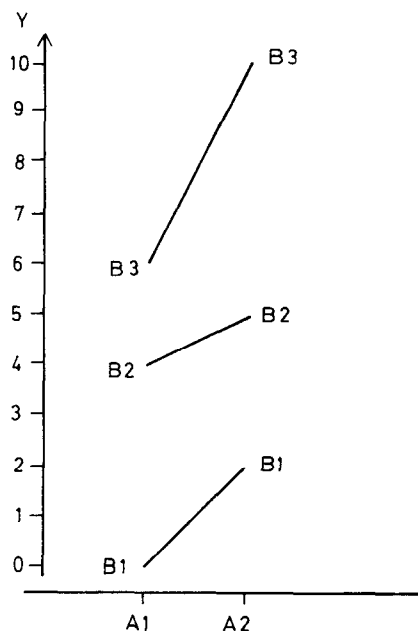


Abb. 8.4: Ordinale Interaktion für Faktor B; Daten aus Tab. 8.4.

die Interaktion für den einen Faktor ordinal, für den anderen aber disordinal ist (vgl. Tab. 8.2), hängt es von der (meist eher zufällig) gewählten Darstellungsform ab, ob sich die Regressionsgeraden schneiden oder nicht (vgl. Abb. 8.1 und 8.2). Durch diese Praxis werden falsche Schlußfolgerungen wie die folgende begünstigt: „It should be emphasized that ‚ordinality‘ and ‚disordinality‘ are properties of graphs“ (Glass & Stanley, 1970, 411); oder Huck & Sutton (1975, 190): „Unfortunately, whether or not an interaction is ordinal can depend upon the way the researcher constructs his graph“; oder auch: „Es muß betont werden, daß ‚Ordinalität‘ und ‚Disordinalität‘ Eigenschaften der graphischen Darstellung von Interaktionen sind“ (McGuigan, 1979, 144).

Selbstverständlich ist der Typ der Interaktion unabhängig von der Art und Weise, in der er graphisch veranschaulicht wird. Zur Vermeidung derartiger Fehlschlüsse und darauf aufbauender Interpretationen bzgl. der Populations- und/oder Situationsvalidität ist es grundsätzlich empfehlenswert, entsprechende Forschungsbefunde auf *beide* Arten darzustellen.

8.5.3 Ein Verfahren zur Unterscheidung zwischen den Interaktionstypen

Sich überkreuzende Regressionsgeraden, die Stichprobenmittelwerte miteinander verbinden, sprechen nicht notwendig für die Disordinalität der Interaktion in der Population; sie stellen lediglich eine notwendige, nicht jedoch hinreichende Bedingung für die Annahme der Hypothese der Populations-Disordinalität dar. Diese Tatsache führt uns zu der Frage, auf welche Art man über den Typ der Interaktion in der Population aufgrund von Stichprobendaten entscheiden kann.

Zur Beantwortung dieser Frage muß zunächst bedacht werden, daß ein statistisch signifikanter F-Wert für den Test der Interaktion lediglich aussagt, daß die Differenzen zwischen den Zellenmittelwerten in der Population nicht gleich Null sind, daß also in der Population keine Null-Interaktion vorliegt.

Die speziellen Muster oder Strukturen von Mittelwerten, die den verschiedenen Interaktionstypen entsprechen, können mit einem einzelnen F-Wert selbstverständlich nicht erfaßt werden. Dies bedeutet, daß zur Beurteilung des Interaktionstyps i. a. weitere Signifikanztests einzusetzen sind. Um ein allzu starkes Anwachsen der beiden Fehlerwahrscheinlichkeiten zu vermeiden, ist anzustreben, die Zahl dieser Tests so gering wie möglich zu halten. Nach diesem Kriterium sind die verschiedenen in der Literatur vorgeschlagenen mehrstufigen Teststrategien zur Beantwortung der Frage nach dem Interaktionstyp nicht befriedigend (vgl. Bracht & Glass, 1968; Plomp, 1974; Bredenkamp, 1975, 1980).

Als „Übergangslösung“ wollen wir an dieser Stelle eine zwar im Sinne Henning & Muthigs (1979, 205) „quasistatistische“, u.E. aber pragmatische Strategie in ihren Grundzügen kurz skizzieren, die unter Ausnutzung der Variabilität der Daten die Information erbringt, von welchem Typ die Interaktion ist und ggf. auf welche Modalitäten die (Dis-)Ordinalität beschränkt ist.

Nachdem der Test der Interaktion bei genügender Teststärke zu einem signifikanten Ergebnis geführt hat, stellt man zunächst für beide Faktoren getrennt fest, ob irgendwo zwischen jeweils zwei Mittelwertspaaren Inversionen vorliegen. Ist dies für einen oder beide Faktoren nicht der Fall, kann man nach unserer Definition auf Ordinalität der Interaktion für diesen Faktor in der Population erkennen.

Haben sich dagegen Inversionen ergeben, bestimmt man die sog. „minimale statistisch signifikante Differenz (LSD)“ (vgl. zur Berechnung u.a. Winer, 1971, 199f.) und die dem praktisch bedeutsamen experimentellen Mindesteffekt entsprechende „minimale praktisch signifikante Differenz (LPD)“.

Anschließend vergleicht man für jeden einfachen Haupteffekt eines Faktors den maximalen und den minimalen empirischen Mittelwert mit der LSD und der LPD. Alle einfachen Haupteffekte, in denen diese Vergleiche nicht zur Überschreitung der *beiden* kritischen Differenzen führen, bleiben von der weiteren Betrachtung ausgeschlossen, da es in ihnen keine signifikanten Differenzen gibt. Für die übrigen einfachen Haupteffekte wird nun nach Inversionen korrespondierender Mittelwerte gesucht, und jede festgestellte Inversion wird mit der LSD und der LPD verglichen. Ergeben diese Vergleiche für mindestens eine Inversion, daß die beiden Mittelwertsdifferenzen statistisch und praktisch signifikant sind, ist die Interaktion für den untersuchten Faktor disordinal.

Es bleibt anzumerken, daß die hier (u.W. in dieser Form erstmals) vorgestellte Verfahrensweise mit steigender Anzahl der Modalitäten zu anwachsenden Fehlerraten führt. Dies Anwachsen ist jedoch wesentlich geringer als das mit der von Bredenkamp (1975) vorgeschlagenen Teststrategie verbundene; und darüber hinaus wird nur dann auf Disordinalität erkannt, wenn mindestens eine *Konjunktion* von Tests zu einem statistisch und praktisch signifikanten Resultat führt. In Abhängigkeit von den jeweils zu prüfenden WH können sich andere Strategien empfehlen (s. Hager, im Druck, a).

8.6 Zusammenfassung

Die Validität einer empirischen Untersuchung zur Prüfung einer wissenschaftlichen Hypothese kann u. a. dadurch vermindert werden, daß man zu falschen Aussagen über die Gültigkeit der aus der wissenschaftlichen Hypothese abgeleiteten statistischen Hypothese gelangt. Die maximalen Wahrscheinlichkeiten α und β für die beiden Fehlentscheidungen über die Gültigkeit einer statistischen Hypothese sollten nach unseren Überlegungen in den Teilen 6 und 7 durch den Experimentator von vornherein auf möglichst niedrige Werte fixiert werden. Ungeachtet dieser Festlegung kann es jedoch aufgrund von Fehlern bei der Prüfung statistischer Hypothesen dazu kommen, daß sich die tatsächlichen Wahrscheinlichkeiten für falsche Entscheidungen erhöhen; diese Fehler haben wir als „Störfaktoren der statistischen Validität“ bezeichnet. Die statistische Validität von Untersuchungen wird häufig dadurch gestört, daß anstelle der von der wissenschaftlichen Hypothese implizierten eine ganz andere statistische Hypothese geprüft wird. Insbesondere müßten in der Psychologie statt der verbreiteten Varianz- und Regressionsanalysen viel öfter spezielle Mittelwertvergleiche wie die Prüfung monotoner Trendhypothesen angestellt werden, zumindest im Anschluß an signifikante globale Tests (Abschnitt 8.1 und 8.2). Um eine Kumulierung von Fehlerwahrscheinlichkeiten weitestgehend zu vermeiden, ist in Abhängigkeit von der geprüften wissenschaftlichen Hypo-

these entweder die Wahrscheinlichkeit α oder aber β für jeden einzelnen Test gegenüber den „konventionellen“ Werten herabzusetzen.

Für die abgeleitete statistische Hypothese muß dann ein angemessenes statistisches Prüfverfahren gewählt werden. Sind die Annahmen der klassischen parametrischen Tests (vgl. Abschnitt 8.2.1) nicht erfüllt, kann es zu einer Erhöhung der Fehler-Wahrscheinlichkeiten kommen, die allerdings i. a. nur unter besonders ungünstigen Umständen erhebliche Ausmaße annehmen dürften (Abschnitt 8.2). Muß man davon ausgehen, daß zwei oder mehr Annahmen nicht aufrechterhalten werden können, sollte man die Daten i. a. nicht mittels eines parametrischen Verfahrens auswerten.

Im Abschnitt 8.4 haben wir einige Methoden angesprochen, die die Varianz der abhängigen Variablen innerhalb der Behandlungspopulationen verringern und dadurch die Präzision der statistischen Hypothesenprüfung erhöhen können: die Parallelisierung der Untersuchungseinheiten (Vpn) hinsichtlich individueller Merkmale, die Kovarianzanalyse, die Beobachtung aller Vpn unter mehreren (oder allen) Modalitäten eines Faktors (Meßwiederholung), die Verwendung homogener Probandengruppen sowie die Konstanthaltung und Elimination situationaler Variablen. Bei der Anwendung dieser Techniken ist jedoch zweierlei zu beachten: Erstens führen gerade die bekanntesten von ihnen (Parallelisierung und Meßwiederholung) nur unter ganz bestimmten Bedingungen tatsächlich zu einer erhöhten Präzision, und zweitens kann der Einsatz dieser Techniken andere Aspekte der experimentellen Validität beeinträchtigen. Bei Meßwiederholungen ist bspw. die interne Validität meist nur unzureichend zu sichern, während Homogenisierung, Konstanthaltung und Elimination leicht zur Beeinträchtigung der Populations- und/oder der Situationsvalidität führen können.

Im Abschnitt 8.5 haben wir auf eine häufig vernachlässigte Störung der statistischen Validität hingewiesen: die falsche Analyse und Interpretation statistischer Interaktionen. Um beurteilen zu können, ob das Vorliegen einer Interaktion die Interpretation der Ergebnisse statistischer Tests über Mittelwerte (oder Ränge) beeinflusst, muß festgestellt werden, von welchem Typ diese Interaktion ist. Die Disordinalität der Interaktion führt zu Einschränkungen der Interpretation hinsichtlich der Populations- oder der Situationsvalidität, während dies bei Vorliegen einer ordinalen Interaktion nicht der Fall ist. Um den Typ einer Interaktion bzgl. eines Faktors mit einem Minimum an statistischen Tests feststellen zu können, wurde ausgehend von einer Definition der Disordinalität eine neue Strategie vorgestellt.

9. Maße der statistischen Assoziation: Die experimentellen Effekte

9.1 Einleitung

In den vorangegangenen Teilen haben wir jegliche Abweichung eines Parameters von seinem Erwartungswert unter Gültigkeit der statistischen Null-Hypothese mit „experimenteller Effekt (EE)“ bezeichnet.

Der Terminus „experimenteller Effekt“ ist demnach ein allgemeiner Ausdruck für systematische Beziehungen, d.h. die statistische Assoziation zwischen den jeweils untersuchten Variablen; in einem Experiment bspw. entsteht diese Assoziation zwischen UV(n) und AV(n) infolge der experimentellen Behandlungen.

Bei der Bestimmung des EE mittels Stichprobendaten muß stets damit gerechnet werden, daß unsystematische Fehler zu Verzerrungen führen, d.h. im Extremfall entweder zu „Scheineffekten“ oder aber zu „fälschlichen Null-Effekten“.

Der Signifikanztest dient unter diesem Aspekt der Beantwortung der Frage, ob ein aus den Daten ermittelter Effekt als durch die unsystematischen Fehler zustande gekommen erklärt werden kann oder ob er eine tatsächlich vorhandene systematische Beziehung zwischen UV und AV reflektiert. Unter sonst gleichen Bedingungen wird ein EE um so eher als „systematisch“ oder „echt“ deklariert, je mehr Vpn zu seiner Entdeckung verwendet wurden.

Da das Vorhandensein eines EE von der Tatsache der statistischen Signifikanz unabhängig ist, sollte seine Größe ungeachtet des Ausganges des Signifikanztests stets aus den erhobenen Daten ermittelt werden (vgl. Abschnitt 7.4).

Spätestens hierbei stellt sich die Frage nach der (formalen) Definition von experimentellen Effekten.

Grundsätzlich sind EE stets unter Verwendung derjenigen Größen definiert, über die die statistischen Hypothesen formuliert werden - also etwa über Mittelwerte, Korrelationskoeffizienten, Ränge oder Wahrscheinlichkeiten. Auf diese Definitionen gehen wir im folgenden detaillierter ein, wobei wir erneut die auf Intervall-Informationen beruhenden EE ausführlicher darstellen werden.

Zuvor jedoch ist eine grundsätzliche Unterscheidung anzusprechen.

9.2 Experimentelle Effekte und praktische Bedeutsamkeit

Die EE wurden in der psychologischen Versuchsplanungsliteratur u.W. von Edwards (1950, 30f.) unter dem Namen „practical significance“ eingeführt und finden sich unter wechselnden Bezeichnungen bei anderen Autoren. Hodges & Lehmann (1954, 261) sprechen etwa von „material significance“, Bolles & Messick (1959) von „statistical utility“, während Kendall & Stuart (1961, 161) die Bezeichnung „magnitude of a difference“ wählen. Hays (1963, 326) nennt den EE „degree (oder auch: „strength“) of true association“; Levy (1967) und Gold (1969) bezeichnen den EE als „Substantive significance“; Vaughan & Corballis (1969) reden von „strength of effect“ und Cohen (1969) kurz von „effect size“. Für den deutschen Sprachraum haben offenbar Lienert & Orlik (1966, 215) die Bezeichnung „praktische Signifikanz“ geprägt, die Bredenkamp (e. g. 1970) später aufgegriffen hat.

Wir plädieren dafür, die Begriffe „experimenteller Effekt“ einerseits und „praktische Bedeutsamkeit“ (oder: „praktische Signifikanz“) andererseits voneinander zu trennen.

Unter „experimenteller Effekt“ (EE) verstehen wir die rein statistische Assoziation, die sich in den Daten manifestiert bzw. aus diesen zu bestimmen ist und die entweder in Roheinheiten (etwa als Mittelwertsdifferenz) oder in normierten Einheiten (etwa als Korrelationskoeffizient; s.U.) angegeben werden kann. Der Ausdruck „experimenteller Effekt“ bezeichnet damit ein ausschließlich statistisches Konzept und bezieht sich auf das Ausmaß der Abweichung eines Parameters von der Erwartung unter einer einfachen Null-Hypothese. Ob man diese Abweichung im konkreten Einzelfall inhaltlich als „praktisch bedeutsam“ im Hinblick auf die zu prüfende wissenschaftliche Hypothese interpretieren will, ist eine Entscheidung, für die es ein statistisches Rationales u.W. nicht gibt, für die also der E andere Kriterien angeben sollte - vgl. zu Einzelheiten Teil 11.

Wir befassen uns im Teil 9 ausschließlich mit den Maßen für die statistische Assoziation bzw. für den experimentellen Effekt in der Population und in der Stichprobe und greifen das Konzept der praktischen Bedeutsamkeit erst im Teil 11 wieder auf.

9.3 Experimentelle Effekte bei parametrischen Hypothesen

Parametrische Hypothesen beziehen sich in der überwiegenden Mehrzahl der Fälle auf Mittelwerte bzw. Regressionskoeffizienten und quadrierte Korrelationen. Die Prüfung dieser Hypothesen erfolgt in aller Regel über die Teststatistiken F und t.

Im Abschnitt 7.3.2 haben wir festgestellt, daß diese Statistiken bei Vorhandensein eines experimentellen Effektes in der Population, EEP, nicht-zentralen Verteilungen folgen. Das Ausmaß der Nicht-Zentralität, d.h. der Abweichung von den zentralen Verteilungen, wird durch die Nicht-Zentralitätsparameter erfaßt. Diese enthalten also Informationen über den EEP.

9.3.1 Maße der Nicht-Zentralität: λ , φ^2 und f^2

Die in diesem Abschnitt vorzustellenden Maße der Nicht-Zentralität sind allesamt unter Verwendung der quadrierten Differenz der K Treatment-Mittelwerte μ_k von ihrem Gesamtmittelwert μ definiert; trifft H_0 zu, müssen alle Mittelwerte μ_k identisch sein. Je größer also diese (quadrierten) Abweichungen sind, desto größer ist auch unter sonst gleichen Bedingungen der EEP.

Ferner ist allen Nicht-Zentralitätsparametern gemeinsam, daß die quadrierten Mittelwertsabweichungen auf die Populationsfehlervarianz σ_e^2 normiert werden, also als Abweichungen relativ zum Fehler ausgedrückt werden. Durch diese Normierung wird die Vergleichbarkeit der in möglicherweise sehr unterschiedlichen Experimenten ermittelten Abweichungen angestrebt, die eine Voraussetzung zur Tabellierung der Gütefunktionen darstellt (vgl. Abschnitt 7.4).

Unter der Annahme gleicher Größe der K (Sub-)Populationen ergibt sich folgende allgemeine Definition der Nicht-Zentralitätsparameter:

$$(9.1) \quad \text{Nicht-Zentralität:} = c \cdot \frac{\sum_{k=1}^K (\mu_k - \mu)^2 / K}{\sigma_e^2} = c \cdot \frac{\sigma_b^2}{\sigma_e^2}$$

Die in der Literatur auffindbaren Nicht-Zentralitätsparameter, die sehr uneinheitlich symbolisiert werden, unterscheiden sich im wesentlichen durch die Größe, die an die Stelle der (allgemeinen) Konstanten c in (9.1) tritt. Der Parameter λ , zuweilen auch als δ^2 bezeichnet, ergibt sich, indem man $c = n \cdot K$ setzt, wobei n die Stichprobengröße pro experimenteller Bedingung und K die Anzahl dieser Bedingungen bezeichnet (vgl. etwa Kirk, 1968, 107) (wir gehen hier wie im folgenden stets von gleichen Stichprobenumfängen aus):

$$(9.2) \quad \lambda = \frac{n \cdot K \cdot \sigma_b^2}{\sigma_e^2} = \frac{N \cdot \sigma_b^2}{\sigma_e^2}$$

Tang (1938), der sich als erster mit der Tabellierung der nicht-zentralen F-Verteilungen befaßt hat, benutzte einen aus λ abgeleiteten Index φ , der der Quotient aus der Standardabweichung der wahren Mittelwertsdifferenzen zu

Lasten des experimentellen Treatments und dem Standardfehler des Mittelwerts ist (vgl. Tang, 1938, 138); für das Quadrat dieses Index' gilt daher bei c n:

$$(9.3) \quad \varphi^2 = \frac{n \sum_{k=1}^K (\mu_k - \mu)^2 / K}{\sigma_e^2} = \frac{n \cdot \sigma_b^2}{\sigma_e^2}$$

Einen weiteren Index für die Nicht-Zentralität hat Cohen (1969, 267-280) vorgeschlagen. Sein Maß f^2 erlaubt die Beschreibung einer Populationssituation unabhängig von der Stichprobengröße n bei c = 1:

$$(9.4) \quad f^2 = \frac{\sum_{k=1}^K (\mu_k - \mu)^2 / K}{\sigma_e^2} = \frac{\sigma_b^2}{\sigma_e^2}$$

Zwischen den drei vorgestellten Nicht-Zentralitätsparametern bestehen folgende wichtige Beziehungen, aus denen sich weitere Zusammenhänge durch einfache algebraische Umformungen ableiten lassen:

$$(9.5) \quad f^2 = \frac{\varphi^2}{n} = \frac{\lambda}{n \cdot K}$$

Neben diesen drei Nicht-Zentralitätsparametern als normierte EEP finden sich in der Literatur weitere, die meist auf die hier angegebenen zurückgeführt werden können - vgl. zu Einzelheiten etwa Winer (1971, 826f.). Bei der Beschäftigung mit diesen Maßen sollte stets beachtet werden, daß die Symbole sehr uneinheitlich verwendet werden.

9.3.2 Korrelationskoeffizienten und -quotienten

Die Nicht-Zentralitätsparameter können als Funktionen interpretiert werden, die die Stärke eines statistischen Zusammenhangs auf den Bereich der positiven reellen Zahlen von 0 bis $+\infty$ abbilden. Diese Art der Abbildung ist insgesamt allerdings weniger gebräuchlich als diejenige in den Bereich der positiven reellen Zahlen von 0 bis 1; hierfür ist etwa der Korrelationskoeffizient r (als normierte Kovarianz) ein Beispiel.

Bei dieser Normierungsart wird die Varianz zu Lasten der verschiedenen Treatments, also $\text{Var}(b)$, nicht mehr auf die Fehlervarianz $\text{Var}(e)$ relativiert, sondern auf die totale Varianz in der AV, nämlich $\text{Var}(t)$, die sich bei Zugrundelegung des ALM (vgl. Abschnitt 8.2) additiv aus den beiden anderen Varianzen

ergibt (vgl. zu den Ableitungen im einzelnen etwa Hays, 1977, Kap. 12). Man nennt die entstehenden Quotienten „Maße der „erklärten“ oder „aufgeklärten“ Varianz“ (Bredenkamp, 1970; zur Kritik dieser Bezeichnung vgl. etwa Guttman, 1977):

$$(9.6) \quad \text{„erklärte“ Varianz:} = \frac{\text{Var}(t) - \text{Var}(e)}{\text{Var}(t)}$$

In Abhängigkeit davon, welche Varianzen man in die allgemeine Formel (9.6) einsetzt, erhält man normierte Maße für den EEP oder für den EES, den experimentellen Effekt in der Stichprobe.

9.3.2.1 Populationsmaße: η^2 , ω^2 und $R_{Y.X}^2$

Ersetzt man die allgemeinen Varianzterme in (9.6) durch die entsprechenden Populationsvarianzen, erhält man:

$$(9.7) \quad \eta^2 = \frac{\sigma_t^2 - \sigma_e^2}{\sigma_t^2} = \frac{\sigma_b^2}{\sigma_t^2} = 1 - \frac{\sigma_e^2}{\sigma_t^2}$$

Diese Quotientenbildung geht auf Pearson (e. g. 1911) zurück und das Resultat wird nach Fisher (1928) in der englisch-sprachigen Literatur „correlation ratio“ genannt. Im deutschen Sprachraum finden sich austauschbar die Bezeichnungen „Korrelationsverhältnis“ oder „Korrelationsindex“ (vgl. etwa Lienert & Raatz, 1971) oder auch „Korrelationsquotient“ und „Eta-Quotient“ (Kerlinger, 1978, 336; 1979, 1042).

η^2 gibt den relativen Anteil der Gesamtvariation in der AV an, der auf die Variation der UV zurückgeführt werden kann, also „systematisch“ ist; es beschreibt jegliche funktionale Beziehung zwischen einer beliebigen UV und einer kontinuierlichen AV - weitere Interpretationen dieses Maßes finden sich etwa bei Cohen (1965, 104f.). Liegt eine quantitativ gestufte (also metrische) UV vor,³¹⁾ ist es sinnvoll, η^2 als ein Maß aufzufassen, das - im Gegensatz zum bekannteren Produkt-Moment-Korrelationskoeffizienten r - den linearen und alle kurvilinearen Zusammenhänge zwischen der UV und der AV zu erfassen erlaubt; zu weiteren Einzelheiten siehe Cohen (1965, 104f.) und Hays (1977, 682-684).

³¹⁾ Als Beispiel denke man an die Geldbeträge, die ein die Theorie der kognitiven Dissonanz prüfender E aussetzt, um seine Vpn zu einer ihrer Meinung widersprechenden Aussage zu veranlassen (Festinger & Carlsmith, 1959), oder aber an Darbietungszeiten in einem Lernexperiment (Bredenkamp & Hager, 1979); vgl. auch Abschnitt 1.1.

Bekannter als η^2 ist in den letzten Jahren das von Hays (1963, 325, 382; 1977, 414) vorgeschlagene ω^2 geworden:

$$(9.8) \quad \omega^2 = \frac{\sigma_t^2 - \sigma_e^2}{\sigma_t^2}$$

Verläßt man das Modell der Varianzanalyse mit fixierten Effekten und geht auf das der multiplen Regression über, läßt sich den einschlägigen Lehrbüchern (etwa Kerlinger & Pedhazur, 1973; Cohen & Cohen, 1975) entnehmen, daß noch ein weiteres Maß verwendet werden kann, um die statistische Assoziation in der Population anzugeben, nämlich $R_{Y.X}^2$, das multiple Korrelationsquadrat. Kodiert man die UV bzw. die Gruppenzugehörigkeit mittels sog. Dummy-Variablen und setzt man voraus, daß den einzelnen Bedingungen formal „Zellenwahrscheinlichkeiten“ zugewiesen werden können (Keren & Lewis, 1979), gilt in der Population die numerische Äquivalenz von $R_{Y.X}^2$, ω^2 und η^2 (Cohen & Cohen, 1975, 187f.; Cohen, 1977).

Um auch in mehrfaktoriellen Plänen die experimentellen Effekte zu Lasten der einzelnen Faktoren bestimmen zu können, definiert man das Maß für den normierten EEP zweckmäßigerweise als quadrierte Partialkorrelation. Hierdurch wird der relative Anteil der Effektvarianz zu Lasten bspw. des Faktors A, also σ_A^2 , an der totalen Varianz σ_t^2 erfaßt, aus der sämtliche Variationsquellen zu Lasten der übrigen Faktoren und der Interaktion(en) auspartialisiert werden - vgl. im einzelnen Kennedy (1970), Cohen (1973a, 1977), Humphreys & Fleishman (1974), Cohen & Cohen (1975) sowie Keren & Lewis (1979). Somit ergibt sich bspw.:

$$(9.9) \quad R_{YA.B,AB,\dots}^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}$$

Wie sich den Definitionen entnehmen und durch algebraische Umformungen zeigen läßt, gelten folgende Beziehungen zwischen den hier angesprochenen Korrelationskoeffizienten und den Nicht-Zentralitätsparametern im vorigen Abschnitt (vgl. auch Cohen, 1977, 410-415, und Hager, im Druck, b):

$$(9.10) \quad f^2 = \frac{R_{Y.X}^2}{1 - R_{Y.X}^2}$$

$$(9.11) \quad R_{Y.X}^2 = \frac{f^2}{1 + f^2}$$

Plant man ein multivariates Experiment zur Prüfung von Hypothesen über Mittelwertsvektoren, müssen die soeben vorgestellten univariaten Maße für den EE durch ihre multivariaten Äquivalente ersetzt werden, die nach den gleichen Prinzipien definiert sind wie die univariaten Maße; nähere Einzelheiten entnehme man den Arbeiten von Cooley & Lohnes (1971, 227), Huberty

(1972), Smith (1972), Stevens (1972a, 1980), Sachdeva (1973), Olson (1974), Shaffer & Gillo (1974), Bredenkamp (1975, 1980), Odeh & Fox (1975, 36-42), Läuter (1978) sowie Cramer & Nicewander (1979).

9.3.2.2 Stichprobenmaße: $\hat{R}_{Y.X}^2$, E^2 , UI

Ersetzt man in der Ausgangsdefinition (9.6) die Varianzterme durch die Stichprobenvarianzen S_t^2 und S_e^2 , erhält man normierte Maße für den EES, die als Stichprobenäquivalente der entsprechenden Populationsmaße aufgefaßt werden können. Auch im Stichprobenfall wird dieser Quotient mit zahlreichen unterschiedlichen Symbolen belegt; Wishart (1932) und Kerlinger (1964) bezeichnen ihn mit E^2 , andere Autoren, etwa Cohen (1965, 105), mit η^2 . Bolles & Messick (1958) und Gaito (1958) haben den Namen „Utilitätsindex“ UI oder U_b vorgeschlagen; dieser Vorschlag wurde neuerdings von Gaebelin & Soderquist (1978) und Soderquist & Hussian (1978) wieder aufgegriffen.

Aus der folgenden Definitionsgleichung (9.13) wird jedoch deutlich, daß die unterschiedlichen Symbole allesamt die quadrierte multiple Korrelation in der Stichprobe, $\hat{R}_{Y.X}^2$, bezeichnen - vgl. Kerlinger & Pedhazur (1973) sowie Cohen & Cohen (1975):

$$(9.12) \quad \hat{R}_{Y.X}^2 = \frac{S_t^2 - S_e^2}{S_t^2} = \frac{S_b^2}{S_t^2} = 1 - \frac{S_e^2}{S_t^2} = \frac{SST - SSE}{SST}$$

(„SST“ bezeichnet die varianzanalytische totale Quadratsumme, „SSE“ die Fehlerquadratsumme.)

Werden nur $K = 2$ Mittelwerte untersucht, ist der Ausdruck (9.12) mit dem quadrierten Punkt-biserialen Korrelationskoeffizienten r_{pbis}^2 identisch (Cohen, 1968; Bredenkamp, 1970).

Darüber hinaus ist $\hat{R}_{Y.X}^2$ nichts anderes als eine monotone Transformation der empirischen Werte der üblichen Teststatistiken F und t , wie bereits Fisher (1928, 672) und Wishart (1932, 441) gezeigt haben; und es gilt (F bezeichne den empirischen Wert der Teststatistik F):

$$(9.13) \quad \hat{R}_{Y.X}^2 = \frac{(K - 1) F}{(K - 1) F + (N - K)}$$

Weitere Maße für den EES haben Levy (1967; vgl. auch Cohen, 1969, 19-21) und Friedman (1968, 1969; zur Kritik Breen & Gaito, 1970) vorgeschlagen, die mit den hier angesprochenen in sehr engem Zusammenhang stehen und deshalb nicht gesondert erörtert werden - vgl. zu Einzelheiten Bredenkamp (1970) sowie Gaito & Firth (1973).

9.3.2.3 Korrekturformeln für $\hat{R}_{Y.X}^2$

Es ist bekannt, daß die Größe der multiplen Korrelation in der Stichprobe, $\hat{R}_{Y.X}$, unter sonst gleichen Bedingungen um so größer ist, je mehr Modalitäten der UV X - also „Prädiktoren“ - zur Vorhersage des „Kriteriums“ - also der Werte der AV - verwendet werden. Selbst wenn die tatsächliche Korrelation zwischen den Prädiktoren und dem Kriterium gleich Null ist, ist im Mittel über alle möglichen Stichproben ein Korrelationsquadrat der Größe

$$(9.14) \quad \hat{R}_{Y.X}^2 = \frac{K - 1}{N - 1}$$

zu erwarten (vgl. Wishart, 1932, 444; Cohen & Cohen, 1975, 107). Zur Reduktion dieser systematischen Verzerrung („Bias“) werden in der Literatur unterschiedliche sog. „(Schrumpfungs-)Korrekturen“ vorgeschlagen, über die zusammenfassend Särndal (1974), Carter (1979) und Huberty & Mourad (1980) informieren.

Von diesen Korrekturformeln greifen wir zwei heraus, die häufiger im Zusammenhang mit Varianz- und Regressionsanalysen benutzt werden.

$$(9.15) \quad \begin{aligned} \hat{R}_{\text{kor.}}^2 = \epsilon^2 &= 1 - (1 - \hat{R}^2) \left(\frac{N - 1}{N - K} \right) = \frac{(K - 1)(F - 1)}{(K - 1)F + (N - K)} = \\ &= \frac{s_t^2 - s_e^2}{s_t^2} \end{aligned}$$

Hierbei gilt: $s_t^2 = [(SST) : (N-1)]$ und $s_e^2 = [(SSE) : (N-K)]$ - siehe zu dieser Formel u.a. Cohen (1965, 105) und Fleiss (1969). Letzterer gibt darüber hinaus eine weitere Korrekturformel an:

$$(9.16) \quad \begin{aligned} \hat{R}_{\text{kor.}}^2 = \hat{\omega}^2 = \hat{\eta}^2 &= 1 - (1 - \hat{R}^2) \left(\frac{N}{N - K + 1 - \hat{R}^2} \right) = \\ &= \frac{(K - 1)(F - 1)}{(K - 1)(F - 1) + N} = \frac{s_t^2 - s_e^2}{s_t^2 + \frac{s_e^2}{N - 1}} \end{aligned}$$

Bredenkamp (1970) hat im einzelnen zeigen können, daß diese unterschiedlichen Korrekturformeln entstehen, indem jeweils verschiedene Schätzwerte für die Populationsvarianzen σ_t^2 , σ_b^2 und σ_e^2 in die Gleichungen (9.5) und (9.6) eingesetzt werden.

Die Kriterien und Methoden zur Ermittlung dieser Schätzwerte, die Wang (1967) und Miller (1979) miteinander verglichen haben, finden sich in einschlägigen Lehrbüchern wie etwa denen von Winer (1962, 1971), Glass & Stanley (1970), Edwards (1971, 1980), Lee (1975), Bortz (1979) sowie Henning & Muthig (1979) und darüber hinaus in den

Arbeiten von Gaito (1960a), Halderson & Glasnapp (1972), Dwyer (1974), Gaebelein & Soderquist (1976) und Hopkins (1976); Schätzwerte für in einigen sog. „Standard-Versuchsplänen“ zu erwartende Varianzen haben zudem Gaito & Turner (1963), Vaughan & Corballis (1969), Fleiss (1969) sowie Dodd & Schultz (1973) veröffentlicht.

Die o. gen. Korrekturen, die als Schätzungen der Populationskorrelation $R_{Y,X}^2$ interpretiert werden können, genügen allesamt nicht dem Kriterium der „Erwartungstreue“ (vgl. hierzu etwa Stilson, 1966, 205f.; Hays, 1977, 272-274; Bortz, 1979, 123f.), obwohl dies verschiedentlich behauptet worden ist. Schätzungen, die diesem Kriterium genügen, liegen bislang nicht vor - vgl. dazu Olkin & Pratt (1958), Cureton (1966), Bredenkamp (1970) und Huberty & Mourad (1980).

Zwar meinen wir, daß grundsätzlich korrigierte Werte für $\hat{R}_{Y,X}^2$ den unkorrigierten vorzuziehen sind, aber andererseits kann es dem E im Einzelfall überlassen bleiben, für welche der möglichen Korrekturen er sich entscheidet, weil die Unterschiede insgesamt Vernachlässigbar gering sind und bei größeren Stichprobenumfängen ohnehin verschwinden; vgl. hierzu die Computer-Simulationen („Monte-Carlo-Studien“)³²⁾ von Särndal (1974), Carroll & Nordholm (1975), Keselman (1975), Carter (1979) sowie von Huberty & Mourad (1980).

Diesen Untersuchungen ist jedoch auch zu entnehmen, daß die Standardfehler der Korrelationsstatistiken meist unverhältnismäßig hoch sind. Da die theoretische Ableitung der entsprechenden Stichprobenverteilungen bislang noch nicht gelungen ist, kann im Einzelfall nicht angegeben werden, wie stark die empirischen Werte für $\hat{R}_{Y,X}^2$ oder $\hat{R}_{korr.}^2$ streuen können (zur Bestimmung approximativer Konfidenzintervalle siehe Venables, 1975, und Fleishman, 1980).

Allerdings sind teilweise für Nicht-Zentralitätsparameter wie etwa f^2 (vgl. Abschnitt 9.3.1) die Standardfehler bestimmt worden. Man kann also den Wert für das korrigierte multiple Korrelationsquadrat in $\hat{f}^2$ umrechnen und für dieses den Standardfehler errechnen - vgl. zu Einzelheiten Fleishman (1980).

Zur direkten Bestimmung des Ausmaßes der Nicht-Zentralität in der Stichprobe als monotone Funktion des multiplen Korrelationsquadrats oder des empirischen Wertes der Teststatistik F (resp. t) können die Gleichungen (9.10) und (9.15) oder (9.16) herangezogen werden.

Darüber hinaus kann der Nachteil des (für Korrelationskoeffizienten typischen) großen Standardfehlers bei der Beurteilung der WH teilweise dadurch ausgeglichen werden, daß die EES der entsprechenden Experimente zusammenfassend miteinander verglichen werden - siehe Abschnitt 11.2.

³²⁾ Zur Einführung in diese Techniken orientiere man sich u.a. bei Hammersley & Handscomb (1964), Collier & Larson (1969), Kleijnen (1974, 1975) sowie Bauknecht, Kohlas & Zehnder (1976).

9.4 Experimentelle Effekte bei nicht-parametrischen Hypothesen

Auch bei ordinalen und nominalen Daten sind die EE in Abhängigkeit von den jeweils geprüften Hypothesen zu definieren; dies führt zu einer Fülle von unterschiedlichen Maßen für den normierten EE. Wir wollen im folgenden nur einige gebräuchliche Konstruktionsprinzipien und Maße aufgreifen.

9.4.1 Experimentelle Effekte bei ordinalen Daten

Das nicht-parametrische Äquivalent der (auf Intervallniveau beruhenden) parametrischen Korrelationskoeffizienten stellen die Rangkorrelationskoeffizienten dar.

Liegen bspw. in einem Zwei-Gruppen-Plan abhängige rangskalierte Daten vor, kann der Zusammenhang zwischen beiden Rangreihen über die rechnerische Vereinfachung des Produkt-Moment-Korrelationskoeffizienten r bestimmt werden, die in der Literatur üblicherweise „Spearman's Rangkorrelationskoeffizient r_s “ genannt wird (vgl. etwa Bortz, 1979, 283-286; ferner Hays, 1977, 788-792).

Für einen Versuchsplan mit K Gruppen und unabhängigen Messungen haben Lienert & Raatz (1971) das nicht-parametrische Äquivalent des parametrischen multiplen Korrelationsquadrates abgeleitet, das sie „Rangkorrelationsverhältnis η_H^2 “ nennen; wir wollen im folgenden die Bezeichnung E_H^2 benutzen, um anzudeuten, daß es sich um ein Maß für den EE in der Stichprobe handelt. Dies folgt aus der Definition von E_H^2 als Funktion des empirischen Wertes der Teststatistik H (vgl. dazu Kruskal & Wallis, 1952):

$$(9.17) \quad E_H^2 = \frac{H}{N - 1}$$

Der Quotient (9.17) gibt den relativen Anteil an der gesamten Rangvarianz an, der durch die Treatments „aufgeklärt“ werden konnte, und entspricht konzeptuell dem parametrischen $\hat{R}_{Y.X}^2$ (vgl. Lienert & Raatz, 1971; Bredenkamp, 1980, 62).

Die bisher vorgestellten Normierungen des in absoluten Einheiten angegebenen EE erfolgten entweder auf die Fehler- oder aber auf die totale Varianz der AV, und es resultierten Maße der Nichtzentralität und der erklärten Varianz (quadrierte Korrelationen). Eine weitere Normierungsmöglichkeit besteht darin, den empirischen Wert einer Teststatistik auf den jeweils theoretisch maximal möglichen Wert zu relativieren. Das Resultat kann man als „Kontingenz-Quotient“ bezeichnen, da der in Kontingenztafeln bestimmbare EE i.a. nach dieser Methode normiert wird (vgl. Abschnitt 9.4.2).

$$(9.18) \quad \begin{aligned} &\text{EE als Kontingenzquotient: } = \\ &= \frac{\text{ermittelter Wert der Teststatistik}}{\text{maximal möglicher Wert der Teststatistik}} \end{aligned}$$

Acock & Stavig (1979; vgl. auch Stavig & Acock, 1980) haben auf dieser Basis für etliche nicht-parametrische Testverfahren Maße für den normierten EE abgeleitet. Als Beispiel ergibt sich für die Rangvarianzanalyse nach Friedman (1937) über abhängige Rangdaten ein Quotient, der dem Konkordanzkoeffizienten W von Kendall (1970) entspricht.

Die als Kontingenzquotienten definierten Maße für den normierten EE sind allesamt mit den quadrierten Korrelationskoeffizienten (s. o.) nicht vergleichbar, obwohl sie sich ebenfalls innerhalb der Grenzen 0 und 1 bewegen (Lienert & Raatz, 1971).

Weitere Maße des statistischen Zusammenhangs zwischen ordinalen Variablen können der im Abschnitt 7.5.3.1 angeführten Literatur entnommen werden; siehe ferner Särndal (1974) und Glass (1978).

9.4.2 Experimentelle Effekte bei nominalen Daten

Das Ausmaß der statistischen Assoziation in einer Vier-Felder-Kontingenztafel wird häufig mittels des „Punkt-Vier-Felder-Korrelationskoeffizienten“, auch kurz „Phi-Koeffizient“ genannt, erfaßt, der ebenfalls eine rechnerische Vereinfachung des Korrelationskoeffizienten r darstellt.

Wenn wir vom empirischen Wert der Teststatistik χ^2 ausgehen, ist der quadrierte Phi-Koeffizient wie folgt definiert (Hays, 1977, 743):

$$(9.19) \quad V_p^2 = \frac{\chi^2}{N}$$

Bei Vorliegen von mehr als 2×2 Kategorien, also einer mehrdimensionalen $A \times B$ -Kontingenztafel, wird eine Verallgemeinerung des Phi-Koeffizienten definiert, die als „Cramérs V “ (oder „ C “) bekannt ist (vgl. Hays, 1977, 745):

$$(9.20) \quad V^2 = \frac{\chi^2}{N (K_{\min} - 1)}$$

($K_{\min}$ bezeichnet die jeweils geringere Anzahl von Abstufungen oder Modalitäten für die beiden Kategorien A und B .)

Auch der Koeffizient V^2 ist im Rahmen des unter (9.18) angegebenen Kontingenzmodells interpretierbar, nicht jedoch als Korrelationskoeffizient (s. o.).

Einen weiteren - wohl den gebräuchlichsten - Ansatz zur Erfassung der Assoziation in Kontingenztafeln stellt der sog. „Kontingenzkoeffizient“ C dar, der wie folgt definiert ist (vgl. Hays, 1977, 745):

$$(9.21) \quad C^2 = \frac{\chi^2}{\chi^2 + N} = \frac{V^2}{1 + V^2}$$

Wie leicht zu erkennen ist, liegt die obere Grenze von C^2 unter Eins, weswegen die Nützlichkeit dieses Maßes im Vergleich zu den anderen häufig angezweifelt wird - vgl. etwa Hays (1977, 745).

Weitere Indices für das Ausmaß an statistischer Assoziation bei nominalen Daten finden sich bei Fleiss (1973), Särndal (1974), Cohen (1977), Glass (1978), Acock & Stavig (1979), Langeheine (1980, 73) sowie Stavig & Acock (1980), ferner in der im Abschnitt 7.5.3.1 genannten Literatur.

9.5 Zur Kritik der Maße der statistischen Assoziation

Gegen die in den vorangegangenen Abschnitten dargestellten oder erwähnten normierten Maße der statistischen Assoziation, insbesondere gegen die Korrelationskoeffizienten, sind verschiedentlich Einwände geltend gemacht worden - vgl. dazu insbesondere Glass & Hakstian (1969), Kennedy (1970) sowie Dooling & Danks (1975). Diesen Einwänden sind insbesondere Halderson & Glasnapp (1972), Soderquist & Hussian (1978) sowie Fleishman (1980) entgegengetreten; wir wollen die Argumente und Gegenargumente an dieser Stelle nicht aufgreifen, zumal aus den bisherigen Ausführungen hinreichend deutlich geworden sein sollte, daß das Konzept der experimentellen Effekte - trotz aller möglichen Schwächen - unverzichtbar ist. Drei dieser Schwächen wollen wir im folgenden kurz ansprechen, weil ihre Kenntnis für das Arbeiten mit den EE wesentlich ist.

- (1) Ein gewichtiger Einwand gegen die EE besteht in der Tatsache, daß sie als normierte Größen definiert werden, wobei hierbei entweder Fehler- oder totale Varianzen verwendet werden, deren Größe interexperimentell sehr starken Schwankungen unterworfen sind. Dies bedeutet, daß ein kleiner Effekt als Quotient aus einer großen quadrierten Mittelwertsdifferenz plus einer großen Fehlervarianz entstehen kann oder in einem präziseren Experiment als Quotient aus einem kleinen absoluten Effekt plus einer geringen Fehlervarianz.

Sofern es die *Festlegung* des EEM vor dem Experiment anbelangt, muß diese unter Verwendung eines normierten EE erfolgen, weil die zur Bestimmung des benötigten Stichprobenumfangs (siehe Abschnitt 10) verfügbaren Tabellen der Gütefunktionen aus praktischen Gründen stets un-

ter Verwendung normierter Maße erstellt werden. Dies geschieht analog der Tabellierung der kritischen Werte der Teststatistiken aus den zentralen Verteilungen - man bezieht sich bspw. immer auf die „Standardnormal-Verteilung“ mit $NV(\mu=0; \sigma_e^2=1)$. Zudem ist in aller Regel vor dem Experiment nicht bekannt, wie groß σ_e^2 sein wird - ohne diese Information können die Tabellen zur Stichprobenumfangsbestimmung nur unter Verwendung normierter EEM benutzt werden - siehe Abschnitt 10.

Darüber hinaus wird sich die Festlegung von absoluten Abweichungen von der Erwartung unter H_0 bei der Planung eines Experiments meist als sehr schwierig erweisen, weil derart genaue Prognosen aus kaum einer wissenschaftlichen Kausalhypothese abgeleitet werden können. Sind dagegen die empirischen Daten erhoben worden, sollte der EES sowohl als Funktion des empirischen Wertes der benutzten Teststatistik wie als absolute Mittelwertsdifferenz oder Treatmentvarianz oder dgl. bestimmt und mitgeteilt werden - dies ist ohne großen Aufwand zu bewerkstelligen, wie die Abschnitte 9.3.2.2 und 9.3.2.3 zeigen.

- (2) Das Ausmaß des EE ist unter sonst gleichen Bedingungen sehr stark abhängig von der Anzahl der untersuchten Treatmentmodalitäten und dem gewählten Versuchsplan. Selbst normierte EE sind interexperimentell nicht direkt miteinander vergleichbar, und festgestellte EES beziehen sich nur auf das durchgeführte Experiment mit seinem speziellen Versuchsdesign und seinen spezifischen Faktormodalitäten. Dennoch ermöglicht eine vergleichende Betrachtung der unter verschiedenen Versuchsbedingungen ermittelten EES eine differenzierte Bewertung der wissenschaftlichen Kausalhypothese, weil etwa festgestellt werden kann, welche Operationalisierungen oder UVn zu relativ kleinen oder aber großen Effekten führen - vgl. Bredenkamp (1980, 52).
- (3) Auf den Einwand, daß die experimentellen Effekte EES - besonders bei kleinen Stichprobenumfängen - mit einem großen Standardfehler behaftet sind, sind wir bereits im Abschnitt 9.3.2.3 eingegangen.

9.6 Zusammenfassung

Zusammenfassend und diesen Teil abschließend sei nochmals hervorgehoben, daß das Konzept des experimentellen Effekts - trotz seiner bestehenden Schwächen - als integrativer Bestandteil eines jeden Signifikanztests unverzichtbar ist. Hierfür gibt es vor allem zwei Gründe:

- (1) Die Tatsache der statistischen Signifikanz besagt lediglich, daß der gefundene EE nicht als zufällig zustande gekommen interpretiert werden sollte, nicht jedoch, wie groß er ist. Diese Information ist aber für die folgende Beurteilung der geprüften Kausalhypothese WH wesentlich (vgl. Abschnitt 7.4 und 11.2).

- (2) Bei jedem Signifikanztest kann die Wahrscheinlichkeit β für einen Fehler 2. Art (und damit auch die Teststärke $1-\beta$) nur kontrolliert und/oder bestimmt werden, wenn angegeben werden kann, wie sehr die empirischen Resultate von ihrer Erwartung unter Gültigkeit der H_0 abweichen. Diese Information ist aber von großer Bedeutung, weil bei einem Fehler 2. Art irrtümliche Falsifikationen der WH zu erwarten sind, wenn wir von der Implikationsbeziehung $WH \rightarrow H_1$ ausgehen.

Zu diesen Punkten haben wir bereits im Zusammenhang mit den Determinanten des Signifikanztests im Abschnitt 7.4.3 ausgeführt, daß der EE als interessierender („praktisch bedeutsamer“) Mindesteffekt EEM zweckmäßigerweise vor dem Experiment vom E nach inhaltlichen Kriterien zu fixieren ist; auf einige Probleme hierbei gehen wir im Abschnitt 11.1 ein. Die Fehlerwahrscheinlichkeit α und β sind ebenfalls vom E zu kontrollieren und gering zu halten. Trifft der E diese Festlegungen, sollte er den Stichprobenumfang so spezifizieren, daß die Hypothesenprüfung unter Einhaltung der vorgegebenen Kriterien durchgeführt werden kann. Dies gilt natürlich unter der Voraussetzung, daß der tatsächliche EE mindestens so groß ist wie der EEM; andernfalls vermindert sich die Teststärke.

Wie die Stichprobenumfangsbestimmung unter diesen Gesichtspunkten im einzelnen erfolgen kann, werden wir im Teil 10 darstellen.

10. Bestimmung des Stichprobenumfanges

10.1 Überblick

Unter „Stichprobenumfang“ verstehen wir die Anzahl der Beobachtungs- oder Untersuchungseinheiten (V_{pn} ; vgl. Abschnitt 3.3.2) pro experimenteller Bedingung (Symbol: n) oder pro Experiment (Symbol: N).

Mit der Bestimmung von Stichprobenumfängen können grundsätzlich drei verschiedene Ziele verfolgt werden (vgl. Mace, 1964; Enderlein, 1972; Rasch, Herrendörfer & Bock, 1974; Rasch et al., 1978):

1. die Prüfung von statistischen Hypothesen;
2. die (Intervall-)Schätzung für Parameter; und
3. die Lösung von Selektionsproblemen.

Der Bearbeitung von Selektionsproblemen sieht sich der experimentell arbeitende Psychologe u.W. so gut wie nie konfrontiert, weswegen wir diesen Bereich der eher praktischen Arbeit nicht behandeln - eine ausführliche Dar-

stellung findet sich etwa bei Mace (1964, insbes. Kap. 7). Die Prinzipien, die bei der Bestimmung von Konfidenzintervallen für Parameter zu beachten sind, lassen sich relativ einfach aus denjenigen ableiten, die wir im folgenden im Zusammenhang mit der Stichprobengrößenbestimmung zur Prüfung statistischer Hypothesen darlegen werden; wir gehen daher nicht weiter auf die unter 2. genannte Fragestellung ein, zu der Mace (1964), Enderlein (1972) sowie Rasch et al. (1978) detaillierte Informationen geben.

Bei der Bestimmung des „optimalen“ Stichprobenumfangs zur Prüfung von statistischen Hypothesen ist zunächst die Frage zu beantworten, hinsichtlich welchen Kriteriums „Optimalität“ erzielt werden soll. Aus dem sog. „Zentralen Grenzwertsatz“ und dem „Gesetz der großen Zahlen“ (vgl. dazu u.a. Haagen & Pertler, 1976, 158-164; Hays, 1977, 278-283; Pfanzagl, 1978, 61-66) sowie aus unseren Ausführungen zur Präzision einer Untersuchung im Abschnitt 8.4 ergibt sich folgendes Optimalitätskriterium: „Wähle eine möglichst (sehr) große Stichprobe, etwa $n > 1000!$ “

Stehen dagegen Kostenüberlegungen im Vordergrund, ergibt sich als entgegengesetzte Forderung: „Wähle eine möglichst (sehr) kleine Stichprobe, etwa $n \leq 5!$ “

Selbstverständlich sind zahlreiche weitere Optimalitätskriterien denkbar, auf die wir jedoch nicht weiter eingehen wollen; man orientiere sich etwa bei Mace (1964) und Rasch, Herrendörfer & Bock (1974).

Aufgrund unserer Erörterungen in den Teilen 6, 7 und 8 ergibt sich die Vorrangigkeit eines anderen Kriteriums: „Wähle die Größe der Stichprobe derart, daß bei vorgegebenem experimentellem Mindesteffekt die Prüfung der statistischen Hypothese bei festgelegten geringen Maximalwerten für die Wahrscheinlichkeiten von Fehlern 1. und 2. Art erfolgen kann!“ Die Erfüllung dieses Kriteriums stellt eine notwendige Bedingung für die strenge Prüfung einer wissenschaftlichen Hypothese dar.

Es wird sich jedoch zeigen, daß sich bei seiner konsequenten Anwendung sehr häufig Stichproben in einer Größenordnung ergeben, die in der Regel nicht zu bewältigen sind - vgl. dazu das erste Beispiel im Abschnitt 10.3.3.2. In diesen Fällen erzwingen allgemeine Kostenüberlegungen i. w. S. (vgl. Pfanzagl, 1978, 106) die Festsetzung einer oberen Grenze für den realisierbaren Stichprobenumfang. Hierdurch ergibt sich oft die Notwendigkeit, die Größe mindestens einer der vier Determinanten des Signifikanztests, die vor der Datenerhebung fixiert wurden, zu modifizieren, und es resultiert möglicherweise eine weniger strenge Prüfung der WH. Ansätze zur Lösung dieses Problems werden teilweise im Abschnitt 10.3.3.2 skizziert und im Abschnitt 11.1.1 nochmals aufgegriffen.

über die unterschiedlichen Möglichkeiten, das Modell der Kosten-Nutzen-Analyse systematisch bei der Bestimmung des Stichprobenumfanges einzusetzen, informieren eingehender die Arbeiten von Cox (1958), Mace (1964, Kap. 11), Overall & Dalal (1965), Cleary & Linn (1969), Dixon & Massey (1969, 89f.), Cochran (1972), Abrahams & Alf (1978) sowie die letzten Jahrgänge des „Biometrical Journal“.

Auf der anderen Seite kann es zuweilen geschehen, daß Stichprobenumfänge in der Größenordnung von $n = 2$ bis $n = 4$ berechnet werden. Derartige Zahlen ergeben sich etwa dann, wenn sehr große experimentelle Effekte bei vergleichsweise großen Irrtumswahrscheinlichkeiten aufgedeckt werden sollen. Ist bspw. die zu entdeckende Effektvarianz σ_b^2 genauso groß wie die Fehlervarianz σ_e^2 , benötigt man für einen mittels t-Test auszuwertenden Zwei-Gruppen-Plan $n = 4$ Vpn pro Gruppe, wenn die Hypothesenprüfung bei $\alpha = 0,20$ und $\beta = 0,20$ zweiseitig erfolgen soll; zur Überprüfung einer gerichteten Hypothese reichen sogar $n = 3$ Vpn aus. In solchen Fällen muß sich der E die Frage stellen, welche Aussagekraft einem Experiment mit derartig geringen Gruppengrößen angesichts unserer Ausführungen über die Populationsvalidität im Abschnitt 4.1 zukommen kann.

Wir werden in den folgenden Abschnitten die wesentlichsten der bislang entwickelten Ansätze der Stichprobenumfangsbestimmung kurz umreißen, wobei wir das Prinzip ausführlich am Beispiel des t-Tests und des varianzanalytischen F-Tests demonstrieren, während wir für eine Reihe weiterer Testverfahren vornehmlich Hinweise auf die relevante Literatur geben.

Nicht eingehen werden wir dabei auf den nur in recht eingeschränkten Zusammenhängen verwendbaren Ansatz, die Stichprobengrößenbestimmung unter Einbeziehung der Reliabilität der AV vorzunehmen (siehe zu diesem Terminus Abschnitt 2.5 und die dort angegebene Literatur). Die vorgeschlagenen Strategien (vgl. Levin & Subkoviak, 1977, 1978; Subkoviak & Levin, 1977; zur Kritik Forsyth, 1978 a, b) gehen von der Annahme aus, es ließe sich eine eindeutige Beziehung zwischen der Reliabilität eines als AV verwendeten Tests und der Teststärke angeben (vgl. dazu Sutcliffe, 1958; Cleary & Linn, 1969; Cleary, Linn & Walster, 1970; Leeb & Weinberg, 1977). Wie die Befunde von Overall & Woodward (1975, 1976) und Fleiss (1976) sowie die theoretischen Erörterungen von Nicewander & Price (1978) und Bredenkamp (1980) aufzeigen, ist diese Annahme nicht generell gültig (vgl. jedoch Sutcliffe, 1980).

10.2 Allgemeine Prinzipien der Stichprobengrößenbestimmung

Zur praktischen Durchführung der Stichprobengrößenbestimmung muß der E vornehmlich Überlegungen zu den folgenden Punkten anstellen:

- (1) Adäquate Umsetzung der wissenschaftlichen Hypothese(n) in statistische
 - siehe Abschnitt 8.1;

- (2) Wahl des ihm unter Berücksichtigung der Teile 1 bis 8 dieser Arbeit geeignet erscheinenden Versuchsplanes mit der Festlegung von Art und Anzahl der Treatmentmodalitäten usw. ;
- (3) Wahl einer Teststatistik - vgl. hierzu die Teile 7 und 8;
- (4) Fixierung der maximalen Wahrscheinlichkeit für einen Fehler 1. Art (Signifikanzniveau α) - vgl. Abschnitt 11.12;
- (5) Fixierung der maximalen Wahrscheinlichkeit für einen Fehler 2. Art (oder der Teststärke $1 - \beta$) - vgl. ebenfalls Abschnitt 11.1.2;
- (6) Angabe der Größe des experimentellen Effektes EEM, der mindestens auftreten muß, damit das Resultat als „praktisch bedeutsam“ interpretiert werden kann - siehe Teil 9 und Abschnitt 11.1.3.

Neben diesen Überlegungen und anschließenden Festlegungen benötigt der E die Werte der nicht-zentralen Verteilungen der von ihm gewählten Teststatistik, ohne die die Stichprobenumfangsbestimmung nicht durchführbar ist. Aus diesen nicht-zentralen Verteilungen läßt sich die Gütefunktion des betr. Tests ableiten, die die Abhängigkeitsbeziehungen zwischen den Determinanten dieses Tests verdeutlicht. Diese Abhängigkeitsbeziehungen wiederum lassen sich graphisch oder tabellarisch darstellen (vgl. dazu die o. g. Abschnitte). Betrachten wir beispielhaft nur die Gütefunktion des F-Tests, so sind hierbei im wesentlichen drei verschiedene Typen von Tabellen resp. Graphen anzutreffen.

1. Bei vorgegebenem Signifikanzniveau α werden die Werte für die Teststärke $1 - \beta$ in Abhängigkeit von der Größe des experimentellen Effektes (meist als λ oder φ^2 angegeben; siehe dazu die folgende Gleichung (10.1)), der Anzahl der Treatment-Modalitäten (i. a. über die Freiheitsgrade df_Z oder u für den Zähler des F-Bruches aufgelistet) und der Stichprobengröße N (in der Regel über die Freiheitsgrade df_N oder v für den Nenner des F-Bruches angegeben) verzeichnet;
2. es wird die Mindestgröße des experimentellen Effektes aufgeführt, die bei vorgegebenen Werten für die übrigen Determinanten aufgedeckt werden kann; oder
3. es wird der benötigte Mindeststichprobenumfang in Abhängigkeit von den anderen Größen angegeben.

Infolge ihres speziellen Aufbaus eignen sich die Tabellen des unter 3. aufgeführten Typs vornehmlich zur Planung von Untersuchungen, während sich die beiden unter 1. und 2. charakterisierten Typen eher bei der Auswertung von Untersuchungen sinnvoll einsetzen lassen. Hat man ein Experiment aus irgendwelchen Gründen nicht nach dem in dieser Arbeit favorisierten Strategie planen können oder aber ist eine Publikation von experimentellen Ergebnissen so unzulänglich, daß sie keine Angaben über die Determinanten der durchgeführten Signifikanztests enthält, müssen diese Angaben aus den vorhandenen Informationen nachträglich ermittelt werden. Ein zumindest ex post facto

bestimmter Wert für den EES und für die Teststärke dient der adäquaten Interpretation der experimentellen Ergebnisse - siehe dazu insbesondere Abschnitt 11.2.

Wenden wir uns nun dem Umgang mit den genannten Tabellen im einzelnen zu!

10.3 Bestimmung des Stichprobenumfanges bei univariaten Varianz- und Regressionsanalysen

Wegen des überwiegenden Gebrauchs, der in der Praxis von den parametrischen Teststatistiken t , χ^2 und F gemacht wird, sind deren nicht-zentrale Verteilungen am ausführlichsten berechnet, tabelliert und in gekürzter tabellarischer oder graphischer Form publiziert worden, u.a. in Mosteller & Bush (1954), Scheffé (1959), Owen (1962), Pearson & Hartley (1962, 1972), Cochran & Cox (1968), Kirk (1968), Cohen (1969, 1977), Dixon & Massey (1969), Bailey (1971), Winer (1971), Enderlein (1972), Guenther (1973), Rasch, Enderlein & Herrendörfer (1973), Lee (1975), Odeh & Fox (1975) sowie Diehl (1979). Leicht zugänglich sind zudem die ausführlichen Tabellen, die Rotton & Schönemann (1978) veröffentlicht haben.

Um mit diesen Tabellen und Graphen direkt operieren zu können, muß der experimentelle Mindesteffekt EEM in Form eines der Nicht-zentralitätsparameter angegeben werden, die sich auf folgende Definition zurückführen lassen (vgl. auch Abschnitt 9.3.1 und 9.3.2.1):

$$(10.1) \quad \varphi^2 = \frac{\sum_{k=1}^K (\mu_k - \mu)^2 / K}{\sigma_e^2 / n} = \frac{\sigma_b^2}{\sigma_e^2 / n} = n \cdot f^2 = \frac{\lambda}{K}$$

Aus der vorstehenden Aussage folgt jedoch nicht, daß der EEM von vornherein notwendigerweise als Nicht-Zentralitätsparameter festzulegen ist. Die darzustellenden Ansätze der Stichprobenumfangsbestimmung unterscheiden sich im wesentlichen gerade in der Art, in der die erste Festlegung in Abhängigkeit von den verfügbaren Vorinformationen erfolgen kann. Ist bspw. σ_e^2 bekannt, muß der EEM nicht als normiertes Maß angegeben werden, wie Formel (10.1) zeigt.

10.3.1 Bei Kenntnis der Populationsvarianz σ_e^2 („Klassischer Ansatz“)

Bei diesem Bestimmungsverfahren wird davon ausgegangen, daß die EEM als Abweichung der einzelnen Treatment-Mittelwerte μ von ihrem Gesamtmittel-

wert μ einzeln oder zumindest als Gesamtquadratsumme gemäß der zu prüfenden WH festgelegt werden können. Ist dies geschehen, muß noch eine Angabe über die Größe der Varianz σ_e^2 erfolgen.

Wie diese zu erlangen ist, wird in der Literatur entweder nicht angesprochen („ . . . given a preliminary estimate of the standard deviation . . .“; Mace, 1964, 90; „Suppose further that we know that . . . (σ_e^2) is approximately . . .“; Dixon & Massey, 1969, 278) oder es wird empfohlen, σ_e^2 aus Daten zu schätzen, die etwa in Vorversuchen erhoben worden sind (Kirk, 1968, 9; Winne, 1968, 1613; Winer, 1971, 222); vgl. zu diesem Problem auch Feldt (1973).

Hat man sich jedoch auf irgendeine Weise eine Schätzung der Varianz σ_e^2 verschaffen können, wählt man die dem vorgegebenem Signifikanzniveau entsprechende Power Chart und setzt nach einer „Versuch-Irrtums-Strategie“ solange verschiedene Werte für n in Formel (10.1) ein, bis ein Wert für φ^2 resultiert, der zur gewünschten Teststärke führt - vgl. zur Vorgehensweise etwa Kirk (1968, 9-11) und Diehl (1979, 26-29). Dieses Verfahren dürfte deshalb in der Praxis nur selten relevant werden, weil erfahrungsgemäß in der Regel Informationen aus gesonderten Vorversuchen oder Pilot-Studien, die unter vergleichbaren Bedingungen durchgeführt wurden, wie dies für den Hauptversuch intendiert ist, nicht verfügbar sind. Beabsichtigt daher der Experimentator, eigene Voruntersuchungen anzustellen, muß er berücksichtigen, daß sich dadurch sein Gesamtversuchspersonenbedarf nicht unwesentlich steigert; er wird vergleichsweise rasch an die Grenzen seiner Möglichkeiten stoßen. Trotz dieser gewichtigen Nachteile wird das soeben vorgestellte Verfahren in der Literatur nicht selten propagiert - vgl. u.a. Kirk (1968, 9f.), Winne (1968), Dixon & Massey (1969, 270-279) sowie Diehl (1979, 25-29).

Das Problem des meist zu großen Stichprobenumfanges ließe sich nicht unwesentlich reduzieren, wenn die in den Vorversuchen erhobenen Daten unverändert auch in den Hauptversuchen Verwendung finden könnten. Ein Verfahren, das dies ermöglicht, wird im folgenden Abschnitt dargestellt.

10.3.2 Bei prä-experimenteller Schätzung der Varianz σ_e^2

(„Two-Stage-Sampling“-Verfahren nach Stein und Rodger)

Auch bei dem „Two-stage-sampling“-Verfahren, das auf Stein (1945) zurückgeht, ist die Populationsfehlervarianz σ_e^2 zunächst unbekannt. Um ihre Größe zu ermitteln, wird ein Vorversuch durchgeführt.

Vor diesem wird neben den übrigen obligatorischen Festlegungen der EEM als $K\sigma_b^2$ angegeben (vgl. dazu Formel (10.1) und Abschnitt 10.2.1); diesen Ausdruck nennt Rodger (1976) „V“.

Anschließend führt der E sein Experiment an K Stichproben der gleichen, aber willkürlichen und vorläufigen Größe n_0 durch und schätzt aus diesen Kn , Daten die Varianz σ_ϵ^2 ; diese Schätzung beruht auf den üblichen Freiheitsgraden der Varianzanalyse, nämlich $df_N = v = K(n_0 - 1)$. Als nächstes bestimmt der E die Stichprobengröße n_1 , für die gilt (vgl. Rodger, 1976, 3):

$$(10.2) \quad n_1 = \frac{\hat{\sigma}_\epsilon^2 \cdot D\beta}{V} + 1$$

Hierbei kennzeichnet $D\beta$ einen speziellen Nicht-Zentralitätsparameter, der von Stein (1945) entwickelt wurde und dessen Werte in Abhängigkeit von den Freiheitsgraden df_z und df_N erstmalig Rodger (1976) in tabellierter Form vorgelegt hat.

Ergibt die Berechnung von n_1 , daß $n_1 \leq n_0$, wird der übliche varianzanalytische F-Test zur Prüfung der generellen Null-Hypothese durchgeführt. Ist jedoch $n_1 > n_0$, sind weitere $n_1 - n_0$ Beobachtungen für jede der K experimentellen Bedingungen anzustellen.

Zur Bestimmung der Quadratsumme zwischen den experimentellen Bedingungen werden alle $K \cdot n_0$ Beobachtungen verwendet; die statistische Prüfung erfolgt allerdings gegen die eingangs bestimmte Varianz $\hat{\sigma}_\epsilon^2$. Der EES wird nach der Datenerhebung aus den Daten bestimmt.

Die Stichprobengröße n_1 hängt bei diesem Verfahren u.a. von der Wahl des Wertes für n_0 ab; Rodger (1967, 192) bemerkt, daß ein Nachteil des zweistufigen Stichprobenverfahrens darin liegt, daß der Wert für n_1 nach der ersten Stufe sich als „unpraktikabel groß“ herausstellen kann. Diese Schwierigkeit kann zuweilen durch die Wahl eines „optimalen“ Wertes für n_0 umgangen werden; Hinweise hierzu entnehme man Rodger (1967, 192) und Bishop (1978).

Beispiele zur praktischen Durchführung dieses Verfahrens sowie weitere Einzelheiten finden sich in Rodger (1967, 1976), Guenther (1973, 217-220) und Bishop (1978); zu einem ganz ähnlichen Ansatz siehe Keppel (1973, 536-540).

Hinweise zur Bestimmung des Stichprobenumfanges für multiple Vergleiche nach einer (signifikanten) Varianzanalyse unter Verwendung der „Two-stage-sampling“-Prozedur geben die Arbeiten von Rodger (1974, 1975a, b, 1978) und von Hochberg & Lachenbruch (1976).

10.3.3 Ohne Kenntnis der Populationsvarianz σ_e^2

Bei den im folgenden zu skizzierenden Verfahren wird das Problem, vor dem Experiment die Größe von σ_e^2 nicht zu kennen, dadurch umgangen, daß der E den EEM als relativen Anteil der Effekt- an der Fehler- oder totalen Varianz angibt. Diese Art der Festlegung ist insofern als unspezifisch zu bezeichnen, als nicht mehr einzelne Mittelwertsdifferenzen oder entsprechende Quadratsummen als „praktisch bedeutsam“ angesetzt werden können, sondern nur noch eine (globale) relative „mittlere Differenz“. Andererseits dürfte eine derartige Festlegung noch am ehesten den in der Regel ebenfalls recht unspezifischen psychologischen Hypothesen entsprechen - vgl. Abschnitt 8.1.1.

10.3.3.1 Festlegung von φ^2

Je nach persönlichen Präferenzen trifft der E bei der hier anzusprechenden Vorgehensweise eine globale Festlegung für die Mindestgröße von φ^2 (vgl. Formel (10.1)) oder aber für f^2 resp. $R_{Y.X}^2$ (siehe dazu Formel (9.4) und (9.9)); die letztgenannten Indices sind anschließend in φ^2 umzurechnen (vgl. Formel (9.10) und (10.1)).

Unter Verwendung der entsprechenden Teststärke-Tabellen kann man dann die benötigte Stichprobengröße nach einem „Versuch-Irrtum-Verfahren“ ermitteln - vgl. zur Durchführung etwa Diehl (1979, 26-36) - oder aber direkt ablesen, wenn man günstiger aufgebaute Tabellen benutzt, wie sie etwa in Odeh & Fox (1975) zu finden sind.

Beispiele für diese und einige aus ihr abgeleitete Bestimmungsstrategien findet man u.a. in den Publikationen von Kirk (1968, 109f.), Levin (1972, 1975), Guenther (1973, 209f., 355-359), Keppel (1973, 529-541) und Diehl (1979, e. g. 29-39).

Es empfiehlt sich bei diesem und den vorhergehenden Ansätzen grundsätzlich, die Tabellen den Graphen vorzuziehen, weil bei letzteren die Ablesegenauigkeit geringer ist.

10.3.3.2 Festlegung von f^2 oder $R_{Y.X}^2$ (Verfahren nach Cohen)

Bei dem abschließend anzusprechenden Verfahren handelt es sich um den Teststärkenanalyse-Ansatz, der von Cohen (1962, 1965, 1969, 1970, 1973b, 1977; Cohen & Cohen, 1975) entwickelt worden ist.

Cohen kommt im Vergleich zu zahlreichen anderen Autoren (s. o.) mit einem leicht zugänglichen Minimum an Tabellen aus, die zudem so angelegt sind, daß

die Stichprobenumfangsbestimmung sowohl für einfache und komplexe varianzanalytisch wie für praktisch alle regressionsanalytisch auszuwertenden Versuchspläne relativ einfach vorgenommen werden kann.

Für die varianzanalytische Prüfung von Hypothesen über Mittelwerte verwendet Cohen das bereits im Abschnitt 9.3.1 eingeführte f^2 (vgl. auch Formel (10.1)) als Index für den experimentellen Effekt; für die entsprechende regressionsanalytische Prüfung von Hypothesen über multiple Korrelationsquadrate wird das Maß $R^2_{Y.X}$ benutzt - siehe dazu Abschnitt 9.3.2.1.

Neben der Verwendung bei der Planung von Experimenten können die in Cohen (1969, 1977) publizierten Tabellen auch zur Auswertung von Experimenten, insbesondere zur nachträglichen Bestimmung der Teststärke, herangezogen werden. Dies ist insofern vorteilhaft, als die dazu verfügbaren Computer-Programme (vgl. etwa Woodward & Overall, 1976) nur recht begrenzt verwendbar sind.

Wir wollen im folgenden trotz des begrenzten uns zur Verfügung stehenden Raumes den Gebrauch dieser Tabellen Cohens an einem Beispiel aufzeigen. Die Beschäftigung mit diesem Beispiel kann selbstverständlich die Lektüre des Buches von Cohen (1977) nicht ersetzen; sie kann jedoch deutlich machen, wie problemlos sich die Stichprobenumfangsbestimmung vornehmen läßt.

Gehen wir der Einfachheit halber davon aus, daß ein Zwei-Gruppen-Experiment zur Prüfung der folgenden statistischen Hypothese H_0 geplant ist:

$$H_0: \mu_1 - \mu_2 \leq 0 \text{ gegen die } H_1: \mu_1 - \mu_2 > 0.$$

Die notwendigen Voraussetzungen (siehe Abschnitt 8.2) werden als erfüllt vorausgesetzt, so daß der t-Test zur Anwendung kommen kann. Der E hat die folgenden Festlegungen getroffen:

$$\alpha = 0,01; \beta = 0,01; R^2_{Y.X} = 0,0099 \text{ (entspricht } f^2 = 0,01).$$

Es sollen die beiden Hypothesen bei gleichen maximalen Fehlerwahrscheinlichkeiten „fair“ gegeneinander getestet werden (vgl. dazu Abschnitt 11.1.2); die bei Fehlentscheidungen möglichen Folgen werden als gleich schwerwiegend angesehen. Es wird eine quadrierte und normierte Mittelwertsdifferenz dann als praktisch bedeutsam interpretiert, wenn sie mindestens 0,99% der Gesamt- oder 1% der Fehlervarianz ausmacht. Diese Werte müssen für die Benutzung der Tabelle 2.4.1 (Cohen, 1977, 54f.) zunächst in den von Cohen gewählten Index d umgerechnet werden, der wie folgt definiert ist:

$$(10.3) \quad d = \frac{\mu_1 - \mu_2}{\sigma_e} = 2f$$

Aus $f = 0,1$ folgt, daß $d = 0,2$; es handelt sich nach Cohens vorgeschlagener Konvention (1977, 24-27) um einen „kleinen Effekt“.

Die vorgegebenen Informationen reichen aus, um direkt aus dem mit „ $\alpha_1 = 0,01$ “ überschriebenen Teil der gen. Tabelle den pro experimenteller Bedingung benötigten Stichprobenumfang abzulesen:

$$n = 1084 \text{ und somit insgesamt } 2n = N = 2168 \text{ Vpn.}$$

Die bei der Größe α vorgenommene Indizierung mit „1“ besagt, daß ein einseitiger Test zugrunde gelegt wird; bei einem vorgesehenen zweiseitigen Test (einer ungerichteten Hypothese) hätte die Ablesung des Wertes für n unter dem mit „ $\alpha_2 = 0,01$ “ überschriebenen Teil der Tab. 2.4.1 vorgenommen werden müssen. Es gilt hier wie auch für die Varianz- und regressionsanalytischen Tabellen, daß einem einseitigen Test bei $\alpha_1 = c$ ein zweiseitiger Test bei $\alpha_2 = 2c$ entspricht und umgekehrt.

üblicherweise ist ein E nicht in der Lage, die soeben bestimmte Anzahl von Vpn verfügbar zu machen. Wenn wir davon ausgehen, daß eine Vergrößerung des zu entdeckenden praktisch bedeutsamen Mindesteffektes zunächst nicht in Frage kommt, kann sich der E mit der Erhöhung der maximalen Irrtumswahrscheinlichkeiten behelfen (vgl. zu dieser Empfehlung jedoch auch Abschnitt 8.3). Er kann also etwa festlegen:

$$\alpha = 0,10 \text{ und } \beta = 0,20.$$

Diese Festlegung impliziert, daß die Risiken, die mit einer Fehlentscheidung zugunsten der statistischen Null-Hypothese verbunden sind, als halb so schwerwiegend angesehen werden wie die, die bei einer falschen Entscheidung für H_1 entstehen können - vgl. Abschnitt 11. Dem mit „ $\alpha_1 = .10$ “ überschriebenen Teil der Tab. 2.4.1 ist unter „d = .20“ und „Power = .80“ zu entnehmen, daß $n = 226$; also sind insgesamt 452 Vpn den experimentellen Behandlungen zu unterziehen. Sind auch diese Versuchspersonenmengen für den Experimentator nicht verfügbar, kann er sich in dieser Situation nur noch auf die Entdeckung eines größeren experimentellen Effektes festlegen, wenn er die Fehlerwahrscheinlichkeiten nicht weiter vergrößern will und wenn eine andere Prüfung der wissenschaftlichen Hypothese nicht möglich ist.

Nehmen wir an, der E entschließt sich, einen Effekt dann als psychologisch bedeutsam zu interpretieren, wenn die Effektvarianz (zwischen den Mittelwerten) mindestens 16% der Fehlervarianz oder 13,79% der Gesamtvarianz beträgt, es sich also nach Cohens vorgeschlagener Konvention um einen „großen Effekt“ handelt (Cohen, 1977, 24-27). Aus dieser Festlegung ergeben sich die folgenden Werte:

$$f^2 = 0,16; R^2_{Y.X} = 0,1379; d = 0,8$$

Unter Verwendung der o.a. liberalen Kriterien bzgl. α und β läßt sich der entsprechenden Tabelle entnehmen, daß der E jetzt nur noch $n = 14$ oder insgesamt 28 Vpn benötigt.

Für die zweiseitige Prüfung einer ungerichteten Hypothese müßten unter den gleichen Bedingungen $n = 20$ ($N = 40$) Vpn an dem Experiment teilnehmen.

In ähnlicher Weise, wie es hier am Beispiel aufgezeigt wurde, gestaltet sich auch die Bestimmung des Stichprobenumfanges für einen mittels t-Test auszuwertenden Versuchsplan mit n Paaren von abhängigen Messungen - siehe dazu im einzelnen Cohen (1977, Kap. 2).

Für Varianz- oder regressionsanalytisch auszuwertende ein- oder mehrfaktorielle Versuchspläne finden sich entsprechende Tabellen auf den Seiten 440-442 des Buches von Cohen (1977); ihre Benutzung gestaltet sich am einfachsten, wenn der experimentelle Effekt als partielles oder semi-partielles multiples Korrelationsquadrat $R^2_{Y \cdot X}$ (vgl. dazu insbesondere Kerlinger & Pedhazur, 1973, Kap. 5; Cohen & Cohen, 1975, e.g. 78-84, 129-134) festgelegt werden kann. Der nachträglichen Bestimmung der Teststärke aus bereits erhobenen Daten dienen die analog aufgebauten Tabellen auf den Seiten 416-418 (Cohen, 1977).

Die Tabellen in Cohen (1977, Kap. 9) eignen sich zur Bestimmung der benötigten Stichprobenumfänge nicht nur für varianzanalytische Versuchspläne, sondern darüber hinaus auch für Trendanalysen (vgl. Cohen & Cohen, 1975, Kap. 6; Bredenkamp, 1980, 56-59), Meßwiederholungspläne (Cohen & Cohen, 1975, Kap. 10; Bredenkamp, 1980, 91-99), Parallelisierungspläne (Cohen & Cohen, 1975, Kap. 10; Bredenkamp, 1980, 71-73) und Kovarianzanalysen (Cohen, 1977, 379f.; Bredenkamp, 1980, 78-80).

10.4 Hinweise zur Stichprobenumfangsbestimmung bei weiteren Gruppen von parametrischen Testverfahren

10.4.1 Varianzanalyse mit zufälligen und gemischten Effekten

Die Unterscheidung zwischen drei varianzanalytischen Modellen, die wir kurz im Abschnitt 8.2.1 angesprochen haben, ist sowohl für die Auswertung als auch für die Planung des Experiments von Bedeutung.

Zuweilen bezieht sich eine wissenschaftliche Hypothese auf eine (sehr) große Menge möglicher Treatmentmodalitäten, die der E aus verschiedenen Gründen nicht in ihrer Gesamtheit untersuchen kann. In diesem Fall besteht für den E die Möglichkeit, aus dieser Population von Bedingungen eine Zufallsstichprobe zu ziehen oder aber die realisierten Modalitäten als eine solche zu interpretieren. Nach der Auffassung zahlreicher Autoren (e. g. Bredenkamp, 1975, 810; Hays, 1977, 542) rechtfertigt dieses Vorgehen die statistische Verallgemeinerung der Resultate auch auf alle nicht untersuchten Bedingungen, die in der WH enthalten sind.

Ein derartiger Faktor ist im Sinne der varianzanalytischen Modelle „zufällig“ (vgl. Abschnitt 8.2.1); üblicherweise stellt der Versuchspersonen-Faktor bei Meßwiederholungen einen zufälligen Faktor dar (vgl. Abschnitt 8.4.3).

Wir haben im Abschnitt 2.3 Gründe angegeben, die in aller Regel eine systematische Auswahl von Modalitäten nahelegen. Darüber hinaus ist der F-Test beim Modell der zufälligen Effekte, bei dem Hypothesen über Varianzen, nicht über Mittelwerte geprüft werden, sehr empfindlich gegenüber Verletzungen seiner Voraussetzungen, insbesondere der der Normalverteilung der

Modalitäten des zufälligen Faktors in der Population (vgl. Lehmann, 1968, 45; Glass & Stanley, 1970, 462; Hays, 1977, 540f., sowie Abschnitt 8.2.4). Aus diesen Gründen halten wir insgesamt das zufällige und das gemischte Modell der VA (vgl. Abschnitt 8.2.1) für wenig empfehlenswert, sofern seine Verwendung nicht aufgrund der Fragestellung zwingend indiziert ist; bspw. wird der Versuchspersonen-Faktor bei Meßwiederholungen univariat varianzanalytisch i.a. unter Annahme dieses Modells ausgewertet.

Entschließt sich der E, trotz der angeführten Bedenken dem Modell der zufälligen Effekte den Vorzug zu geben, muß er die Auswertung der Daten entsprechend vornehmen, da teilweise andere Prüfvarianzen im Nenner des F-Bruches resultieren als beim Modell der fixierten Effekte - vgl. zu den Einzelheiten etwa Hays (1977, Kap. 13).

Darüber hinaus folgt der F-Bruch bei zufälligen Effekten auch unter Gültigkeit der statistischen Alternativhypothese den *zentralen* F-Verteilungen. Aufgrund dieser Tatsache muß auch die Bestimmung der Teststärke und der benötigten Stichprobengröße nicht mehr über die nicht-zentralen, sondern über die zentralen F-Verteilungen erfolgen. Auf die Einzelheiten gehen wir nicht ein, sondern verweisen auf John (1971, 53f.), Lindman (1975, 118-135), Hays (1977, 539f.) und Henning & Muthig (1979, 241-248), die das Vorgehen demonstrieren.

10.4.2 Nicht-orthogonale Varianzanalysen

Die bisherigen Stichprobengrößenbestimmungen erfolgten stets derart, daß gleiche Umfänge pro Treatmentmodalität (und pro Zelle des Versuchsplanes) resultierten. Da bei psychologischen Experimenten in der Regel davon ausgegangen werden kann, daß der Erwartungswert der Korrelation zwischen den untersuchten Faktoren oder UV, gleich Null ist, stellt die Gleichbesetzung der Zellen mit Untersuchungseinheiten das adäquate Vorgehen dar (vgl. etwa Yates, 1933).

Bei der Datenerhebung kann jedoch aus den verschiedensten Gründen der Fall eintreten, daß Vpn ausfallen („experimentelle Mortalität“), so daß ungleiche Zellenbesetzungen resultieren. Dieser Ausfall hat zur Folge, daß die Korrelation zwischen den experimentellen Faktoren nicht mehr gleich Null ist (vgl. zur genaueren Charakterisierung der Nicht-Orthogonalität etwa Steyer, 1979, e.g. 82). In diesem Fall sind auch die Quadratsummen der Varianzanalyse nicht mehr additiv (vgl. Steyer, 1979, 79f., 89f.).

Die Frage, auf welche Art Experimente auszuwerten sind, deren Faktoren nicht orthogonal zueinander sind, wird - ausgehend von einem Artikel von Overall & Spiegel (1969) - kontrovers diskutiert. Die wesentlichsten Aspekte dieser Diskussion hat Steyer (1979) zusammengefaßt und vertieft (vgl. auch Bredenkamp, 1980, 74-78).

Unter den Publikationen neueren Datums sind insbesondere Milliken & McDonald (1976), Hamer & Hosking (1977), Lewis & Keren (1977), Blair & Higgins (1978), Herr

& Gaebelein (1978), Jennings (1978) sowie Cramer & Appelbaum (1980) zu erwähnen; empirische Untersuchungen zur Angemessenheit der verschiedenen in der Literatur vorgeschlagenen Auswertetechniken haben u.a. Levy, Narula & Abrami (1975), Narula, Abrami & Levy (1976) (zur Kritik dieser Arbeiten siehe Hummel, 1977) und Steyer (1979) vorgelegt.

Den genannten Arbeiten ist zu entnehmen, daß die Verwendung der üblichen approximativen varianzanalytischen Techniken (vgl. u.a. Winer, 1971) sehr häufig nicht anzuraten ist. Statt dessen sollte auf die regressionsanalytischen Verfahren zurückgegriffen werden, die die Korrelationen zwischen den UVn zu erfassen erlauben und in der Regel exakte Tests der geprüften Hypothesen ermöglichen. Die Korrelationen können dabei je nach Ausgangslage entweder auspartialisiert oder aber in die Auswertung einbezogen werden; bzgl. der Einzelheiten verweisen wir insbesondere auf Bredenkamp (1975, 809; 1980, 74-79), ferner auf die entsprechenden Kapitel in Gaensslen & Schubö (1973, 1976), Kerlinger & Pedhazur (1973) sowie Cohen & Cohen (1975).

Für die Interpretation der Daten ist ungeachtet der statistischen Auswertung jedoch zu bedenken, daß ein nicht-zufälliger (d.h. z.B. von der Art des Treatments abhängiger) *Ausfall* von Vpn zu Beeinträchtigungen der internen Validität führt (vgl. Jurs & Glass, 1971; sowie Abschnitt 3.3.2), während eine *nachträgliche Orthogonalisierung* der Faktoren die Populationsvalidität herabsetzen kann (siehe Jurs & Glass, 1971; Bredenkamp, 1975, 809; und Abschnitt 4.1).

10.4.3 Multivariate Varianz- und Regressionsanalysen

Werden in einem Experiment die Auswirkungen der unabhängigen Variable(n) nicht nur auf eine, sondern auf mehrere abhängige Variable untersucht, liegt eine multivariate Versuchsanordnung vor (vgl. Abschnitt 1.1). Wir haben im Abschnitt 8.4.6.2.2 einige Gründe genannt, die zur Auswertung von multivariaten Versuchsplänen das Testkriterium V von Bartlett (1937, 1939) und Pillai (1955) nahelegen.

Dieses V stellt die Summe der kanonischen Korrelationsquadrate dar; sind die Prädiktorvariablen orthogonal, ergibt sich V als die Summe der multiplen Korrelationsquadrate zwischen je einer der UVn $\mathbf{X}_z$ und jeweils allen AVn $\mathbf{Y}_q$ (vgl. Kerlinger & Pedhazur, 1973; Bredenkamp, 1980).

Für die Stichprobenumfangsbestimmung für Tests von multiplen Korrelationsquadraten können die entsprechenden Tabellen von Cohen (1977, 440-442) verwendet werden, so daß sich unter Verwendung des V-Kriteriums keine grundsätzlichen Schwierigkeiten bei der Planung multivariater Versuchspläne ergeben. Welche Überlegungen hierbei im einzelnen anzustellen sind und wie zweckmäßigerweise vorzugehen ist, schildert ausführlicher Bredenkamp (1980).

Bei der Planung des Stichprobenumfangs für ein einfaktoriell-multivariates Experiment, das mittels des T^2 -Tests nach Hotelling (1931) ausgewertet werden soll, können zudem die von Läuter (1978) vorgelegten Tabellen verwendet werden.

Entschließt man sich zur Anwendung eines der übrigen multivariaten Analyseverfahren (vgl. dazu die im Abschnitt 7.5.3.3 angegebene Literatur), kann die Stichprobenumfangsbestimmung derzeit nur approximativ über die o.a. Tabellen erfolgen, weil ausführliche exakte Tabellen für diese Verfahren noch nicht verfügbar sind (vgl. Olson, 1974; Stevens, 1980).

10.5 Hinweise zur Stichprobenumfangsbestimmung bei nicht-parametrischen Verfahren

10.5.1 *Nominale Daten*

Zur Auswertung nominaler Daten, die etwa in Kontingenztafeln angeordnet sind, wird in Abhängigkeit von der Anzahl der untersuchten Merkmalskategorien vorwiegend die Binomial- oder die Multinomialverteilung verwendet - vgl. Abschnitt 7.5.3.1.

Ist die Anzahl der zu untersuchenden Merkmale so festgelegt, daß zur Auswertung einer der Binomialtests in Frage kommt, kann die Bestimmung des (insgesamt) benötigten Stichprobenumfangs über die speziellen Tabellen erfolgen, die etwa Fleiss (1973, 176-194, ferner Kap. 3), Cohen (1977, Kap. 4 und 5) sowie Bortz, Österreich & Vogelbusch (1979) vorgelegt haben. Auch für einen Vorzeichen-Test läßt sich der Stichprobenumfang unter Verwendung dieser Tabellen bestimmen.

Will man die entsprechende Auswertung approximativ über einen der vielseitigen χ^2 -Tests vornehmen, die sowohl zur Analyse der o. gen. Kontingenztafeln als auch u.a. zur Prüfung der Anpassungsgüte verwendet werden können, findet man bei Cohen (1977, Kap. 7) ebenfalls entsprechende Tabellen zur Ermittlung des benötigten Mindeststichprobenumfangs.

Die Bestimmung des im Experiment aufgedeckten EE erfolgt jeweils unter Verwendung eines der im Abschnitt 9.4.2 vorgestellten resp. angesprochenen Maße.

10.5.2 *Ordinale Daten*

Für die insgesamt recht zahlreichen Techniken zur Auswertung von ordinalen oder Rangdaten liegen Tabellen zur Bestimmung des Stichprobenumfangs

und damit allgemein zur Teststärkenanalyse derzeit noch nicht vor, weil die Ableitung der zu ihrer Erstellung benötigten nicht-zentralen Verteilungen von Rangstatistiken noch weitgehend aussteht (vgl. jedoch zu ersten Ansätzen Kraemer, 1974 und Henze, 1979).

Dennoch sollte man auch bei der vorgesehenen Verwendung dieser Gruppe von Verfahren nicht auf die notwendigen Festlegungen bzgl. der Maximalwerte für beide Fehlerarten und den experimentellen Effekt verzichten.

Das Festlegen des experimentellen Mindesteffektes und die Determination des Stichprobenumfanges kann dann näherungsweise über die für das analoge parametrische Verfahren benutzten Effektgrößenindices und Bestimmungstabellen vorgenommen werden.

Diese vorläufige Lösung scheint uns gerechtfertigt, wenn man sich vergegenwärtigt, daß nicht-parametrische Tests fast stets dann effizienter sind als ihre parametrischen Homologa, wenn die zur Anwendung der letzteren notwendigen Voraussetzungen nicht gegeben sind (vgl. zu Einzelheiten die Abschnitte 7.5.3.2 und 8.2).

Will man mit großer Sicherheit eine festgesetzte Mindestteststärke nicht unterschreiten, empfiehlt sich folgendes Vorgehen: Man bestimmt den bei vorgegebenem Signifikanzniveau, Mindesteffekt und festgelegter Teststärke benötigten Stichprobenumfang über die entsprechenden parametrischen nicht-zentralen Verteilungen, wie sie etwa in Cohen (1977) tabelliert sind. Das Ergebnis dividiert man durch die für das vorgesehene nicht-parametrische Verfahren angegebene Maßzahl der relativen asymptotischen Effizienz, die der im Abschnitt 7.5.3 angegebenen Literatur entnommen werden kann - Beispiele für diese Vorgehensweise finden sich in Bredenkamp (1980, 61 f., 73f.). Der experimentelle Effekt läßt sich anschließend etwa unter Verwendung eines der im Abschnitt 9.4.1 angegebenen Maße abschätzen.

10.6 Abschließende Bemerkungen zur Stichprobengrößenbestimmung

Der vorangegangene Teil 10 dieses Artikels sollte den Planer und Auswerter von psychologischen Experimenten mit den unterschiedlichen Ansätzen bekannt machen, aufgrund derer die Größe von Stichproben nach einem vorgegebenen Kriterium bestimmt werden kann.

Den vorgestellten Verfahren ist gemeinsam, daß sie auf die Kontrolle und Minimierung der Wahrscheinlichkeiten für falsche Entscheidungen abzielen. Man kann derartige Ansätze als einen wesentlichen Aspekt des umfassenderen

Konzepts der Teststärkenanalyse auffassen, über die Cohen (1977, Kap. 1) ausführlich informiert.

Zu den vielseitigen Möglichkeiten zur praktischen Anwendung dieses Konzepts zählen dabei nicht nur die hier skizzierten Verfahren zur Ermittlung von Stichprobenumfängen, sondern auch die mittels der Tabellen Cohens (1977) leicht mögliche Bestimmung der Teststärke bei vorgegebener Stichprobengröße, also nach der Datenerhebung und statistischen Hypothesenprüfung. Diese Anwendung ist als wesentlicher Bestandteil von statistischen Reanalysen anzusehen.

Abschließend bleibt nur noch zu hoffen, daß Verfahren der Art, wie sie hier vorgestellt worden sind, in absehbarer Zeit die sporadisch propagierten „Faustregeln“ (vgl. etwa Cowles, 1974; Pfanzagl, 1978, 101-106; McGuigan, 1979, 287) ersetzen werden.

11. Eine Strategie zur Entscheidung über wissenschaftliche Hypothesen mittels Signifikanztests

11.1 Stadium der Planung des Experiments

11.1.1 Überblick

Wir gehen im folgenden davon aus, daß der Experimentator (E) aus der wissenschaftlichen Hypothese WH eine statistische Alternativhypothese H_1 abgeleitet hat, unsere Überlegungen sind aber auch auf jene selteneren Fälle zu übertragen, in denen eine Nullhypothese impliziert wird (vgl. Teil 6 und 8.1). Soll eine statistische Hypothese über einen Signifikanztest geprüft werden, hat sich der E auch mit den Determinanten eines solchen Tests auseinanderzusetzen: das Signifikanzniveau α , die Wahrscheinlichkeit β für einen Fehler 2. Art, den tatsächlichen experimentellen Effekt EEP sowie die Präzision des Experiments, die definiert ist durch den Standardfehler der benutzten Teststatistik, der seinerseits von der Stichprobengröße N und (bei parametrischen Verfahren) von der Fehlervarianz σ_c^2 abhängt. Zwischen diesen Größen bestehen eine Reihe von Beziehungen, die im Teil 7 dargestellt worden sind. Dort ist auch bereits deutlich geworden, daß viele empirisch arbeitende Psychologen diesen Abhängigkeitsbeziehungen nur geringe oder gar keine Beachtung schenken. Insbesondere führt die Mißachtung der Fehlerwahrscheinlichkeit β dazu, daß Entscheidungen über statistische Hypothesen getroffen werden, die mit einer relativ hohen Wahrscheinlichkeit falsch sind. Da dies sowohl zu falschen Schlüssen über die Gültigkeit psychologischer Hypothesen und Theorien führen kann wie zu ungerechtfertigten Handlungsvorschlägen für praktische Si-

tuationen, ist die fortlaufende Mißachtung fundamentaler Gesetzmäßigkeiten bei der Anwendung von Signifikanztests nicht zu tolerieren. Dies um so weniger, als bereits mit geringem Aufwand dafür Sorge getragen werden kann, daß der relative Anteil von falschen Entscheidungen aufgrund von statistischen Hypothesenprüfungen vom Forscher selbst durch die Kontrolle *beider* statistischen Fehlerwahrscheinlichkeiten niedrig gehalten werden kann. Dieses Ziel wird durch folgende *Planungsstrategie* erreicht:

Der Experimentator legt die ihm als maximal zulässig erscheinende Irrtumswahrscheinlichkeit 1. Art, α , auf einen kleinen Wert fest, um auf diese Weise die Gefahr gering zu halten, ein fälschlich die Hypothese WH bestätigendes Ergebnis zu erhalten.

Er fixiert die Wahrscheinlichkeit β für einen Fehler 2. Art ebenfalls auf einen kleinen Wert, um die Gefahr fälschlicher Falsifikationen der WH gering zu halten.

Er gibt an, wie groß der experimentelle Effekt EEM mindestens sein muß, um für die Prüfung der WH relevant („praktisch bedeutsam“) zu sein (s. Teil 9).

Der Experimentator kann dann den Stichprobenumfang N bestimmen, der notwendig ist, um den erwünschten Mindesteffekt EEM bei festgelegter Teststärke $1 - \beta$ und gewähltem Signifikanzniveau α als signifikant ausweisen zu können, falls dieser tatsächlich vorliegt (s. Teil 10).

Hält der Experimentator die ermittelte notwendige Stichprobengröße N aus ökonomischen Gründen (im weitesten Sinne) für zu groß, sollte er die Werte für α , β und EEM unabhängig voneinander und in verschiedenen Kombinationen erhöhen und dazu das jeweils notwendige N bestimmen. Nach den Erfordernissen des Einzelfalles kann er sich dann für eine Kombination von vertretbaren Werten für diese vier Größen entscheiden. Diese Entscheidung sollte samt ihrer Begründung in jedem Forschungsbericht explizit erwähnt werden.

In einigen Ausnahmefällen dagegen kann die vorgestellte Strategie zu dem Ergebnis führen, daß weniger als etwa fünf V_{pn} in jeder experimentellen Bedingung ausreichend sind (vgl. Abschnitt 10.1). Wir halten es bei derart kleinen Stichprobenumfängen für kaum möglich, zu einer hinreichenden Sicherung der internen Validität (vgl. Abschnitt 3.3.1) und der Populationsvalidität (vgl. Teil 4) zu gelangen.

11.1.2 Zur Festlegung der beiden Fehlerwahrscheinlichkeiten

Bezüglich der Festlegung der Wahrscheinlichkeit für einen Fehler 1. Art haben sich bestimmte Konventionen durchgesetzt, nämlich die überwiegende Ver-

Wendung von $\alpha = 0,05$ und $\alpha = 0,01$. Diese Präferenz ist z.T. darauf zurückzuführen, daß die kritischen Werte auch der häufiger gebrauchten Teststatistiken oft nur für diese Wahrscheinlichkeiten veröffentlicht werden. Zahlreiche Simulationsstudien zur Robustheit parametrischer Tests (vgl. Abschn. 8.2) legen jedoch nahe, das Signifikanzniveau *nicht ohne zwingenden Grund* zu klein werden zu lassen, falls die Vermutung besteht, daß erhebliche Abweichungen von der Normalverteilungsannahme vorliegen. Die Stichprobenverteilungen der meisten parametrischen Teststatistiken sind nämlich besonders in den extremen Wertebereichen („Enden“) sehr empfindlich gegenüber derartigen Abweichungen (vgl. Abschn. 7.5.1). Man sollte deshalb eher $\alpha = 0,05$ als $\alpha = 0,01$ wählen; dies gilt insbesondere bei einseitigen Tests. Eine Erhöhung des Signifikanzniveaus etwa auf $\alpha = 0,10$ oder $\alpha = 0,20$ kann notwendig werden, wenn der berechnete Stichprobenumfang praktisch nicht zu realisieren ist (s. o.). Dagegen kann eine Verringerung des α -Niveaus angezeigt sein, wenn zur Prüfung einer wissenschaftlichen Hypothese mehrere statistische Hypothesenprüfungen durchgeführt werden (s. Abschn. 8.3).

Für die Festlegung der maximalen Wahrscheinlichkeit für einen Fehler 2. Art gibt es keine vergleichbar weithin anerkannten Konventionen. Cohen (1965; 1969, 51-54) hat jedoch vorgeschlagen, die Teststärke $1 - \beta$ nicht geringer als 0,80 werden zu lassen. Dies bedeutet, daß „auf lange Sicht“ stets vier von fünf zutreffenden Alternativhypothesen aufgrund der experimentellen Daten als richtig ausgewiesen werden sollten (vgl. auch Brewer, 1972, 394). Einem „fairen“ Test entspräche die Festlegung $\alpha = \beta$ (vgl. Bredenkamp, 1969b, 283; Bailey, 1971, 333-336). Hierbei haben beide statistischen Hypothesen unter ihrer Richtigkeit die gleiche Chance, sich zu bewähren. Wählt man dabei allerdings beide Wahrscheinlichkeiten sehr klein (etwa $\alpha = \beta = 0,01$), hat diese Festlegung in der Regel zur Folge, daß sehr große Stichproben untersucht werden müßten (vgl. Abschn. 10.3.3.2). Bei der daher meist unumgänglichen Festlegung unterschiedlicher Werte für α und β muß bedacht werden, daß damit zumindest implizit eine Gewichtung der zu prüfenden Hypothesen vorgenommen wird. Legen wir bspw. die Werte $\alpha = 0,05$ und $\beta = 0,20$ fest, wird damit ausgesagt, daß ein Fehler 1. Art viermal so schwer wiegt wie ein Fehler 2. Art (vgl. hierzu im einzelnen Cohen, 1977, 56; Krause & Metzler, 1978, 254f.; ferner Lehmann, 1958 sowie Rasch et al., 1978, 90f.). Bei der Wahl der maximalen Werte für die statistischen Fehlerwahrscheinlichkeiten sollte demnach stets berücksichtigt werden, welche Konsequenzen sich aus der irrtümlichen Annahme oder Zurückweisung einer Hypothese ergeben können. Diese Überlegungen können u.U. auch zu „ungewöhnlichen“ Festlegungen wie $\alpha = 0,20$ und $\beta = 0,05$ führen (vgl. Abschn. 8.3).

11.1.3 Zur Festlegung des experimentellen Mindesteffektes EEM

Wie oben begründet, sollte der E neben α und β auch angeben, wie groß der Experimentelle Effekt als Maß für die statistische Assoziation mindestens sein muß, um für ihn nach bestimmten inhaltlichen Kriterien „praktisch bedeutsam“ zu sein. Für dieses Ausmaß des statistischen Zusammenhangs zwischen zwei (oder mehr) Variablen sind ebenfalls Konventionen vorgeschlagen worden (Cohen, 1969, 1977): So wird beispielsweise ein $r^2 = 0,01$ als „kleiner“, ein $r^2 = 0,09$ als „mittlerer“ und ein $r^2 = 0,25$ als „großer Effekt“ bezeichnet (vgl. auch Abschn. 10.2.3.2).

Da aus der Mehrzahl der derzeit vorliegenden psychologischen Hypothesen keine Aussagen über die Größe der bei ihrer Gültigkeit zu erwartenden statistischen Assoziationen abgeleitet werden können, muß gerade zu Beginn eines Forschungsprogramms zwangsläufig auf solche Konventionen zurückgegriffen werden. Später können dann die in vergleichbaren Experimenten zur Prüfung der gleichen Theorie oder Hypothese aufgedeckten Effekte zur Grundlage der Planung weiterer Untersuchungen gemacht werden (s. Abschn. 11.2).

11.1.4 Zur Frage der Willkür bei der Planung von Experimenten

Die eben skizzierte Planungsstrategie kann insofern kritisiert werden, als die Werte für α , β und EEM, die notwendigerweise vor der Datenerhebung festzulegen sind, in aller Regel stark von der Willkür des einzelnen Experimentators abhängen. Gegen diesen Einwand führt Bredenkamp (1969 b, 278f.) aus: „Die willkürlich aussehende Fixierung einer praktisch bedeutsamen Differenz und der maximalen Wahrscheinlichkeit eines Typ-II-Fehlers führt dazu, daß durch die (nicht zu begründende) Wahl des Stichprobenumfangs für jedes β eine Differenz implizite als bedeutsam festgelegt worden ist. . . . doch hat die explizite Festlegung einer solchen Differenz den Vorteil, daß anderen Forschern mitgeteilt werden muß, wie groß ein Mittelwertsunterschied zu sein hat, wenn die Alternativhypothese mit der Wahrscheinlichkeit $1 - \beta$ akzeptiert werden soll.“

11.2 Stadium der Entscheidung über die Kausalhypothese

Wir werden nun darstellen, wie man aufgrund der Ergebnisse von Signifikanztests zu Entscheidungen über Beibehaltung oder Falsifikation wissenschaftlicher Hypothesen gelangen kann. Wir halten es für unabdingbar, diese Entscheidung nicht nur auf dem Wert der Teststatistik beruhen zu lassen, sondern dabei auch den Wert EES für die Größe des aufgedeckten experimentellen Effektes zu beachten, der sich aus den jeweils vorliegenden Daten berechnen läßt.

Wir erläutern unsere *Entscheidungsstrategie* zunächst für die Prüfung einer gerichteten statistischen Hypothese durch einen einseitigen Test (z.B. t-Test), wobei aus der wissenschaftlichen Hypothese weiterhin H_1 folgen soll.³³⁾ In diesem Fall sind 4 Situationen zu unterscheiden:

(1) Die Teststatistik fällt in den von H_0 abgedeckten Bereich. Bei einer $H_1 : \mu_1 - \mu_2 > 0$ heißt das beispielsweise, daß die Teststatistik t einen Wert kleiner oder gleich Null annimmt. Wir nehmen dann die H_0 an. Sollten derartige Ergebnisse wiederholt auftreten, kann man sich zur Falsifikation der wissenschaftlichen Hypothese entschließen.

(2) Die Teststatistik fällt in den Rejektionsbereich, und der EES ist größer als der festgelegte Wert EEM. In diesem Fall wird die H_1 akzeptiert, und die WH hat sich (vorläufig) bewährt.

(3) Der Wert der Teststatistik fällt in den Rejektionsbereich, EES ist jedoch geringer als EEM. Wie im Fall (2) wird die H_1 angenommen, denn die Wahrscheinlichkeit, daß dies irrtümlich geschieht, ist höchstens gleich α , also einem (sehr) geringen Wert. Der Ausgang des Experiments spricht grundsätzlich für die wissenschaftliche Hypothese; allerdings sollte in einem Folgeexperiment untersucht werden, ob der kleinere EE replizierbar ist. Ist dies nicht oder nur sporadisch der Fall, sollte man die WH als schlecht bewährt ansehen. Auf die andere Möglichkeit, daß der geringere Effekt wiederholbar ist, gehen wir im Anschluß an Situation (4) ein.

(4) Der Wert der Teststatistik weist auf eine Abweichung von der H_0 in der von H_1 spezifizierten Richtung hin, ist aber zu klein, um zu einem signifikanten Ergebnis zu führen. Nach unserer Entscheidungsstrategie im Signifikanztest nehmen wir dann die H_0 an; das bedeutet, daß der experimentelle Effekt als zufällige Abweichung interpretiert wird. Dies spricht zunächst einmal gegen die Kausalhypothese.

Weitere Prüfungen sollten mit einem kleineren EEM, einer größeren Zahl von Versuchspersonen und dadurch mit einer höheren Präzision durchgeführt werden. Wird dabei insgesamt das zuerst erhaltene Ergebnis repliziert, muß das aber nicht unbedingt zur Falsifikation der wissenschaftlichen Hypothese führen. Ergeben sich nämlich mit einer gewissen Regelmäßigkeit nur sehr kleine Abweichungen von der Nullhypothese in Richtung auf die Alternativ-

³³⁾ Wir erweitern und modifizieren damit den Gedanken eines dreigeteilten Stichprobenraums, der erstmals von Neyman & Pearson (1933 b, 493) entwickelt wurde und der dann mit unterschiedlichen Akzentuierungen nacheinander von Lehmann (1950, 23f.), Kaiser (1960), Meehl (1967), Kleiter (1969), Bredenkamp (1972, 80-84) und Bortz, Österreich & Vogelbusch (1979) wieder aufgegriffen und weiterentwickelt worden ist (zur ausführlichen Darstellung siehe Westermann & Hager, 1982).

hypothese, kann der E zwischen zwei Möglichkeiten entscheiden, und zwar unabhängig davon, ob die Abweichungen überwiegend „signifikant“ sind (obiger Fall 3) oder nicht (Fall 4): Er kann die wissenschaftliche Hypothese falsifizieren, weil sie zu wenig an Datenvarianz aufklärt (Bredenkamp, 1972, 83-84), oder er behält sie bei, allerdings mit der Einschränkung, daß ihr nur geringe statistische Assoziationen entsprechen.

Bei dieser Entscheidung ist auch zu berücksichtigen, inwieweit für die durchgeführten Untersuchungen die besprochenen Aspekte der experimentellen Validität gegeben sind. So kann es z.B. von besonderem Interesse sein, zu prüfen, inwieweit sich die Effekte auch unter „natürlichen“ Bedingungen (Feldexperimente, nicht-reaktive Operationalisierungen usw.) zeigen.

Ferner ist bei Falsifikationsentscheidungen stets auch zu berücksichtigen, ob eine alternative Hypothese zur Erklärung der empirischen Ergebnisse gefunden werden kann. Damit derartige Entscheidungen nachvollziehbar und auch kritisierbar werden, ist es unbedingt erforderlich, für jedes Experiment als spezielles und wichtiges Ergebnis auch die Größe des aufgedeckten experimentellen Effektes mitzuteilen.

Muß die wissenschaftliche Hypothese über einen *zweiseitigen Test* geprüft werden, tritt zusätzlich folgendes Problem auf: Die Nullhypothese (z. B. $\mu = 0$) definiert keine Parametermenge, die eindeutig für H_0 spricht, sondern nur genau *einen* Wert. Da die Signifikanztests in der Regel von kontinuierlichen Variablen ausgehen, ist dann die Wahrscheinlichkeit, einen Wert der Teststatistik zu erhalten, der genau dem in H_0 spezifizierten Wert entspricht, gleich Null. Eine Annahme von H_0 (und damit eine Ablehnung von H_1) ist deshalb bei zweiseitigen Tests nur möglich, wenn man um den in H_0 spezifizierten Parameterwert eine Menge von Stichprobenergebnissen festlegt, die - falls sie auftreten - auch noch für H_0 sprechen sollen. Beim hier vorgeschlagenen Signifikanztest in der Neyman-Pearson-Tradition erfolgt dies durch eine vorherige Festlegung von EEM, α und β . Dadurch wird ein Bereich von möglichen Werten der Stichprobenstatistik (oder äquivalent der Effektgröße) festgelegt, der zur Annahme von H_0 führt. Dieser Bereich liegt zwischen dem unteren und dem oberen kritischen Wert der Teststatistik und entspricht Effektgrößen zwischen Null und einem Wert EEK, der bei adäquat fixierter Teststärke notwendigerweise kleiner ist als der gewählte EEM.

Wir unterscheiden beim zweiseitigen Signifikanztest drei Entscheidungssituationen:³⁴⁾

³⁴⁾ Der Einfachheit halber gelte weiterhin, daß aus der wissenschaftlichen Hypothese die statistische Alternativhypothese folgt, obwohl dieser Fall weniger häufig auftritt (vgl. Abschn. 8.1.1, Punkt 2). Die hier dargestellten Entscheidungsstrategien sind aber problemlos auf den Fall der Implikationsbeziehung $WH \rightarrow H_0$ übertragbar; vgl. dazu Westermann & Hager (1982).

(1) Der Test führt nicht zu einem signifikanten Ergebnis, EES ist also kleiner als EEK. H_0 ist in diesem Falle anzunehmen, und bei zuverlässigen Resultaten dieser Art kann die WH falsifiziert werden.

(2) Der empirische Wert der Teststatistik fällt in den Rejektionsbereich, und EES ist nicht kleiner als EEM. Es wird die H_1 angenommen, und die WH hat sich (vorläufig) bewährt.

(3) Der empirische Wert der Teststatistik fällt zwar in den Rejektionsbereich, aber EES ist kleiner als EEM. Nach den Entscheidungsregeln des Signifikanztests muß auch in diesem Fall die H_1 akzeptiert werden. Allerdings kann die bei der Prüfung gerichteter Hypothesen ausgesprochene Empfehlung, weitere Tests mit einem entsprechend verringerten Wert für EEM durchzuführen, in dieser Situation nicht wiederholt werden, weil dadurch der für H_0 sprechende Parameterbereich zwangsläufig ebenfalls verringert würde. Eine derartige Entscheidung kann vom Experimentator nur fallspezifisch und gut begründet getroffen werden. Treten nun in weiteren Hypothesenprüfungen überwiegend experimentelle Effekte auf, die zu signifikanten Abweichungen von der H_0 führen, aber kleiner als EEM sind, stellt sich die Frage nach der Interpretation dieser Befunde. Wie im Falle der einseitigen Tests kann diese Frage auch hier nicht nach allgemeinen Kriterien beantwortet werden: Man kann sich entscheiden, die wissenschaftliche Hypothese mit der Einschränkung als bewährt anzusehen, daß sie nur (relativ) geringen statistischen Assoziationen entspricht, oder aber diese (zu) geringen Effekte als Basis für eine Falsifikation akzeptieren.

Prinzipiell die gleichen Überlegungen wie zu zweiseitigen Tests gelten auch bei Prüfungen statistischer Hypothesen über Verfahren, bei denen die Nullhypothese zwar einseitig getestet wird, mit ihrer Gültigkeit aber auch nur ein einziger Wert vereinbar ist. Beim (einseitigen) Test der (ungerichteten) Nullhypothese der Varianzanalyse über die F-Verteilung beispielsweise ist dieser Wert $F = df_N / (df_N - 2)$ (Hays, 1977, 445). Zu Einzelheiten siehe Westermann & Hager (1982).

Insgesamt soll eine Entscheidung über die Bewährung oder Falsifikation einer wissenschaftlichen Hypothese aufgrund der Prüfung einer statistischen Hypothese also nie allein auf der Signifikanz oder Nicht-Signifikanz des Stichprobenergebnisses beruhen, sondern stets auch auf dem erhaltenen experimentellen Effekt, der ja ein Maß für die Stärke der statistischen Assoziation zwischen UV(n) und AV(n) darstellt. Eine Angabe über die Größe des EES sollte daher in Forschungsberichten *stets* enthalten sein, und zwar aus den im Abschnitt 9.5 genannten Gründen sowohl in absoluten wie in relativen oder normierten Einheiten. Diese Angabe ist besonders wichtig, wenn der Experimentator seine wissenschaftliche Hypothese beibehalten hat, obwohl die empirisch festgestellten statistischen Assoziationen nur gering waren.

Die routinemäßige Angabe der aufgedeckten experimentellen Effekte würde es auch erleichtern, die Ergebnisse mehrerer Untersuchungen zur Prüfung der gleichen Hypothese zusammenzufassen: Man könnte einen mittleren Wert für die statistische Assoziation berechnen, die empirisch zwischen den durch die Hypothese in Zusammenhang gebrachten Variablen bestehen (Bredenkamp, 1972, 83; Glass, 1978; Pillemer & Light, 1980; s.a. Abschn. 1.3). Des weiteren könnten nach unterschiedlichen Kriterien (Art der Operationalisierung, Population, Situation usw.) gebildete Teilmengen der vorliegenden Untersuchungen auf diese Weise zusammengefaßt werden. Auf dieser breiten Basis wäre dann zu entscheiden, ob die wissenschaftliche Kausalhypothese sich bewährt hat, ob sie falsifiziert werden sollte oder ob ihr Geltungsbereich einzuschränken ist.

Literatur

- Abrahams, N. M. & Alf, E. F. Jr. 1978. Relative costs and statistical power in the extreme groups approach. *psychometrika*, 43, 11-17.
- Acock, A. C. & Stavig, G. R. 1979. A measure of association for nonparametric statistics. *Social Forces*, 57, 1381-1386.
- Adam, J. & Enke, H. 1972. Analyse mehrdimensionaler Kontingenztafeln mit Hilfe des Informationsmaßes von Kullback. *Biometrische Zeitschrift*, 14, 305-323.
- Adams, E. R., Fagot, R. F. & Robinson, R. E. 1965. A theory of appropriate statistics. *Psychometrika*, 30, 99-127.
- Alf, E. F. Jr. & Abrahams, N. M. 1972. Comment on component randomization tests. *Psychological Bulletin*, 77, 223-224.
- Alf, E. F. Jr. & Abrahams, N. M. 1973. Reply to Edgington. *Psychological Bulletin*, 80, 86-87.
- Alimena, B. S. 1962. A method of determining unbiased distribution in the latin square. *Psychometrika*, 27, 315-317.
- Allerbeck, K. R. 1978. Meßniveau und Analyseverfahren: Das Problem „strittiger“ Intervallskalen. *Zeitschrift für Soziologie*, 7, 199-214.
- Amir, Y. 1969. Contact hypothesis in ethnic relations. *Psychological Bulletin*, 71, 319-342.
- Anderson, N. H. 1961. Scales and statistics: Parametric and non-parametric. *Psychological Bulletin*, 58, 305-316.
- Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. & Tukey, J. W.: 1972. Robust estimates of location: Survey and advances. Princeton: Princeton University Press.
- Anscombe, F. J. & Tukey, J. W. 1963. The examination and analysis of residuals. *Technometrics*, 5, 141-160.

- Armitage, P. & Remington, R. D. 1970. Experimental design. In: McArthur, J. W. & Colton, Th. (Eds): Statistics in endocrinology. Cambridge, Mass.: The MIT Press, 3-31.
- Arnold, W. (Ed.) 1972'. Psychologisches Praktikum. Band 1. Stuttgart: Fischer.
- Aronson, E. & Carlsmith, J. M. 1968². Experimentation in socialpsychology. In: Lindzey, G. & Aronson, E. (Eds): Handbook of social psychology. Vol. II. Reading, Mass.: Addison-Wesley, 1-79.
- Aspin, A. A. 1948. An examination and further development of a formula occuring in the problem of comparing two mean values. Biometrika, 35, 88-96.
- Aspin, A. A. 1949. Tables for use in comparisons whose accuracy involves two variances separately estimated. Biometrika, 36, 290-296.
- Auslitz, K., Hesse, H. G. & Rieder, A. 1975. Möglichkeiten und Probleme der Anwendung des Allgemeinen Linearen Modells auf Fragestellungen der Varianz-, Regressions- und Kovarianzanalyse. Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung und Gesellschaft zur Förderung Pädagogischer Forschung, Mitteilungen und Nachrichten, 79/80, 1-78.
- Bailey, D. E. 1971. Probability and Statistics. Modells for Research. New York: Wiley.
- Bakan, D. 1966. The test of significance in psychological research. Psychological Bulletin, 66, 423-437.
- Baker, B. O., Hardyck, C. D. & Petrionovich, L. F. 1966. Weak measurements vs. strong statistics: An empirical critique to S. S. Stevens proscription on statistics. Educational and Psychological Measurement, 26, 291-309.
- Baker, F. B. & Collier, R. O. Jr. 1966. Some empirical results on variance ratios under permutation in the completely randomized design. Journal of the American Statistical Association, 63, 813-820.
- Baker, F. B. & Collier, R. O. Jr. 1968. An empirical study into factors affecting the F-test under Permutation for the randomized block design. Journal of the American Statistical Association, 63, 902-911.
- Barber, T. X. 1976. Pitfalls in human research. New York: Pergamon.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D. 1972. Statistical inference under order restrictions. New York: Wiley.
- Barnett, V. 1973. Comparative statistical inference. London: Wiley.
- Barnett, V. & Lewis, T. 1978. Outliers in statistical data. Chichester: Wiley.
- Bartenwerfer, H. & Raatz, U. 1979. Einführung in die Psychologie, Bd. 6: Methoden der Psychologie. Wiesbaden: Akademische Verlagsgesellschaft.
- Bartholomew, D. J. 1959 (a). A test of homogeneity for ordered alternatives. Biometrika, 46, 36-48.
- Bartholomew, D. J. 1959 (b). A test of homogeneity for ordered alternatives II. Biometrika, 46, 328-335.
- Bartholomew, D. J. 1961 (a). Ordered tests in the analysis of variance. Biometrika, 48, 325-332.

- Bartholomew, D. J. 1961 (b). A test of homogeneity of means under restricted alternatives. *Journal of the Royal Statistical Society, Series B*, 23, 239-281.
- Bartlett, M. S. 1937. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society, Series A*, 160, 268-282.
- Bartlett, M. S. 1939. A note on tests of significance in multivariate analysis. *Proceedings of the Cambridge Philosophical Society*, 35, 180-185.
- Bartlett, M. S. 1947. The use of transformations. *Biometrics*, 3, 39-52.
- Beauchamp, K. B. & May, R. B. 1964. Replication report: Interpretation of levels of significance by psychological researchers. *Psychological Reports*, 14, 272.
- Bauknecht, K. Kohlas, J. & Zehnder, C. A. 1976. *Simulationstechnik*. Berlin: Springer.
- Behnken, D. W. & Draper, N. R. 1972. Residuals and their variance Patterns. *Technometrics*, 14, 101-111.
- Behrens, W. V. 1929. Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher*, 68, 807-837.
- Bern, D.J. 1972. Self-perception theory. In: Berkowitz, L. (Ed.): *Advances in experimental social psychology*. Vol. 6. New York: Academic Press, 1-66.
- Bern, D. J. 1975. *Meinungen, Einstellungen, Vorurteile*. Köln: Bensinger & Sauerländer.
- Benninghaus, H. 1974, 1979³. *Deskriptive Statistik*. Stuttgart: Teubner.
- Berchtold, H. 1979. A modified Mann-Whitney test with improved asymptotic relative efficiency. *Biometrical Journal*, 21, 649-655.
- Berenson, M. L. 1978. A simple distribution-free test for trend in one-way layouts. *Educational and Psychological Measurement*, 38, 905-912.
- Berg, J. A. (Ed.) 1967. *Response set in personality measurement*. Chicago: Aldine.
- Bernhardson, C. S. 1975. Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics*, 31, 229-232.
- Betz, M. A. & Gabriel, K. R. 1978. Type IV errors and analysis of simple effects. *Journal of Educational Statistics*, 3, 121-143.
- Bevan, M. F., Benton, J. Q. & Myers, J. L. 1974. The robustness of the F-test to violations of continuity and form of treatment population. *British Journal of Mathematical and Statistical Psychology*, 27, 199-204.
- Bickel, P. J. 1976. Another look at robustness: A review of reviews and some developments. *Scandinavian Journal of Statistics*, 3, 145-168.
- Bishop, T. A. 1978. A Stein two-sample procedure for the general model with unequal variances. *Communications in Statistics, A* 7(5), 495-507.
- Blair, R. C. & Higgins, J. J. 1978. Tests of hypotheses for unbalanced factorial designs under various regression/coding method combinations. *Educational and Psychological Measurement*, 38, 621-631.

- Blair, R. C., Higgins, J. J. & Smitley, W. D. S. 1980. On the relative power of the U and t test. *British Journal of Mathematical and Statistical Psychology*, 33, 114-120.
- Bliss, C. J. 1967, 1970. *Statistics in biology*. Vol. 1: 1967. Vol. 2: 1970. New York: McGraw-Hill.
- Bock, R. D. 1963, 1967. Multivariate analysis of variance of repeated measurements. In: Harris, C. W. (Ed.): *Problems in measuring change*. Madison: The University of Wisconsin Press, 85-103.
- Bock, R. D. 1975. *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- Bock, R. D. & Haggard, E. A. 1968. The use of multivariate analysis of variance in behavioral research. In: Whitla, D. K. (Ed.): *Handbook of measurement and assessment in behavioral sciences*. Reading, Mass.: Addison-Wesley, 100-142.
- Boersma, F. D., de Jonge, J. J. & Stellwagen, W. R. 1964. A power comparison of the F and L tests - I. *Psychological Review*, 71, 505-513.
- Boik, R. J. 1979. The rationale of Scheffé's method and the simultaneous test procedure. *Educational and Psychological Measurement*, 39, 49-56.
- Bolles, R. C. & Messick, S. 1958. Statistical Utility in experimental inference. *Psychological Reports*, 4, 223-227.
- Boneau, C. A. 1962. A comparison of the power of the U and t tests. *Psychological Review*, 69, 246-256.
- Borich, G. O. & Godbout, R. C. 1974. Extreme groups design and calculation of statistical power. *Educational and Psychological Measurement*, 34, 663-675.
- Bortz, J. 1977, 1979². *Lehrbuch der Statistik für Sozialwissenschaftler*. Berlin: Springer.
- Bortz, J., Österreich, R. & Vogelbusch, W. 1979. Die Ermittlung optimaler Stichprobenumfänge für die Durchführung von Binomial-Tests. *Archiv für Psychologie*, 131, 267-292.
- Bowman, K. O., Beauchamp, J. J. & Shenton, L. R. 1977. The distribution of the t-statistic under non-normality. *International Statistical Review*, 45, 233-242, 256.
- Box, G. E. P. 1950. Problems in the analysis of growth and wear curves. *Biometrics*, 6, 362-389.
- Box, G. E. P. 1953. Non-normality and tests on variances. *Biometrika*, 40, 318-335.
- Box, G. E. P. 1954 (a). Some theorems on quadratic forms applied in the study of analysis of variance problems. I: Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 25, 290-302.
- Box, G. E. P. 1954 (b). Some theorems on quadratic forms applied in the study of analysis of variance problems. II: Effect of inequality of variance and correlation of errors in the two-way-classification. *Annals of Mathematical Statistics*, 25, 484-498.

- Box, G. E. P. 1968. Experimental design: Response surfaces. In: Sills, D. L. (Ed.): International Encyclopedia of the Social Sciences. Vol. 5. Boston: Crowell Collier and Macmillan, 259-259.
- Box, G. E. P. & Cox, D. R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211-252.
- Box, G. E. P., Hunter, W. G. Jr. & Hunter, J. S. 1978. *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York: Wiley-Interscience.
- Box, G. E. P. & Jenkins, G. M. 1970, 1976². *Time series analysis, forecasting and control*. San Francisco: Holden-Day.
- Box, G. E. P. & Tiao, G. C. 1973. *Bayesian inference in statistical analysis*. Reading, Mass.: Addison-Wesley.
- Bozarth, J. D. & Roberts, R. R. Jr. 1972. Signifying significant significance. *American Psychologist*, 27, 774-775.
- Bracht, G. H. 1970. Experimental factors related to aptitude-treatment interactions. *Review of Educational Research*, 40, 627-645.
- Bracht, G. H. & Glass, G. V. 1968. The external validity of experiments. *American Educational Research Journal*, 5, 437-474.
- Bradley, J. V. 1968. *Distribution-free statistical tests*. Englewood Cliffs, N. J.: Prentice-Hall.
- Bradley, J. V. 1972. Nonparametric statistics. In: Kirk, R. E. (Ed.): *Statistical issues*. Belmont, Cal.: Wadsworth, 329-338.
- Bradley, J. V. 1977. A common Situation conducive to bizarre distribution shapes. *The American Statistician*, 31, 147-150.
- Bradley, J. V. 1978. Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Bradley, J. V. 1980 (a). Nonrobustness in one-sample Z and t tests: a large scale sampling study. *Bulletin of the Psychonomic Society*, 15, 29-32.
- Bradley, J. V. 1980 (b). Nonrobustness in classical tests on means and variances: A large-scale sampling study. *Bulletin of the Psychonomic Society*, 15, 275-278.
- Bradley, J. V. 1980 (c). Nonrobustness in Z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society*, 16, 333-336.
- Brand, M. 1976. *The nature of causation*. Urbana: University of Illinois Press.
- Bredenkamp, J. 1968. F-Tests zur Prüfung von Trends und Trendunterschieden. *Zeitschrift für experimentelle und angewandte Psychologie*, 15, 239-272.
- Bredenkamp, J. 1969 (a). Experiment und Feldexperiment. In: Graumann, C. F. (Hrsg.): *Handbuch der Psychologie*. Bd. 7: Sozialpsychologie. I. Halbband: Theorien und Methoden. Göttingen: Hogrefe, 332-374.
- Bredenkamp, J. 1969 (b). über die Anwendung von Signifikanztests bei Theorie-testenden Experimenten. *Psychologische Beiträge*, 11, 275-285.

- Bredenkamp, J. 1970. Über Maße der praktischen Signifikanz. *Zeitschrift für Psychologie*, 177, 310-318.
- Bredenkamp, J. 1971. Bemerkungen zu den Trendanalysen nach Ferguson. *Zeitschrift für experimentelle und angewandte Psychologie*, 18, 386-391.
- Bredenkamp, J. 1972. *Der Signifikanztest in der psychologischen Forschung*. Frankfurt am Main: Akademische Verlagsgesellschaft.
- Bredenkamp, J. 1974. Nonparametrische Prüfung von Wechselwirkungen. *Psychologische Beiträge*, 16, 398-416.
- Bredenkamp, J. 1975. Varianzanalytische und regressionsanalytische Verfahren in der Curriculumevaluation. In: Frey, K. (Hrsg.): *Curriculum-Handbuch*. München: Piper, 786-822.
- Bredenkamp, J. 1979. Das Problem der externen Validität pädagogisch-psychologischer Untersuchungen. In: Brandstätter, J., Reinert, G. & Schneewind, K. A. (Hrsg.): *Pädagogische Psychologie. Probleme und Perspektiven*. Stuttgart: Klett-Cotta, 267-289.
- Bredenkamp, J. 1980. *Theorie und Planung psychologischer Experimente*. Darmstadt: Steinkopff.
- Bredenkampf, J. & Feger, H. 1970. Kriterien für die Entscheidung über die Aufnahme empirischer Arbeiten in die *Zeitschrift für Sozialpsychologie*. *Zeitschrift für Sozialpsychologie*, 1, 43-47.
- Bredenkamp, J. & Hager, W. 1979. Experimentelle Befunde zur modifizierten Invarianzhypothese und zur Hypothese einer konstanten Langzeitgedächtnisspanne. *Psychologische Beiträge*, 21, 382-400.
- Breen, L. & Gaito, J. 1970. Comments on Friedman's r_m procedure. *Psychological Bulletin*, 73, 309-310.
- Brewer, J. K. 1972. On the power of statistical tests in the American Educational Research Journal. *American Educational Research Journal*, 9, 391-401.
- Bridgman, P. W. 1927. *The logic of modern physics*. New York: Macmillan.
- Brillinger, D. R. 1975. *Time series, data analysis and theory*. New York: Holt, Rinehart & Winston.
- Brocke, B. 1978. *Technologische Prognosen*. Freiburg: Alber.
- Brocke, B. 1979. Aspekte einer Methodologie der angewandten Sozial- und Verhaltenswissenschaften. *Zeitschrift für Sozialpsychologie*, 10, 2-29.
- Broekmann, N. C. 1973. Note on the influence of an exceptional subject. *Psychometrika*, 38, 611-613.
- Brown, D. J. 1975. Down with the linear model. *American Educational Research Journal*, 491-505.
- Brown, M. B. & Forsythe, A. B. 1974 (a). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16, 129-132.
- Brown, M. B. & Forsythe, A. B. 1974 (b). The ANOVA and multiple comparisons for the data with heterogeneous variances. *Biometrics*, 30, 719-724.

- Brown, M. B. & Forsythe, A. B. 1974 (c). Robust tests for equality of variances. *Journal of the American Statistical Association*, 69, 364-367.
- Bünning, H. & Trenkler, G. 1978. *Nichtparametrische statistische Methoden*. Berlin: de Gruyter.
- Bugelski, B. R. 1960. *A first course in experimental psychology*. New York: Holt.
- Bungard, W. (Hrsg.) 1980. *Die „gute“ Versuchsperson denkt nicht*. München: Urban und Schwarzenberg.
- Bungard, W. & Lück, H. E. 1974. *Forschungsartefakte und nicht-reaktive Meßverfahren*. Stuttgart: Teubner.
- Bunge, M. 1967. (a, b). *Scientific Research*. I. The search for system. 1967 (a). II. The search for truth. 1967 (b). New York: Springer.
- Burnett, T. D. & Barr, D. R. 1977. A non-metric analogy of analysis of covariance. *Educational and Psychological Measurement*, 37, 341-348.
- Busemeyer, J. R. 1980. Importance of measurement theory, error theory, and experimental design for testing the significance of interactions. *Psychological Bulletin*, 88, 237-244.
- Callaway, J. N., Nowicki, S. & Duke, M. P. 1980. Overt expression of experimental expectancies, interaction with subject expectancies and Performance in a psychomotor task. *Journal of Research in Personality*, 14, 27-39.
- Campbell, D. T. 1957. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297-312.
- Campbell, D. T. 1969. Prospective: Artifact and control. In: Rosenthal, R. L. & Rosnow, R. C. (Eds): *Artifact in behavioral research*. New York: Academic Press, 351-382.
- Campbell, D. T. & Fiske, D. W. 1959. Convergent and discriminant Validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Campbell, D. T. & Stanley, J. C. 1963, 1973. Experimental and quasi-experimental designs for research in teaching. In: Gage, N. L. (Ed.): *Handbook of research in teaching*. Chicago: Rand McNally, Kap. 5. Deutsche Bearbeitung in: Ingenkamp, K. (Hrsg.): *Strategien der Unterrichtsforschung*. Weinheim: Beltz, 101-193.
- Carlsmith, J. M., Ellsworth, P. C. & Aronson, E. 1976. *Methods of research in social psychology*. Reading, Mass. : Addison-Wesley.
- Carlson, R. 1976. Discussion: The logic of tests of significance. *Philosophy of Science*, 43, 116-128.
- Carmer, S. G. & Swanson, M. R. 1971. Detection of differences between means: A Monte Carlo study of five pairwise multiple comparison procedures. *Agronomy Journal*, 63, 940-945.
- Carmer, S. G. & Swanson, M. R. 1973. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association*, 68, 66-74.

- Carnap, R. 1956, 1960. The methodological Character of theoretical concepts. In: Feigl, H. & Scriven, M. (Eds): Minnesota studies in the philosophy of science. Vol. I. Minneapolis: University of Minnesota Press, 33-76. Deutsche Übersetzung: Theoretische Begriffe der Wissenschaft: Eine logische und methodologische Untersuchung. Zeitschrift für philosophische Forschung, 14, 209-233 und 571-598.
- Carroll, R. M. & Nordholm, L. A. 1975. Sampling characteristics of Kelley's ϵ and Hays' ω^2 . Educational and Psychological Measurement, 35, 541-554.
- Carter, D. S. 1979. Comparison of different shrinkage formulars in estimating population multiple correlation coefficients. Educational and Psychological Measurement, 39, 261-266.
- Carver, R. P. 1968. Note on a schema for proper utilization of multiple comparisons in research. American Educational Research Journal, 5, 730-732.
- Carver, R. P. 1978. The case against statistical significance testing. Harvard Educational Review, 48, 378-399.
- Cascio, W. F., Valenzi, E. R., Silbey, V. 1978. Validation and statistical power: Implications for applied research. Journal of Applied Psychology, 63, 589-595.
- Cascio, W. F., Valenzi, E. R. & Silbey, V. 1980. More on Validation and statistical power. Journal of Applied Psychology, 65, 135-138.
- Chase, L. J. & Baran, S. J. 1976. An assessment of quantitative research in mass communication. Journalism Quarterly, 53, 308-311.
- Chase, L.J. & Chase, R. B. 1976. A statistical power analysis of applied psychological research. Journal of Applied Psychology, 61, 234-237.
- Chase, L. J. & Tucker, R. K. 1975. A power-analytic examination of contemporary communication research. Speech Monographs, 42, 29-41.
- Chase, L. J. & Tucker, R. K. 1976. Statistical power: Derivation, development and data-analytic implications. Psychological Record, 26, 473-486.
- Chatfield, C. 1975. The analysis of time series: Theory and practice. London: Chapman & Hall.
- Church, J. D. & Wike, E. L. 1976. The robustness of homogeneity of variance tests for asymmetric distributions: A Monte Carlo study. Bulletin of the Psychonomic Society, 7, 416-420.
- Church, J. D. & Wike, E. L. 1979. A Monte Carlo study of nonparametric multiple-comparison tests for a two-way layout. Bulletin of the Psychonomic Society, 14, 95-98.
- Church, J. D. & Wike, E. L. 1980. Two Monte Carlo studies of Silverstein's nonparametric multiple comparison tests. Psychological Reports, 46, 403-407.
- Clauß, G. & Ebner, H. 1978⁶. Grundlagen der Statistik. Berlin: Volk und Wissen.
- Cleary, T. A. & Linn, R. L. 1969. Error of measurement and the power of a statistical test. British Journal of Mathematical and Statistical Psychology, 22, 49-55.

- Cleary, T. A., Linn, R. L. & Walster, G. W. 1970. Effect of reliability and validity on power of statistical tests. In: Borgatta, E. F. & Bohrnstedt, G. W. (Eds): *Sociological methodology*. 1970. San Francisco: Jossey-Bass, 130-138.
- Cochran, W. G. 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 3, 22-38.
- Cochran, W. G. 1951. Testing a linear relation among variances. *Biometrics*, 7, 17-32.
- Cochran, W. G. 1957. Analysis of covariance: Its nature and uses. *Biometrics*, 13, 261-281.
- Cochran, W. G. 1968 (a). The design of experiments. In: Sills, D. L. (Ed.): *International Encyclopedia of the Social Sciences*. Vol. 5. Boston: Crowell Collier and Macmillan, 245-254.
- Cochran, W. G. 1968 (b). Errors of measurement in statistics. *Technometrics*, 10, 637-666.
- Cochran, W. G. 1970. Some effects of errors of measurement on multiple correlation. *Journal of the American Statistical Association*, 65, 22-34.
- Cochran, W. G. 1977³, 1972. *Sampling Techniques*. New York: John Wiley. Deutsch: Stichprobenverfahren. Berlin: De Gruyter.
- Cochran, W. G. & Bliss, C. I. 1970. Analysis of variance. In: McArthur, J. W. & Colton, Th. (Eds): *Statistics in endocrinology*. Cambridge, Mass.: The MIT Press, 33-78.
- Cochran, W. G. & Cox, G. M. 1957², 1968³. *Experimental designs*. New York: Wiley.
- Cohen, J. 1962. The statistical power of abnormal - social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.
- Cohen, J. 1965. Some statistical issues in psychological research. In: Wolman, B. B. (Ed.): *Handbook of clinical psychology*. New York: MacGraw-Hill, 95-121.
- Cohen, J. 1968. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70, 426-443.
- Cohen, J. 1969, 1977². *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Cohen, J. 1970. Approximate power and sample size determination for common one-sample and two-sample hypothesis tests. *Educational and Psychological Measurement*, 30, 811-831.
- Cohen, J. 1973 (a). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, 33, 107-112.
- Cohen, J. 1973 (b). Statistical power analysis and research results. *American Educational Research Journal*, 10, 225-230.
- Cohen, J. 1980. Trend analysis the easy way. *Educational and Psychological Measurement*, 40, 565-568.
- Cohen, J. & Cohen, P. 1975. *Applied multiple regression/correlation analysis for the Behavioral Sciences*. Hillsdale, N. J.: Lawrence Erlbaum.

- Collier, R. O. Jr. & Baker, F. B. 1966. Some Monte Carlo results on the power of the F-test under permutation in the simple randomized block design. *Biometrika*, 53, 199-203.
- Collier, R. O. Jr., Baker, F. B., Mandeville, G. K. & Hayes, T. F. 1967. Estimates of test size for several test procedures based on conventional variance ratio in the repeated measures design. *Psychometrika*, 32, 339-353.
- Collier, R. O. Jr. & Larson, R. C. 1969. Monte Carlo methods. *Review of Educational Research*, 39, 734-739.
- Cook, S. W. & Selltitz, C. 1964. A multiple-indicator approach to attitude measurement. *Psychological Bulletin*, 62, 36-55.
- Cook, T. D. & Campbell, D. T. 1976. The design and conduct of quasi-experiments and true experiments in field settings. In: Dunnette, M. D. (Ed.): *Handbook of industrial and organizational psychology*. Chicago: Rand McNally, 223-326.
- Cook, T. D. & Campbell, D. T. 1979. *Quasi-experimentation. Design and analysis issues for field settings*. Chicago: Rand McNally.
- Cooley, W. W. & Lohnes, P. R. 1971. *Multivariate data analysis*. New York: Wiley.
- Coombs, C. H., Dawes, R. M. & Tversky, A. 1975. *Mathematische Psychologie*. Weinheim: Beltz.
- Cooper, H. M. & Rosenthal, R. 1980. Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 87, 442-449.
- Cowles, M. P. 1974. N = 35: A rule of thumb for psychological researchers. *Perceptual and Motor Skills*, 38, 1135-1138.
- Cox, D. R. 1957. The use of a concomitant variable in selecting an experimental design. *Biometrika*, 44, 150-158.
- Cox, D. R. 1958. *Planning of experiments*. New York: Wiley.
- Cox, D. R. 1961. The design of experiments: The control of error. *Journal of the Royal Statistical Society, Series A*, 124, 44-48.
- Craig, J. R., Eison, C. L., Metzger, L. P. 1976. Significance tests and their Interpretation: An example utilizing published research and ω^2 . *Bulletin of the Psychonomic Society*, 7, 280-282.
- Cramer, E. M. & Appelbaum, M. I. 1980. Nonorthogonal Analysis of Variance - once again. *Psychological Bulletin*, 87(1), 51-57.
- Cramer, E. M. & Bock, R. R. 1966. Multivariate analysis. *Review of Educational Research*, 36, 604-617.
- Cramer, E. M. & Nicewander, W. A. 1979. Some symmetric, invariant measures of multivariate association. *Psychometrika*, 44, 43-54.
- Cronbach, L. J. 1957. The two disciplines of scientific psychology. *American Psychologist*, 12, 671-689.
- Cronbach, L. J. 1975. Beyond the two disciplines of scientific psychology. *American Psychologist*, 30, 116-127.

- Cronbach, L. J. & Furby, L. 1970. How should we measure „change“ - or should we? *Psychological Bulletin*, 74, 68-80.
- Cronbach, L. J. & Meehl, P. E. 1955. Construct Validation in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Cureton, E. E. 1966. On correlation coefficients. *Psychometrika*, 31, 605-607.
- D'Agostino, R. B. 1972. Relation between the chi-squared and ANOVA tests for testing the equality of k independent dichotomous populations. *American Statistician*, 26(3), 30-32.
- Darlington, R. B. 1968. Multiple regression in psychological research and practice. *Psychological Bulletin*, 69, 161-182.
- David, H. A., Lachenbruch, P. A. & Brandis, H. P. 1972. The power function of range and studentized range tests in normal samples. *Biometrika*, 59, 161-168.
- Davidson, M. L. 1972. Univariate versus multivariate tests in repeated measures experiments. *Psychological Bulletin*, 77, 446-452.
- David, D. J. 1969. Flexibility and power in comparisons among means. *Psychological Bulletin*, 71, 441-444.
- Dawes, R. M. 1977. *Grundlagen der Einstellungsmessung*. Weinheim: Beltz.
- Dayton, C. M. & Schafer, W. D. 1973. Extended tables of t and Chi Square for Bonferroni tests with unequal error allocation. *Journal of the American Statistical Association*, 68, 78-83.
- Dénes, J. & Keedwell, A. D. 1974. *Latin squares and their applications*. Budapest: Akadémiai Kiado.
- Derrick, T. 1976. The criticism of inferential statistics. *Educational Research*, 19(1), 35-40.
- Diehl, J. M. 1977, 1979³. *Varianzanalyse*. Frankfurt am Main: Fachbuchhandlung für Psychologie.
- Dierkes, M. 1977. Die Analyse von Zeitreihen und Longitudinalstudien. In: Koolwijk, J. van & Wieken-Mayser, M. (Hrsg.): *Techniken der empirischen Sozialforschung*. Bd. 7. München: Oldenbourg, 111-169.
- Digman, J. M. 1966. Interaction and nonlinearity in multivariate experiment. In: Cattell, R. B. (Ed.): *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 459-475.
- Dipboye, R. L. & Flanagan, M. F. 1979. Research settings in industrial and organizational psychology. Are findings in the field more generalizable than in the laboratory? *American Psychologist*, 34, 141-150.
- Dixon, W. J. & Massey, F. J. Jr. 1951, 1957², 1969³. *Introduction to statistical analysis*. New York: McGraw-Hill.
- Dixon, W. J. & Tukey, J. W. 1968. Approximate behavior of the distribution of Winsorized t (Trimming/Winsorization II). *Technometrics*, 10, 83-98.

- Dodd, D. H. & Schultz, R. F. Jr. 1973. Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin*, 79, 391-395.
- Dooling, D. J. & Danks, J. H. 1975. Going beyond tests of significance: Is psychology ready? *Bulletin of the Psychonomic Society*, 5, 15-17.
- Draper, N. R. & Hunter, W. G. 1969. Transformations: Some examples revisited. *Technometrics*, 11, 23-40.
- Draper, N. R. & Smith, H. 1966. *Applied regression Analysis*. New York: Wiley.
- Dunn, O. J. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56, 52-64.
- Dunn, O. J. & Massey, F. J. 1965. Estimation of multiple contrasts using t-distributions. *Journal of the American Statistical Association*, 60, 573-583.
- Dunnett, C. W. 1970. Multiple comparisons. In: McArthur, J. W. & Colton, Th. (Eds.): *Statistics in endocrinology*. Cambridge, Mass.: The MIT Press, 79-103.
- Dwyer, J. H. 1974. Analysis of variance and the magnitude of effects: A general Approach. *Psychological Bulletin*, 81, 731-737.
- Edgington, E. S. 1964(a). A tabulation of inferential statistics used in psychology journals. *American Psychologist*, 19, 202-203.
- Edgington, E. S. 1964(b). Randomization test. *Journal of Psychology*, 57, 445-449.
- Edgington, E. S. 1965. The assumption of homogeneity of variance for the t-test and nonparametric tests. *Journal of Psychology*, 59, 177-179.
- Edgington, E. S. 1966. Statistical inference and nonrandom samples. *Psychological Bulletin*, 66, 485-487.
- Edgington, E. S. 1969(a). Approximate randomization tests. *Journal of Psychology*, 72, 143-149.
- Edgington, E. S. 1969(b). Statistical inference: The distribution-free approach. New York: McGraw-Hill.
- Edgington, E. S. 1973. The random-sampling assumption in: „Comment on component-randomization tests“. *Psychological Bulletin*, 80, 84-85.
- Edgington, E. S. 1974. A new tabulation of statistical procedures used in APA journals. *American Psychologist*, 29, 25-26.
- Edwards, A. L. 1950, 1960², 1968³, 1971. *Experimental design in psychological research*. New York: Holt, Rinehart & Winston. *Versuchsplanung in der psychologischen Forschung*. Weinheim: Beltz. Frankfurt am Main: Fachbuchhandlung für Psychologie.
- Edwards, A. L. 1957(a). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.
- Edwards, A. L. 1957(b). *The social desirability variable in personality assessment and research*. New York: Holt, Rinehart & Winston.
- Edwards, A. L. 1970. *The measurement of personality traits by scales and inventories*. New York: Holt. Rinehart & Winston.

- Edwards, A. W. F. 1972. Likelihood. Cambridge: Cambridge University Press.
- Edwards, W., Lindman, H. & Savage, L. J. 1963. Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Eimer, E. 1978. Varianzanalyse. Stuttgart: Kohlhammer.
- Eisenhart, C. 1947. The Assumptions underlying the analysis of variance. *Biometrics*, 3, 1-21.
- Einot, I. & Gabriel, K. R. 1975. A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, 70, 574-583.
- Ekbohm, G. 1976. Testing the equality of several means with small samples. *Biometrische Zeitschrift*, 18, 547-553.
- Elashoff, J. D. 1969. Analysis of covariance: A delicate instrument. *American Educational Research Journal*, 6, 383-401.
- Elashoff, R. M. 1968. Errors: Effects of errors in statistical assumptions. In: Sills, D. L. (Ed.): *International Encyclopedia of the Social sciences*. Vol. 5. Boston: Crowell Collier & Macmillan, 132-142.
- Elston, R. C. 1961. On additivity in the analysis of variance. *Biometrics*, 17, 209-219.
- Elston, R. C. & Bush, N. 1964. The hypotheses that can be tested when there are interactions in the analysis of variance model. *Biometrics*, 20, 681-698.
- Enderlein, G. (Hrsg.) (Autorenkollektiv). 1972. *Biometrische Versuchsplanung*. Berlin: VEB Deutscher Landwirtschaftsverlag.
- Engelhardt, W. 1979. Nonparametrische Prüfung von Wechselwirkungen: Verteilungsfunktionen der Interaktionsstatistik. *Psychologische Beiträge*, 21, 223-236.
- Erlebacher, A. 1977. Design and analysis of experiments contrasting the within- and between-subjects manipulation of the independent variable. *Psychological Bulletin*, 84, 212-219.
- Erlebacher, A. 1978. The analysis of multifactor experiments designed to contrast the within- and between-subjects manipulation of the independent variable. *Behavior Research Methods and Instrumentation*, 10, 833-840.
- Evans, S. H. & Anastasio, E. J. 1968. Misuse of covariance when treatment effect and covariate are confounded. *Psychological Bulletin*, 69, 225-234.
- Everitt, B. S. 1979. A Monte Carlo investigation of the robustness of Hotelling's one- and two-sample T^2 tests. *Journal of the American Statistical Association*, 74, 48-51.
- Federer, W. T. 1955. *Experimental design*. New York: Macmillan
- Field, H. S. & Armenakis, A. A. 1974. On use of multiple tests of significance in psychological research. *Psychological Reports*, 35, 427-431.
- Feir-Walsh, B. E. & Toothaker, L. E. 1974. An empirical comparison of the ANOVA F-test, normal scores test and Kruskal Wallis test under Violation of assumptions. *Educational and Psychological Measurement*, 34, 789-799.
- Feldt, L. S. 1958. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, 23, 335-353.

- Feldt, L. S. 1961. The use of extreme groups to test for the presence of a relationship. *Psychometrika*, 26, 307-316.
- Feldt, L. S. 1973. What size samples for methods/materials experiments? *Journal of Educational Measurement*, 10, 221-226.
- Fennessey, J. 1968. The general linear model: A new perspective on some familiar topics. *American Journal of Sociology*, 74(1), 1-27.
- Ferguson, G. A. 1965. *Nonparametric trend analysis*. Montreal: McGill University Press.
- Festinger, L. 1978. *Theorie der kognitiven Dissonanz*. Bern: Huber.
- Festinger, L. & Carlsmith, J. M. 1959. Cognitive consequences of forced compliance. *Journal of abnormal and social Psychology*, 58, 203-210.
- Fietkau, H.-J. 1973. *Zur Methodologie des Experimentierens in der Psychologie*. Meisenheim: Hain.
- Finn, J. D. 1969. Multivariate analysis of repeated measures data. *Multivariate Behavioral Research*, 4, 391-413.
- Finn, J. D. 1974. *A general model for multivariate analysis*. New York: Holt, Rinehart & Winston.
- Fischer, G. 1974. *Einführung in die Theorie psychologischer Tests*. Bern: Huber.
- Fisher, R. A. 1925¹, 1950¹, 1954¹². *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
- Fisher, R. A. 1928. The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society, London, A*, 121, 654-673.
- Fisher, R. A. 1935. The fiducial argument in statistical inference. *Annals of Eugenics*, 6, 391-398.
- Fisher, R. A. 1935, 1949⁵, 1951⁶, 1953. *The design of experiments*. Edinburgh: Oliver & Boyd.
- Fisher, R. A. 1956. *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd.
- Fisher, R. A. & Mackenzie, W. A. 1923. Studies in crop Variation. II. The manurial response of different potato varieties. *Journal of Agricultural Science*, 13, 311-331.
- Fisher, R. A. & Yates, F. 1938, 1963⁶. *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver Boyd.
- Fisz, M. 1970. *Wahrscheinlichkeitsrechnung und mathematische Statistik*. Berlin: VEB Deutscher Verlag der Wissenschaften.
- Fleishman, A. I. 1980. Confidence intervals for correlation ratios. *Educational and Psychological Measurement*, 40, 659-670.
- Fleiss, J. L. 1969. Estimating the magnitude of experimental effects. *Psychological Bulletin*, 72, 273-276.
- Fleiss, J. L. 1973. *Statistical methods for rates and proportions*. New York: Wiley.

- Fleiss, J. L. 1976. Comment on Overall and Woodward's asserted paradox concerning the measurement of change. *Psychological Bulletin*, 83, 774-775.
- Forsyth, R. A. 1978(a). A Note on „Planning an experiment in the Company of measurement error“ by Levin and Subkoviak. *Applied Psychological Measurement*, 2, 377-381.
- Forsyth, R. A. 1978(b). Some additional comments on „Planning an experiment in the Company of measurement error“. *Applied Psychological Measurement*, 2, 386-387.
- French, J. R. P. 1956, 1972⁸. Feldexperimente: Änderungen in der Gruppenproduktivität. In: König, R. (Hrsg.): *Beobachtung und Experiment in der Sozialforschung*. Köln: Kiepenheuer & Witsch, 259-273.
- Frey, D. 1978. Die Theorie der kognitiven Dissonanz. In: Frey, D. (Hrsg.): *Kognitive Theorien der Sozialpsychologie*. Bern: Huber, 243-292.
- Fricke, R. 1977. Möglichkeiten zur zusammenfassenden Darstellung von unabhängigen Forschungsergebnissen zur Lehrer-Schüler-Interaktion. *Zeitschrift für erziehungswissenschaftliche Forschung*, 11, 208-215.
- Friedman, H. 1968. Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, 70, 245-251.
- Friedman, H. 1969. A simplified table for the estimation of magnitude of experimental effect. *Psychonomic Science*, 14, 193-195.
- Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675-701.
- Friedman, M. 1940. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11, 86-92.
- Fromkin, H. L. & Streufert, S. 1976. Laboratory experimentation. In: Dunnette, M. D. (Ed.): *Handbook of industrial and organizational psychology*. Chicago: Rand McNally, 415-465.
- Gabriel, K. R. 1964. A procedure for testing the homogeneity of all sets of means in analysis of variance. *Biometrics*, 20, 459-477.
- Gabriel, K. R. 1969. Simultaneous tests procedures - some theory of multiple comparisons. *Annals of Mathematical Statistics*, 40, 224-250.
- Gabriel, K. R. 1978. A simple method of multiple comparisons of means. *Journal of the American Statistical Association*, 73, 724-729.
- Gabriel, R. M. & Glavin, G. B. 1978. Multivariate data analysis in empirical research: A look on the bright side. *Pavlovian Journal of Biological Science*, 13, 93-112.
- Gabriel, R. M. & Hopkins, K. D. 1974. Relative merits of MANOVA, repeated measures ANOVA, and univariate ANOVAs for research utilizing multiple criterion measures. *Journal of Special Education*, 8, 377-389.
- Gadenne, V. 1976. *Die Gültigkeit psychologischer Untersuchungen*. Stuttgart: Kohlhammer.

- Gaebelein, J. W. & Soderquist, D. R. 1978. The Utility of within-subjects variables: Estimates of Strength. *Educational and Psychological Measurement*, 38, 351-360.
- Gaebelein, J. W., Soderquist, D. R. & Powers, W. A. 1976. A note on variance explained in the mixed analysis of variance model. *Psychological Bulletin*, 83, 1110-1112.
- Gaensslen, H. & Schubö, W. 1973, 1976². *Einfache und komplexe statistische Analyse*. München: Reinhardt.
- Gaito, J. 1958. The Bolles-Messick coefficient of Utility. *Psychological Reports*, 595-598.
- Gaito, J. 1959(a). Multiple comparisons in analysis of variance. *Psychological Bulletin*, 56, 392-393.
- Gaito, J. 1959(b). Non-parametric methods in psychological research. *Psychological Reports*, 5, 115-125.
- Gaito, J. 1960(a). Expected mean squares in analysis of variance techniques. *Psychological Reports*, 7, 3-10.
- Gaito, J. 1960(b). Scale classification and statistics. *Psychological Review*, 67, 277-278.
- Gaito, J. 1973. Repeated measurements designs and tests of null-hypothesis. *Educational and Psychological Measurement*, 33, 69-75.
- Gaito, J. 1977(a). Directional and nondirectional alternative hypotheses. *Bulletin of the psychonomic Society*, 9, 371-372.
- Gaito, J. 1977(b). Equal and unequal n and equal and unequal intervals in trend analyses. *Educational and Psychological Measurement*, 37, 283-289.
- Gaito, J. 1978. Multiple comparisons within ANOVA using orthogonal or nonorthogonal components. *Educational and Psychological Measurement*, 38, 901-904.
- Gaito, J. 1980. Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 564-567.
- Gaito, J. & Firth, J. 1973. Procedures for estimating magnitude of effects. *Journal of Psychology*, 83, 151-161.
- Gaito, J. & Turner, E. D. 1963. Error terms in trend analyses. *Psychological Bulletin*, 60, 464-474.
- Garnes, P. A. 1966. Comments on „A power comparison of the F and L tests - I“. *Psychological Review*, 73, 372-375.
- Garnes, P. A. 1971(a). Inverse relation between the risks of type I and type II errors and suggestions for the unequal n case in multiple comparisons. *Psychological Bulletin*, 97-102.
- Garnes, P. A. 1971(b). Multiple comparisons of means. *American Educational Research Journal*, 8, 531-565.
- Garnes, P. A. 1973. Type IV errors revisited. *Psychological Bulletin*, 80, 304-307.

- Garnes, P. A. 1978(a). A three-factor model encompassing many possible statistical tests on independent groups. *Psychological Bulletin*, 85, 168-182.
- Garnes, P. A. 1978(b). A four-factor structure for parametric tests on independent groups. *Psychological Bulletin*, 85, 661-672.
- Garnes, P. A. 1978(c). Nesting, crossing, type IV errors, and the role of statistical models. *American Educational Research Journal*, 15, 253-258.
- Garnes, P. A. & Howell, J. F. 1976. Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, 1, 113-125.
- Garnes, P. A., Keselman, H. K. & Clinch, J. J. 1979(a). Multiple comparisons for variance heterogeneity. *British Journal of Mathematical and Statistical Psychology*, 32, 133-142.
- Garnes, P. A., Keselman, H. K. & Clinch, J. J. 1979(b). Test for homogeneity of variance in factorial design. *Psychological Bulletin*, 86, 978-984.
- Garnes, P. A. & Lucas, P. A. 1966. Power of the analysis of variance of independent groups on non-normal and normally transformed data. *Educational and Psychological Measurement*, 26, 311-327.
- Garnes, P. A., Winkler, H. B. & Probert, D. A. 1972. Robust test for homogeneity of variance. *Educational and Psychological Measurement*, 32, 887-909.
- Gardner, P. L. 1975. Scales and statistics. *Review of Educational Research*, 45, 43-57.
- Garten, H.-K. 1980. Zur inhaltlichen Problematik von Aptitude-Treatment-Interaktion. *Zeitschrift für erziehungswissenschaftliche Forschung*, 14, 13-35.
- Gartside, P. S. 1972. A study of methods for comparing several variances. *Journal of the American Statistical Association*, 67, 342-346.
- Geary, R. C. 1936. The distribution of „Student's“ ratio for non-normal samples. *Journal of the Royal Statistical Society, Supplement*, 3, 178-189.
- Geary, R. C. 1947. Testing for normality. *Biometrika*, 34, 209-242.
- Gebert, A. 1977. Prüfung von Wechselwirkungen in 2^2 -faktoriellen Blockplänen mittels simultaner W-Tests. *Psychologische Beiträge*, 19, 121-129.
- Gebhardt, F. 1966. Verteilung und Signifikanzschranken des 3. und 4. Stichprobenmomentes bei normalverteilten Variablen. *Biometrische Zeitschrift*, 8, 219-241.
- Geisser, S. & Greenhouse, S. W. 1958. An extension of Box's results on the use of the F distribution in multivariate analysis. *The Annals of Mathematical Statistics*, 29, 885-891.
- Gibbons, J. D. 1971. *Nonparametric statistical inference*. New York: McGraw-Hill.
- Gibbons, J. D. & Pratt, J. W. 1975. P-Values: Interpretation and methodology. *The American Statistician*, 29(1), 20-25.
- Glass, G. V. 1966. Testing homogeneity of variances. *American Educational Research Journal*, 3, 187-190.
- Glass, G. V. 1976. Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3-8.

- Glass, G. V. 1978. Integrating findings: The meta-analysis of research. *Review of Research in Education*, 5, 351-379.
- Glass, G. V. & Hakstian, A. R. 1969. Measures of association in comparative experiments: Their developments and interpretation. *American Educational Research Journal*, 6, 40-14.
- Glass, G. V., Peckham, P. D. & Sanders, J. R. 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42, 237-288.
- Glass, G. V. & Stanley, J. C. 1970. *Statistical methods in education and psychology*. New Jersey: Prentice Hall.
- Glass, G. V., Willstott, V. L. & Gottman, J. M. 1975. *Design and analysis of time-series-experiments*. Boulder, Col.: University Press.
- Gniech, G. 1976. *Störeffekte in psychologischen Experimenten*. Stuttgart: Kohlhammer.
- Gokhale, D. V. & Kullback, S. 1978. *The information in contingency tables*. New York: Dekker.
- Gold, D. 1969. Statistical tests and Substantive significance. *The American Sociologist*, 4, 42-46.
- Golhar, M. B. 1972. The errors of first and second kinds in Welch-Aspin's Solution of the Behrens-Fisher problem. *Journal of Statistical Computation and Simulation*, 1, 209-224.
- Goodman, L. A. (Ed.) 1978. *Analyzing qualitative/categorical data*. Cambridge, Mass.: Abt Books.
- Gray, H. L. & Schucany, W. R. 1972. *The generalized jackknife statistic*. New York: Dekker.
- Greenhouse, S. W. & Geisser, S. 1959. On methods in the analysis of profile data. *Psychometrika*, 24, 95-112.
- Greenwald, A. G. 1975. Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1-20.
- Greenwald, A. G. 1976. Within-subjects designs: To use or not to use? *Psychological Bulletin*, 83, 314-320.
- Grizzle, J. E., Starmer, C. F. & Koch, G. 1969. Analysis of categorical data by linear models. *Biometrics*, 25, 489-504.
- Groeben, N. & Westmeyer, H. 1975. *Kriterien psychologischer Forschung*. München: Juventa.
- Gruijter, D. N. M. de & Kamp, L.J. T. van der (Eds). 1976. *Advances in psychological and educational measurement*. London: Wiley.
- Guenther, W. C. 1964. *Analysis of variance*. Englewood Cliffs, N. J.: Prentice Hall.
- Guenther, W. C. 1965, 1973². *Concepts of statistical inference*. New York: McGraw-Hill. Tokyo: McGraw-Hill Kogakuska.

- Guilford, J. P. & Fruchter, B. 1956, 1973⁵. Fundamental statistics in psychology and education. New York: McGraw-Hill. Tokyo: McGraw-Hill.
- Guttman, L. 1977. What is not what in statistics. *The Statistician*, 26, 81-107.
- Haagen, K. & Pertler, R. 1976. Methoden der Statistik für Psychologen. Band I. Stuttgart: Kohlhammer.
- Haagen, K. & Seifert, H.-G. 1979. Methoden der Statistik für Psychologen. Band II. Stuttgart: Kohlhammer.
- Haberman, S. J. 1978. Analysis of qualitative data. Vol. 1: Introductory topics. 1978. New York: Academic Press.
- Hacking, J. 1965, 1976². Logic of statistical inference. London: Cambridge University Press.
- Hager, W. Zur statistischen Prüfung von Interaktionshypothesen. In: Lürer, G. (Hrsg.): Bericht über den 33. Kongreß der DGfPs 1982 in Mainz. Göttingen: Hogrefe (im Druck, a).
- Hager, W. Über univariate parametrische Maße der wissenschaftlichen Signifikanz. *Zeitschrift für Psychologie* (im Druck, b).
- Hager, W., Lübbecke, B. & Hübner, R. Verletzung der Annahmen bei Zwei-Stichproben-Lokationstests. *Zeitschrift für experimentelle und angewandte Psychologie* (im Druck).
- Hager, W. & Westermann, R. Zur Wahl und Prüfung statistischer Hypothesen in psychologischen Untersuchungen. *Zeitschrift für experimentelle und angewandte Psychologie* (im Druck, a).
- Hager, W. & Westermann, R. Entscheidungen über statistische und wissenschaftliche Hypothesen: Probleme bei mehrfachen Signifikanztests zur Prüfung einer wissenschaftlichen Hypothese. *Zeitschrift für Sozialpsychologie* (im Druck, b).
- Hakstian, A. R., Roed, J. C. & Lind, J. C. 1979. Two Sample T^2 procedure and the assumption of homogenous covariance matrices. *Psychological Bulletin*, 86, 1255-1263.
- Halderson, J. S. & Glasnapp, D. R. 1972. Generalized rules für calculating the magnitude of an effect in factorial and repeated measures ANOVA designs. *American Educational Research Journal*, 9, 301-310.
- Hall, I. J. 1972. Some comparisons of tests for equality of variances. *Journal of Statistical Computation and Simulation*, 1, 183-194.
- Hamer, R. M. & Hosking, J. D. 1977. The nonorthogonal analysis of variance. *Representative Research in Social Psychology*, 8, 71-87.
- Hamilton, B. L. 1976. A Monte Carlo test of the robustness of parametric and non-parametric analysis of covariance against unequal regression slopes. *Journal of the American Statistical Association*, 71, 864-869.
- Hamilton, B. L. 1977. An empirical investigation of the effects of heterogenous regression slopes in analysis of covariance. *Educational and Psychological Measurement*, 37, 701-712.

- Hammersley, J. M. & Handscomb, D. C. 1964. Monte Carlo methods. New York: Wiley.
- Hampel, F. R. 1980. Robuste Schätzungen: Ein anwendungsorientierter Überblick. *Biometrical Journal*, 22, 3-21.
- Harnatt, J. 1975. Der statistische Signifikanztest in kritischer Betrachtung. *Psychologische Beiträge*, 17, 595-612.
- Harnatt, J. 1979. Anmerkungen zur Kritik psychologischer Forschungsmethoden. *Psychologische Beiträge*, 21, 203-211.
- Harris, C. W. (Ed.) 1963. Problems in measuring Change. Madison: The University of Wisconsin Press.
- Harris, D. B., Bisbee, C. I. & Evans, S. H. 1971. Further comments - Misuse of analysis of covariance. *Psychological Bulletin*, 75, 220-222.
- Harris, R. J. 1975. A primer of multivariate statistics. New York: Academic Press.
- Harter, H. L. 1957. Error rates and sample sizes for range tests in multiple comparisons. *Biometrics*, 13, 511-536.
- Hartley, H. 1950. Use of range in analysis of variance. *Biometrika*, 37, 271-280.
- Havlicek, L. L. & Peterson, N. L. 1974. Robustness of the t test: A guide for researchers on effect of Violation of assumptions. *Psychological Report*, 34, 1095-1114.
- Hawkins, D. M. 1979. Fractiles of an extended multiple outlier test. *Journal of Statistical Computation and Simulation*, 8, 227-236.
- Hays, W. L. 1963, 1972, 1973², 1977², 1981³. Statistics for the social sciences. London: Holt, Rinehart & Winston.
- Heckhausen, H. 1969. Allgemeine Psychologie in Experimenten. Göttingen: Hogrefe.
- Hedayat, A. & Afsarinejad, K. 1975. Repeated measurements designs, I. In: Srivastava, J. N. (Ed.): A Survey of statistical design and linear models. Amsterdam: North-Holland, 229-242.
- Heermann, E. F. & Braskamp, L. A. (Eds). 1970. Readings in statistics for the behavioral sciences. Englewood Cliffs, N. J.: Prentice Hall.
- Hegemann, V. & Johnson, D. E. 1976. The power of two tests for nonadditivity. *Journal of the American Statistical Association*, 71(356), 945-948.
- Hempel, C. G. 1965. Aspects of scientific explanation. In: Hempel, C. G. (Ed.): Aspects of scientific explanation and other essays in the philosophy of science. New York: Free Press, 331-396.
- Hempel, C. G. 1974. Grundzüge der Begriffsbildung in der empirischen Wissenschaft. Düsseldorf: Bertelsmann Universitätsverlag.
- Hempel, C. G. & Oppenheim, P. 1948. Studies in the logic of explanation. *Philosophy of Science*, 15, 135-175.
- Henning, H. J. 1978. Interaktion, Transformation und empirische Bedeutsamkeit. *Archiv für Psychologie*, 130, I.: 120-138; II: 236-264.
- Henning, H. J. & Muthig, K. 1979. Grundlagen konstruktiver Versuchsplanung. München: Kösel.

- Henze, F. H.-H. 1979. The exact noncentral distributions of Spearman's r and other related correlation coefficients. *Journal of the American Statistical Association*, 74, 459-464.
- Herr, D. G. & Gaebelin, J. 1978. Nonorthogonal two-way analysis of variance. *Psychological Bulletin*, 85, 207-216.
- Herrmann, T. 1969, 1976³. *Lehrbuch der empirischen Persönlichkeitsforschung*. Göttingen: Hogrefe.
- Herrmann, T. 1973. *Persönlichkeitsmerkmale*. Stuttgart: Kohlhammer.
- Hersen, M. & Barlow, D. H. 1976. *Single-case experimental design*. New York: Pergamon.
- Higbee, K. L. & Wells, M. G. 1972. Some research trends in social psychology during the 1960s. *American Psychologist*, 27, 963-966.
- Hochberg, Y. 1976. A modification of the T-method of multiple comparisons for a oneway-layout with unequal variances. *Journal of the American Statistical Association*, 71, 200-203.
- Hochberg, Y. & Lachenbruch, P. A. 1976. Two-stage multiple comparison procedures based on the studentized range. *Communications in Statistics*, A5(15), 1447-1453.
- Hodges, J. L. & Lehmann, E. L. 1954. Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society*, B, 16, 261-268.
- Hodges, J. L. & Lehmann, E. L. 1956. The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*, 27, 324-335.
- Hogg, R. V. 1974. Adaptive robust procedures: A partial review and some suggestions for further applications and theory. *Journal of the American Statistical Association*, 69, 909-937.
- Hogg, R. V. 1977. An introduction to robust procedures. *Communications in Statistics*, A6(9), 789-794.
- Hogg, R. V. 1979. Statistical robustness: One view of its use in applications today. *The American Statistician*, 33(3), 108-115.
- Hogg, R. V. & Craig, A. T. 1970³, 1971². *Introduction to mathematical statistics*. New York: McMillan. London: Collier McMillan International Edition.
- Hohander, M. & Wolfe, D. A. 1973. *Nonparametric statistical methods*. New York: Wiley.
- Hollingsworth, H. H. 1980. An analytical investigation of the effects of heterogeneous regression slopes in the analysis of covariance. *Educational and Psychological Measurement*, 40, 611-618.
- Holm, K. (Hrsg.) 1975, 1976, 1977. *Die Befragung*. Bde. 1-5. München: Francke.
- Holzkamp, K. 1964. *Theorie und Experiment in der Psychologie*. Berlin: de Gruyter.
- Hopkins, K. D. 1976. A simplified method for determining for expected mean squares and error terms in the analysis of variance. *Journal of Experimental Education*, 45, 13-18.

- Hopkins, K. D. & Anderson, B. L. 1973. A guide to multiple-comparison techniques: Criteria for selecting the „Method of choice“. *Journal of Special Education*, 7, 319-328.
- Horsnell, G. 1953. The effect of unequal group variances on the F-test for the homogeneity of group means. *Biometrika*, 40, 128-136.
- Hotelling, H. 1931. The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2, 360-378.
- Howell, J. F. & Garnes, P. A. 1973. The robustness of the analysis of variance and the Tukey WSD test under various patterns of heterogeneous variances. *Journal of Experimental Education*, 41, 33-37.
- Howell, J. F. & Garnes, P. A. 1974. The effects of variance heterogeneity on simultaneous multiple comparison procedures with equal sample size. *British Journal of Mathematical and Statistical Psychology*, 27, 72-81.
- Hoyle, M. H. 1973. Transformations - An introduction and a bibliography. *International Statistical Review*, 41, 203-223.
- Hsu, L. M. 1978. A Poisson method of controlling the maximum tolerable number of Type I errors. *Perceptual and Motor Skills*, 46, 211-218.
- Huber, P. J. 1972. The 1972 Wald lecture: Robust statistics: A review. *Annals of Mathematical Statistics*, 43, 1041-1067.
- Hubert, L. J. 1973. The use of orthogonal polynomials for trend analysis. *American Educational Research Journal*, 10, 241-244.
- Huberty, C. J. 1972. Multivariate indices of strength of association. *Multivariate Behavior Research*, 7, 523-526.
- Huberty, C. J. & Mourad, S. A. 1980. Estimation in multiple correlation/prediction. *Educational and Psychological Measurement*, 40, 101-112.
- Huck, S. W. & Sandler, H. M. 1973. A note on the Solomon 4-group design: Appropriate statistical analyses. *Journal of Experimental Education*, 42, 54-55.
- Huck, S. W. & Sutton, C. O. 1975. Some comments concerning the use of monotonic transformation to remove the interaction in two-factor ANOVA's. *Educational and Psychological Measurement*, 35, 789-791.
- Huitema, B. E. 1980. The analysis of covariance and alternatives. New York: Wiley.
- Hummel, T. J. 1977. Comment on „An empirical comparison of the methods of least squares and unweighted means . . .“. *International Statistical Review*, 45, 299-301.
- Hummel, T. J. & Sligo, J. R. 1971. Empirical comparison of univariate and multivariate analysis of variance procedures. *Psychological Bulletin*, 76, 49-57.
- Hummell, H. J. & Ziegler, R. (Hrsg.) 1976. *Korrelation und Kausalität*. Bd. 1-3. Stuttgart: Enke.
- Humphreys, L. G. & Fleishman, A. 1974. Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. *Journal of Educational Psychology*, 66, 464-472.

- Hurlburt, R. T. & Spiegel, D. K. 1976. Dependence of F ratios sharing a common denominator mean Square. *American Statistician*, 30, 74-78.
- Huynh, H. 1978. Some approximate tests for repeated measurement designs. *Psychometrika*, 43, 161-175.
- Huynh, H. & Feldt, L. S. 1970. Conditions under which mean square ratios in repeated measurement designs have exact F-distributions. *Journal of the American Statistical Association*, 65, 1582-1589.
- Huynh, H. & Feldt, L. S. 1976. Estimations of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, 1, 69-82.
- Huynh, H. & Mandeville, G. K. 1979. Validity conditions in repeated measures designs. *Psychological Bulletin*, 86, 969-973.
- Irle, M. 1975. *Lehrbuch der Sozialpsychologie*. Göttingen: Hogrefe.
- Jacobs, K. W. 1976. A table for the determination of experimentwise error rate (α) from independent comparisons. *Educational and Psychological Measurement*, 36, 899-903.
- James, G. S. 1951. The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, 38, 329-329.
- Jennings, E. 1967. Fixed effects analysis of variance by regression analysis. *Multivariate Behavioral Research*, 2, 95-108.
- Jennings, E. 1978. Fixed effects analysis of variance with unequal cell sizes. *Journal of Experimental Education*, 46, 42-51.
- John, J. A. & Quenouille, M. H. 1977². *Experiments: Design and analysis*. London: Griffin.
- John, P. W. M. 1971. *Statistical design and analysis of experiments*. New York: Macmillan.
- Johnson, D. E. & Graybill, F. A. 1972. An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association*, 67, 862-868.
- Jonckheere, A. R. 1954(a). A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41, 133-145.
- Jonckheere, A. R. 1954(b). A test of significance for the relation between m rankings and k ranked categories. *British Journal of Statistical Psychology*, 7, 93-100.
- Jurs, S. G. & Glass, G. V. 1971. The effect of experimental mortality on the internal and external validity of the randomized comparative experiment. *Journal of Experimental Education*, 40, 62-66.
- Kaiser, H. F. 1960. Directional statistical decisions. *Psychological Review*, 67, 160-167.
- Katzer, J. & Sodt, J. 1973. An analysis of the use of statistical testing in communication research. *Journal of Communication*, 23, 251-265.

- Kazdin, A. E. 1980. Research design in clinical psychology. New York: Harper & Row.
- Kemp, K. E. & Conover, W. J. 1973. Robustness and power of the t-test compared with some nonparametric alternatives when sampling from a Poisson distribution. *Journal of Statistical Computation and Simulation*, 2, 293-307.
- Kempthorne, O. 1952, 1973⁶. Design and analysis of experiments. New York: Wiley. Huntington, N. Y.: Krieger.
- Kempthorne, O. 1955. The randomization theory of statistical inference. *Journal of the American Statistical Association*, 50, 946-967.
- Kendall, M. G. 1970⁴. Rank correlation methods. London: Griffin.
- Kendall, M. G. 1973. Time-series. London: Griffin.
- Kendall, M. G. 1975. Multivariate analysis. London: Griffin.
- Kendall, M. G. & Stuart, A. The advanced theory of statistics. Vol. I, II und III. London: Griffin, Vol. 1: 1958, 1963², 1969³; Vol. II: 1961, 1973²; Vol. III: 1966.
- Kennedy, J. J. 1970. The eta coefficient in complex ANOVA designs. *Educational and Psychological Measurement*, 30, 885-889.
- Kenny, D. A. 1979. Correlation and causality. New York: Wiley.
- Keppel, G. 1973. Design and analysis: A researcher's handbook. Englewood Cliffs, N. Y.: Prentice-Hall.
- Keren, G. & Lewis, C. 1979. Partial omega Squared for ANOVA designs. *Educational and Psychological Measurement*, 39, 119-128.
- Kerlinger, F. N. 1964, 1973. Foundations of behavioral research. New York: Holt, Rinehart & Winston. Deutsch: Grundlagen der Sozialwissenschaften. Weinheim und Basel: Beltz, Band 1: 1975, 1978², Band 2: 1979.
- Kerlinger, F. N. & Pedhazur, E. J. 1973. Multiple regression in behavioral research. New York: Holt, Rinehart & Winston.
- Keselman, H. J. 1975. A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared. *Canadian Psychological Review*, 16, 44-48.
- Keselman, H. J. 1976. A power investigation of the Tukey multiple comparison statistic. *Educational and Psychological Measurement*, 36, 97-104.
- Keselman, H., Garnes, P. A. & Rogan, J. C. 1979a. Protecting the Overall rate of type I errors for pairwise comparisons with an omnibus test statistic. *Psychological Bulletin*, 86, 884-888.
- Keselman, H., Garnes, P. A. & Rogan, J. C. 1979b. An addendum to: „A comparison of the modified Tukey and Scheffé methods of multiple comparisons for pairwise contrasts.“ *Journal of the American Statistical Association*, 74, 626-627.
- Keselman, H. J., Garnes, P. A. & Rogan, J. 1980. Type I and type II errors in simultaneous and two-stage multiple comparison procedures. *Psychological Bulletin*, 88, 356-358.
- Keselman, H. J., Murray, R. & Rogan, J. C. 1976. Effect of very unequal group sizes

- on Tukey's multiple comparison test. *Educational and Psychological Measurement*, 36, 263-270.
- Keselman, H. J. & Rogan, J. C. 1977. An evaluation of some non-parametric and parametric tests for multiple comparisons. *British Journal of Mathematical and Statistical Psychology*, 30, 125-133.
- Keselman, H. J. & Rogan, J. C. 1978. A comparison of the modified Tukey and Scheffé methods of multiple comparisons for pairwise contrasts. *Journal of the American Statistical Association*, 73, 47-52.
- Keselman, H. J., Rogan, J. C. & Feir-Walsh, B. J. 1977. An evaluation of some nonparametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology*, 30, 213-221.
- Keselman, H. J., Rogan, J. C., Mendoza, J. L. & Breen, L. C. 1980. Testing the validity conditions of repeated measures F-tests. *Psychological Bulletin*, 87, 479-481.
- Keselman, H. J. & Toothaker, L. E. 1973. Error rates for multiple comparison methods: Some evidence concerning the misleading conclusions of Petrinovich and Hardyck. *Psychological Bulletin*, 80, 31-32.
- Keselman, H. J. & Toothaker, L. E. 1974. Comparison of Tukey's T-method and Scheffé's S-method for various numbers of all possible differences of average contrasts under violation of assumptions. *Educational and Psychological Measurement*, 34, 511-519.
- Keselman, H. J., Toothaker, L. E. & Shooter, M. 1975. An evaluation of two unequal n_k forms of the Tukey multiple comparison statistic. *Journal of the American Statistical Association*, 70, 584-587.
- Kimball, A. W. 1951. On dependent tests of significance in the analysis of variance. *Annals of Mathematical Statistics*, 22, 600-602.
- Kirk, R. E. 1968. *Experimental design: Procedures for the behavioral sciences*. Belmont, Cal.: Brooks/Cole (Wadsworth).
- Kirk, R. E. (Ed.) 1972. *Statistical issues. A Reader for the behavioral sciences*. Belmont, Cal. : Wadsworth.
- Klauer, K. J. 1973. *Das Experiment in der pädagogischen Forschung*. Düsseldorf: Schwann.
- Kleijnen, J. P. C. 1974, 1975. *Statistical techniques in Simulation*. Part I und II. New York: Dekker.
- Kleiter, G. 1969. Krise der Signifikanztests in der Psychologie. *Jahrbuch für Psychologie, Psychotherapie und medizinische Anthropologie*, 17, 144-163.
- Kleiter, G. D. 1981. *Bayes-Statistik*. Berlin: de Gruyter.
- Knapp, T. R. 1978. Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, 85, 410-416.
- Koch, J.-J. 1976. „Guter Eindruck“ und Attitüden. Ein Beitrag zum Problem der „Verfälschbarkeit“ von Einstellungsinventaren. *Archiv für Psychologie*, 128, 135-149.

- Kohr, R. L. & Garnes, P. A. 1974. Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *Journal of Experimental Education*, 43, 61-69.
- Kohr, R. L. & Garnes, P. A. 1977. Testing complex a priori contrasts on means from independent samples. *Journal of Educational Statistics*, 2, 207-216.
- Konijn, H. S. 1973. *Statistical theory of sample survey design and analysis*. Amsterdam: North-Holland.
- Kraemer, H. C. 1974. The non-null distribution of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 69, 114-117.
- Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. 1971. *Foundations of measurement*. Vol. I: Additive and polynomial representations. New York: Academic Press.
- Kranz, H. T. 1979. *Einführung in die klassische Testtheorie*. Frankfurt am Main: Verlagsbuchhandlung für Psychologie.
- Kratochwill, T. R. 1978. *Single subject research*. New York: Academic Press.
- Krause, B. & Metzler, P. 1978. Zur Anwendung der Inferenzstatistik in der psychologischen Forschung. *Zeitschrift für Psychologie*, 186, 244-267.
- Krauth, J. 1980. Ein Vergleich der Konfigurationsfrequenzanalyse mit der Methode der log-linearen Modelle. *Zeitschrift für Sozialpsychologie*, 11, 233-247.
- Krauth, J. & Lienert, G. A. 1973. *Die Konfigurationsfrequenzanalyse (KFA)*. Freiburg: Alber.
- Kreyszig, E. 1973. *Statistische Methoden und ihre Anwendungen*. Göttingen: Vandenhoeck & Ruprecht.
- Kriz, J. 1973, 1978. *Statistik in den Sozialwissenschaften*. Reinbek: Rowohlt (rororo 1080).
- Kroll, R. M. & Chase, L. J. 1975. Communication disorders: A power analytic assessment of recent research. *Journal of Communication Disorders*, 8, 237-247.
- Krüger, H.-P. 1977. Simultane U-Tests zur exakten Prüfung von Haupt- und Wechselwirkungen an 2²-faktoriellen Versuchsplänen. *Psychologische Beiträge*, 19, 110-120.
- Kruglanski, A. W. 1975. The human subject in the psychology experiment: Fact and artifact. In: Berkowitz, L. (Ed.): *Advances in experimental social psychology*. Vol. 8. New York: Academic Press, 101-147.
- Kruskal, W. H. & Wallis, W. A. 1952. Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association*, 47, 583-621.
- Ku, H. H. & Kullback, S. 1968. Interaction in multidimensional contingency tables: An information theoretic approach. *Journal of Research of the National Bureau of Standards - Mathematical Sciences*, 72B, 159-199.
- Küchler, M. 1979. *Multivariate Analyseverfahren*. Stuttgart: Teubner.
- Küttner, M. 1976. Ein verbesserter deduktiv-nomologischer Erklärungs-begriff. *Zeitschrift für allgemeine Wissenschaftstheorie*, 7, 274-297.

- Labovitz, S. 1968. Criteria for selecting a significance level: A note on the sacredness of .05. *The American Sociologist*, 3, 220-222.
- Läuter, J. 1978. Sample size requirements for the T^2 -Test of MANOVA (Tables for one-way classification). *Biometrical Journal*, 20, 389-406.
- La Forge, R. 1967. Confidence intervals or tests of significance in scientific research? *Psychological Bulletin*, 68, 446-447.
- Lakatos, J. 1970. Falsification and the methodology of scientific research programs. In: Lakatos, I. & Musgrave, A. (Eds): *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press, 91-196. Deutsch: *Falsifikation und die Methodologie wissenschaftlicher Forschungsprogramme*. In: Lakatos, I. & Musgrave, A. (Hrsg.): *Kritik und Erkenntnisfortschritt*. Braunschweig: Vieweg, 1974.
- Lana, R. E. 1969. Pretest sensitization. In: Rosenthal, R. L. & Rosnow, R. C. (Eds): *Artifact in behavioral research*. New York: Academic Press, 121-141.
- Lana, R. E. & Lubin, A. 1963. The effect of correlation on the repeated measures design. *Educational and Psychological Measurement*, 23, 729-739.
- Lane, D. M. & Dunlap, W. P. 1978. Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, 31, 107-112.
- Langeheine, R. 1980. *Log-lineare Modelle zur multivariaten Analyse qualitativer Daten*. München: Oldenbourg.
- Lantermann, E.-D. 1976. Zum Problem der Angemessenheit eines inferenzstatistischen Verfahrens. *Psychologische Beiträge*, 18, 99-109.
- Layard, M. W. J. 1973. Robust large-sample tests for homogeneity of variances. *Journal of the American Statistical Association*, 68, 195-198.
- Lee, E. T., Desu, M. M. & Gehan, E. A. 1975. A Monte Carlo study of the power of some two-sample tests. *Biometrika*, 62, 425-432.
- Lee, M. C. 1961. Interactions, configurations, and nonadditive models. *Educational and Psychological Measurement*, 21, 797-805.
- Lee, W. 1975. *Experimental design and analysis*. San Francisco: Freeman.
- Leeb, S. & Weinberg, S. 1977. Unreliability of difference scores: A clarification of the issues. *Psychological Reports*, 40, 931-935.
- Lehmann, E. L. 1950. Some principles of the theory of testing hypotheses. *Annals of Mathematical Statistics*, 21, 1-26.
- Lehmann, E. L. 1958. Significance level and power. *Annals of Mathematical Statistics*, 29, 1167-1176.
- Lehmann, E. L. 1968. Hypothesis testing. In: Sills, D. L. (Ed.): *International Encyclopedia of the Social Sciences*. Vol. 7. New York: Macmillan and Free Press, 40-47.
- Lehmann, E. L. 1975. *Nonparametrics: Statistical methods based on ranks*. San Francisco: Holden-Day.
- Lehmann, E. L. & Shaffer, J. P. 1977. On a fundamental theorem in multiple comparison. *Journal of the American Statistical Association*, 72, 576-578.

- Leiser, E. 1978. Einführung in die statistischen Methoden der Erkenntnisgewinnung. Köln: Pahl-Rugenstein.
- Leslie, R. I. & Brown, B. M. 1966. Use of range in testing heterogeneity of variance. *Biometrika*, 53, 221-227.
- Levene, H. 1960. Robust test for equality of variances. In: Olkin, I. (Ed.): Contributions to probability and statistics. Stanford, Cal.: Stanford University Press, 278-292.
- Levin, J. R. 1971. Some new approaches and approximations to statistical power in educational research. Unpublished Manuscript. Madison: University of Wisconsin.
- Levin, J. R. 1975. Determining sample size for planned and post hoc analysis of variance comparisons. *Journal of Educational Measurement*, 12, 99-108.
- Levin, J. R. & Marascuilo, L. A. 1972. Type IV errors and interactions. *Psychological Bulletin*, 78, 368-374.
- Levin, J. R. & Marascuilo, L. A. 1973. Type IV errors and Games. *Psychological Bulletin*, 80, 308-309.
- Levin, J. R. & Subkoviak, M. J. 1977. Planning an experiment in the company of measurement error. *Applied Psychological Measurement*, 1, 331-338.
- Levin, J. R. & Subkoviak, M. J. 1978. Correcting „Planning an experiment in the Company of measurement error“. *Applied Psychological Measurement*, 2(3), 382-385.
- Levy, K. J. 1975(a). Some multiple range tests for variances. *Educational and Psychological Measurement*, 35, 599-604.
- Levy, K. J. 1975(b). Comparing the variances of several treatments with a control. *Educational and Psychological Measurement*, 35, 793-796.
- Levy, K. J. 1975(c). An empirical comparison of several multiple range test for variances. *Journal of the American Statistical Association*, 70, 186-183.
- Levy, K. J. 1975(d). An empirical comparison of the Z-variance and Box-Scheffé tests for homogeneity of variance. *Psychometrika*, 40, 519-524.
- Levy, K. J. 1978(a). An empirical study of the tube-root test for homogeneity of variance with respect to the effects of non-normality and power. *Journal of Statistical Computation and Simulation*, 7, 71-78.
- Levy, K. J. 1978(b). Some empirical power results associated with Welch's robust analysis of variance technique. *Journal of Statistical Computation and Simulation*, 8, 43-48.
- Levy, K. J. 1978(c). An empirical comparison of the ANOVA F-test with alternatives which are more robust against heterogeneity of variance. *Journal of Statistical Computation and Simulation*, 8, 49-57.
- Levy, K. J. 1978(d). A priori contrasts under conditions of variance heterogeneity. *Journal of Experimental Education*, 47, 42-45.
- Levy, K. J. 1979(a). Nonparametric large sample many-one comparisons. *Journal of Experimental Education*, 48, 45-49.

- Levy, K. J. 1979(b). Nonparametric large-sample pairwise comparisons. *Psychological Bulletin*, 86, 371-375.
- Levy, K. J., Narula, S. C. & Abrami, P. F. 1975. An empirical comparison of the methods of least squares and unweighted means for the analysis of disproportionate cell data. *International Statistical Review*, 43, 335-338.
- Levy, P. 1967. Substantive significance of significant differences between two groups. *Psychological Bulletin*, 67, 37-40.
- Lewis, C. & Keren, G. 1977. You can't have your cake and eat it too: Some considerations of the error term. *Psychological Bulletin*, 84, 1150-1154.
- Li, C. C. 1964. *Introduction to experimental statistics*. New York: McGraw-Hill.
- Lieberman, B. (Ed.) 1971. *Contemporary problems in Statistics*. New York: Oxford University Press.
- Lienert, G. A. 1962. über die Anwendung von Variablen-Transformationen in der Psychologie. *Biometrische Zeitschrift*, 4, 145-181.
- Lienert, G. A. *Verteilungsfreie Methoden in der Biostatistik*. Bd. 1 und 2. Meisenheim am Glan: Anton Hain. Bd. 1: 1973², Bd. II: 1978², Tafelband: 1975.
- Lienert, G. A. & Marascuilo, L. A. 1980. Comparing treatment-induced changes for k independent samples of paired observations. *Biometrical Journal*, 22, 763-777.
- Lienert, G. A. & Raatz, U. 1971. Das Rangkorrelationsverhältnis Eta-H-Quadrat als nicht-lineares Abhängigkeitsmaß. *Biometrische Zeitschrift*, 13, 407-413.
- Lindman, H. R. 1974. *Analysis of variance in complex experimental designs*. San Francisco: Freeman.
- Lindquist, E. F. 1953. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.
- Linn, R. L. & Slinde, J. A. 1977. Significance of pre- and post-test change. *Review of Educational Research*, 47, 121-150.
- Lippman, L. G. & Taylor, C. J. 1972. Multiple comparisons in complex ANOVA designs. *Journal of General Psychology*, 86, 221-223.
- Lissitz, R. W. & Chardos, S. 1975. A study of the effect of the Violation of the assumption of independent sampling upon the type I error rate of the two-group t-test. *Educational and Psychological Measurement*, 45, 353-359.
- Loftus, G. R. 1978. On interpretation of interactions. *Memory and Cognition*, 6, 312-319.
- Lord, F. M. 1953. On the statistical treatment of football numbers. *American Psychologist*, 8, 750-751.
- Lord, F. M. 1967. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304-305.
- Lord, F. M. & Novick, M. R. 1968. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.
- Lubin, A. 1961. The interpretation of significant interaction. *Educational and Psychological Measurement*, 21, 807-817.

- Lunney, G. H. 1969. Individual size for multiple t-tests. *American Educational Research Journal*, 6, 701-703.
- MacCorquodale, K. & Meehl, P. E. 1948. On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55, 95-107.
- Mace, A. E. 1964. Sample size determination. New York: Holt, Rinehart & Winston.
- Mann, H. B. & Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50-60.
- Marascuilo, L. A. 1966. Large-sample multiple comparisons. *Psychological Bulletin*, 65, 280-290.
- Marascuilo, L. A. & Levin, J. R. 1970. Appropriate post hoc comparisons for interaction and nested hypothesis in analysis of variance designs: The elimination of type IV errors. *American Educational Research Journal*, 7, 397-421.
- Marascuilo, L. A. & Levin, J. R. 1976. The simultaneous investigation of interaction and nested hypotheses in two-factor analysis of variance designs. *American Educational Research Journal*, 13, 61-65.
- Marascuilo, L. A. & McSweeney, M. 1967. Nonparametric post hoc comparisons for trend. *Psychological Bulletin*, 67, 401-412.
- Marascuilo, L. A. & McSweeney, M. 1977. Nonparametric and distribution-free methods for the social sciences. Monterey, Cal.: Brooks/Cole.
- Marks, M. R. 1951. Two kinds of experiments distinguished in terms of statistical operations. *Psychological Review*, 58, 179-184.
- Martin, C. G. 1976. Comment on Levy's „An empirical comparison of the Z-variance and Box-Scheffé tests for homogeneity of variance“. *Psychometrika*, 41, 551-556.
- Martin, C. G. & Garnes, P. A. 1977. ANOVA tests for homogeneity of variance: nonnormality and unequal samples. *Journal of Educational Statistics*, 2, 187-206.
- Massaro, D. W. 1975. Experimental psychology and information processing. Chicago: Rand McNally.
- Matheson, D. W., Bruce, R. L. & Beauchamp, K. L. 1970. Introduction to experimental psychology. New York: Holt, Rinehart & Winston.
- Mauchly, J. W. 1940. Significance test for sphericity of a normal n-variate distribution. *Annals of Mathematical Statistics*, 11, 204-209.
- May, R. B. & Konkin, P. R. 1970. A nonparametric test of an ordered hypothesis for k independent samples. *Educational and Psychological Measurement*, 30, 251-258.
- McCall, R. B. 1970. The use of multivariate procedures in developmental psychology. In: Mussen, P. H. (Ed.): Carmichael's manual of child psychology. Vol. I. New York: Wiley, 1366-1377.
- McCall, R. B. & Appelbaum, M. J. 1973. Bias in the analysis of repeated-measures designs: Some alternative approaches. *Child Development*, 44, 401-415.

- McGuigan, F. J. 1979. Einführung in die experimentelle Psychologie (bearbeitet von J. M. Diehl). Frankfurt am Main: Fachbuchhandlung für Psychologie.
- McHugh, R. B. 1963. Comments on „Scales and statistics: parametric and nonparametric“. *Psychological Bulletin*, 60, 350-355.
- McNemar, Q. 1949, 1955, 1962³. *Psychological statistics*. New York: Wiley.
- McNemar, Q. 1960. At random: Sense and nonsense. *American Psychologist*, 15, 295-300.
- McSweeney, M. & Katz, B. M. 1978. Nonparametric statistics: Use and nonuse. *Perceptual and Motor Skills*, 46, 1023-1032.
- McSweeney, M. & Marascuilo, L. A. 1969. Nonparametric methods. *Review of Educational Research*, 39, 728-734.
- Meehl, P. E. 1967. Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103-115.
- Meehl, P. E. 1970. Nuisance variables and the ex-Post-facto design. In: Radner, M. & Winokur, S. (Eds): *Minnesota studies in the philosophy of science*. Vol. IV: *Analyses of theories and methods of physics and psychology*. Minneapolis: University of Minnesota Press, 373-402.
- Mehta, J. S. & Srinivasan, R. 1970. On the Behrens-Fisher problem. *Biometrika*, 57, 649-655.
- Melton, A. W. 1962. Editorial. *Journal of Experimental Psychology*, 64, 553-557.
- Mendenhall, W. 1968. *Introduction to linear models and the design and analysis of experiments*. Belmont, Cal.: Wadsworth.
- Mendoza, J. L. 1980. A significance test for multisample sphericity. *Psychometrika*, 45, 495-498.
- Mendoza, J. L., Toothaker, L. E. & Crain, B. R. 1976. Necessary and sufficient conditions for F-ratios in the LxJxK factorial design with two repeated factors. *Journal of the American Statistical Association*, 71, 992-993.
- Mendoza, J. L., Toothaker, L. E. & Nicewander, W. A. 1974. A Monte Carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures design. *Multivariate Behavioral Research*, 9, 165-178.
- Menges, G. 1968, 1972². *Grundriß der Statistik*. Teil 1: Theorie. Opladen: Westdeutscher Verlag.
- Meredith, W. M., Fredericksen, S. H. & McLaughlin, D. 1974. Statistics and data-analysis. *Annual Review of Psychology*, 25, 453-505.
- Mertens, W. 1975. *Sozialpsychologie des Experiments*. Hamburg: Hoffmann & Campe.
- Miller, G. R. 1970. Research setting: laboratory studies. In: Emmert, P. & Brooks, W. D. (Eds): *Methods of research in communications*. New York: Houghton Mifflin, 77-104.
- Miller, J. J. 1979. Maximum likelihood estimation of variance components - A Monte Carlo study. *Journal of Statistical Computation and Simulation*, 8, 175-190.

- Miller, R. G. Jr. 1966. Simultaneous statistical inference. New York: McGraw-Hill.
- Miller, R. G. 1968. Jackknifing variances. *Annals of Mathematical Statistics*, 39, 567-582.
- Miller, R. G. 1974. The jackknife - a review. *Biometrika*, 61, 1-15.
- Miller, R. G. 1977. Developments in multiple comparison methods. *Journal of the American Statistical Association*, 72, 779-788.
- Milliken, G. A. & McDonald, L. L. 1976. Linear models and their analysis in the case of missing or incomplete data: a unifying approach. *Biometrische Zeitschrift*, 18, 381-396.
- Mood, A. M., Graybill, F. A. & Boes, D. C. 1974³. Introduction to the theory of statistics. Tokyo: McGraw-Hill.
- Moosbrugger, H. 1978. Multivariate statistische Analyseverfahren. Stuttgart: Kohlhammer.
- Morrison, D. F. 1967, 1976². Multivariate statistical methods. New York: McGraw-Hill.
- Morrison, D. E. & Henkel, R. E. (Eds). 1970. The significance test controversy. A reader. Chicago: Aldine.
- Moses, L. E. & Oakford, R. V. 1963. Tables of random permutations. Stanford, Cal.: Stanford University Press.
- Mosteller, F. 1968. Errors: Nonsampling errors. In: Sills, D. L. (Ed.): *International Encyclopedia of the Social Sciences*. Vol. Boston: Crowell Collier and Macmillan, 113-132.
- Mosteller, F. & Bush, R. R. 1954. Selected quantitative techniques. In: Lindzey, G. (Ed.): *Handbook of social psychology*. Vol. 1. Cambridge, Mass.: Addison-Wesley, 289-334.
- Mosteller, F. & Tukey, J. W. 1968. Data analysis, including statistics. In: Lindzey, G. & Aronson, E. (Eds): *Handbook of social psychology*. Vol. 2. Reading, Mass.: Addison-Wesley, 80-203.
- Mosteller, F. & Tukey, J. W. 1977. Data analysis and regression. Reading, Mass.: Addison-Wesley.
- Myers, J. L. 1972². Fundamentals of experimental design. Boston: Allyn & Bacon.
- Namboodiri, N. K. 1972. Experimental designs in which each subject is used repeatedly. *Psychological Bulletin*, 77, 54-64.
- Namboodiri, N. K., Carter, L. F. & Blalock, H. M. Jr. 1975. Applied multivariate analysis and experimental designs. New York: McGraw-Hill.
- Narula, S. C., Abrami, P. F. & Levy, K. J. 1976. An empirical comparison of several methods for analyzing disproportionate subclass data in a two-way classification ANOVA. *International Statistical Review*, 44, 341-348.
- Neave, H. R. & Granger, C. W. J. 1968. A Monte Carlo study comparing various two-sample-tests for differences in mean. *Technometrics*, 10, 509-522.
- Nelson, P. L. & Toothaker, L. E. 1975. An empirical study of Jonckheere's non-

- parametric test of ordered alternatives. *British Journal of Mathematical and Statistical Psychology*, 28, 167-176.
- Neyman, J. 1952². Lectures and conferences on mathematical statistics and probability. Washington: Graduate School U. S. Department of Agriculture.
- Neyman, J. & Pearson, E. S. 1933(a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, 231, 289-337.
- Neyman, J. & Pearson, E. S. 1933(b). The testing of statistical hypotheses in relation to probabilities. *Proceedings of the Cambridge Philosophical Society*, 29, 492-510.
- Neyman, J. & Pearson, E. S. 1936. Contributions to the theory of testing statistical hypotheses. Pt. I: Unbiased critical regions of type A and type A₁. *Statistical Research Memoirs*, 1-37.
- Neyman, J. & Pearson, E. S. 1938. Contribution to the theory of testing statistical hypotheses. Pt. II: Certain theorems on unbiased critical regions of type A. *Statistical Research Memoirs*, 2, 25-36.
- Neyman, J. & Pearson, E. S. 1938. Contribution to the theory of testing statistical hypotheses. Pt. III: Unbiased tests of simple statistical hypotheses specifying the values of more than one unknown Parameter. *Statistical Research Memoirs*, 2, 36-57.
- Nicewander, W. A. & Price, J. M. 1978. Rependent variable reliability and the power of significance tests. *Psychological Bulletin*, 85, 405-409.
- Nunnally, J. C. 1960. The place of statistics in psychology. *Educational and Psychological Measurement*, 20, 641-650.
- Oakes, W. F. 1972. External validity and the use of real people as subjects. *American Psychologist*, 27, 959-962.
- O'Brien, R. G. 1978. Robust techniques for testing heterogeneity of variance effects in factorial designs. *Psychometrika*, 43, 327-342.
- O'Brien, R. G. 1979. An improved ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, 74, 877-880.
- Odeh, R. E. 1972. On the power of Jonckheere's k-sample rank tests against ordered alternatives. *Biometrika*, 59, 467-471.
- Odeh, R. E. & Fox, M. 1975. Sample size choice. New York: Dekker.
- Oliver, R. L. & Berger, P. K. 1980. Advisability of pretest designs in psychological research. *Perceptual and Motor Skills*, 51, 463-471.
- Olkin, I. & Pratt, J. W. 1958. Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 29, 201-211.
- Olson, C. L. 1974. Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 69, 894-908.
- Olson, C. L. 1976. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83, 579-586.

- Olson, C. L. 1979. Practical considerations in choosing a MANOVA test statistic: A rejoinder to Stevens. *Psychological Bulletin*, 86, 1350-1352.
- O'Neill, R. & Wetherill, G. B. 1971. The present state of multiple comparison methods (with discussion). *Journal of the Royal Statistical Society, Series B*, 33, 218-241.
- Opp, K.-D. & Schmidt, P. 1976. Einführung in die Mehrvariablenanalyse. Reinbek: Rowohlt.
- Orne, M. T. 1962. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776-783.
- Orth, D. 1974. Einführung in die Theorie des Messens. Stuttgart: Kohlhammer.
- Overall, J. E. & Dalal, S. N. 1965. Design of experiments to maximize power relative to cost. *Psychological Bulletin*, 64, 339-350.
- Overall, J. E. & Klett, C. J. 1972. Applied multivariate analysis. New York: McGraw-Hill.
- Overall, J. E. & Spiegel, D. K. 1969. Concerning least squares analysis of experimental data. *Psychological Bulletin*, 72, 311-322.
- Overall, J. E. & Woodward, J. A. 1974. A simple test for heterogeneity of variance in complex factorial designs. *Psychometrika*, 39, 311-318.
- Overall, J. E. & Woodward, J. A. 1975. Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, 82, 85-86.
- Overall, J. E. & Woodward, J. A. 1976. Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin*, 83, 776-777.
- Overall, J. E. & Woodward, J. A. 1977(a). Nonrandom assignment and the analysis of covariance. *Psychological Bulletin*, 84, 588-594.
- Overall, J. E. & Woodward, J. A. 1977(b). Common misconceptions concerning the analysis of covariance. *Multivariate Behavioral Research*, 12, 171-185.
- Owen, D. B. 1962. Handbook of statistical tables. Reading, Mass.: Addison-Wesley.
- Page, E. B. 1963. Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 58, 216-230.
- Patry, J. L. 1979. Feldforschung in den Sozialwissenschaften. *Zeitschrift für Klinische Psychologie und Psychotherapie*, 27, 317-335.
- Patry, J. L. (Ed.) 1982. Laborforschung/Feldforschung. Bern: Huber.
- Pearson, E. S. 1931. The analysis of variance in cases of non-normal Variation. *Biometrika*, 23, 114-133.
- Pearson, E. S. 1966. Alternative tests for heterogeneity of variance; some Monte Carlo results. *Biometrika*, 53, 229-234.
- Pearson, E. S. & Hartley, H. O. Vol. 1: 1954, 1962²; Vol. II: 1972. *Biometrika tables for statisticians*. Vol. I und II. Cambridge: Cambridge University Press.

- Pearson, E. S. & Please, N. W. 1975. Relation between the shape of population distributions and the robustness of four simple test statistics. *Biometrika*, 62, 223-241.
- Pearson, K. 1911. On a correction to be made to the correlation ratio η . *Biometrika*, 8, 254-256.
- Pedhazur, E. J. 1977. Coding subjects in repeated measures designs. *Psychological Bulletin*, 84, 298-305.
- Perlmutter, J. & Myers, J. L. 1973. A comparison of two procedures for testing multiple contrasts. *Psychological Bulletin*, 79, 181-184.
- Petermann, F. 1978. *Veränderungsmessung*. Stuttgart: Kohlhammer.
- Petermann, F. 1981. Möglichkeiten der Einzelfallanalyse in der Psychologie. *Psychologische Rundschau*, 32, 31-48.
- Petermann, F. & Hehl, F. J. (Hrsg.) 1979. *Einzelfallanalyse*. München: Urban & Schwarzenberg.
- Petrinovich, L. F. & Hardyck, C. D. 1969. Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. *Psychological Bulletin*, 71, 43-54.
- Pfanzagl, J. Bd. 1: 1972⁵, Bd. II: 1978⁵. *Allgemeine Methodenlehre der Statistik*. Bd. I und II. Berlin: de Gruyter.
- Pfanzagl, J. 1968. *Theory of Measurement*. Würzburg: Physica.
- Philips, L. D. 1974. *Bayesian statistics for social scientists*. New York: Crowell.
- Pillai, K. C. S. 1955. Some new test criteria in multivariate analysis. *Annals of Mathematical Statistics*, 26, 117-121.
- Pillemer, D. B. & Light, R. J. 1980. Synthesizing outcomes: How to use research evidence from many studies. *Harvard Educational Review*, 50, 176-195.
- Pitman, E. J. G. 1948. *Notes on non-parametric statistical inference*. New York: Columbia University Press.
- Plomp, T. 1974. The statistical basis for aptitude-treatment interactions research: Definitions and techniques. In: Verreck, W. A. (Ed.): *Methodological problems in research and development in higher education*. Amsterdam: Swets & Zeitlinger, 293-323.
- Poor, D. S. 1973. Analysis of variance for repeated measures designs: Two approaches. *Psychological Bulletin*, 80, 204-209.
- Popper, K. R. 1976⁶. *Logik der Forschung*. Tübingen: J. C. B. Mohr (Paul Siebeck).
- Popper, K. R. 1979. *Ausgangspunkte*. Hamburg: Hoffmann und Campe.
- Posten, H. O. 1978. The robustness of the two sample t-test over the Pearson System. *Journal of Statistical Computation and Simulation*, 6, 295-311.
- Pratt, J. W. 1964. Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association*, 59, 665-680.

- Preiser, S. 1977. Die experimentelle Methode. In: Strube, G. (Hrsg.): Die Psychologie des 20. Jahrhunderts. Bd. 5: Binet und die Folgen. Zürich: Kindler, 102-150.
- Puri, M. L. 1965. Some distribution-free k-sample rank tests of homogeneity against ordered alternatives. *Communications on Pure and Applied Mathematics*, 18, 51-63.
- Puri, M. L. & Sen, P. H. 1971. *Nonparametric methods in multivariate analysis*. New York: Wiley.
- Raghavarao, D. 1971. *Constructions and combinatorial problems in design of Experiments*. New York: Wiley.
- Ramsey, P. H. 1978. Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, 73, 479-485.
- Ramsey, P. H. 1980. Choosing the most powerful pairwise multiple comparison procedure in multivariate analysis of variance. *Journal of Applied Psychology*, 317-326.
- Ramseyer, G. C. & Tcheng, T.-K. 1973. The robustness of the studentized range statistic to violations of the normality and homogeneity of variance assumptions. *American Educational Research Journal*, 10, 235-240.
- Rasch, D., Enderlein, G. & Herrendörfer, G. 1973. *Biometrie*. Berlin: Deutscher Verlag der Wissenschaften.
- Rasch, D., Herrendörfer, G., Bock, J. & Busch, K. 1978. *Verfahrensbibliothek Versuchsplanung und -auswertung*. Band 1 und 2. Berlin: VEB Deutscher Landwirtschaftsverlag.
- Ray, W. S. 1966. Logic for a randomization test. *Behavioral Science*, 11, 405-406.
- Redding, W. C. 1970. Research setting: Field studies. In: Emmert, P. & Brooks, W. D. (Eds): *Methods of research in communication*. Boston: Houghton Mifflin, 105-159.
- Renn, H. 1975. *Nichtparametrische Statistik*. Stuttgart: Teubner.
- Rennert, M. 1977. Einige Anmerkungen zur Verwendung von Differenzen bei der Veränderungsmessung. *Psychologische Beiträge*, 19, 100-109.
- Restle, F. & Greeno, J. G. 1970. *Introduction to mathematical psychology*. Reading: Addison-Wesley.
- Revenstorf, D. 1979. *Zeitreihenanalyse für klinische Daten*. Weinheim: Beltz.
- Robinson, J. 1973(a). The large sample power of Permutation tests for randomization models. *Annals of Statistics*, 1, 291-296.
- Robinson, J. 1973(b). The analysis of covariance under a randomization model. *Journal of the Royal Statistical Society, Ser. B*, 35, 368-376.
- Rodger, R. S. 1967. Type II errors and their decision basis. *British Journal of Mathematical and Statistical Psychology*, 20, 187-204.
- Rodger, R. S. 1973. Confidence intervals for multiple comparisons and the misuse of the Bonferroni inequality. *British Journal of Mathematical and Statistical Psychology*, 26, 58-60.

- Rodger, R. S. 1974. Multiple contrasts, factors, error rate and power. *British Journal of Mathematical and Statistical Psychology*, 27, 178-198.
- Rodger, R. S. 1975(a). The number of non-zero, post-hoc contrasts from ANOVA and error rate. *British Journal of Mathematical and Statistical Psychology*, 28, 71-78.
- Rodger, R. S. 1975(b). Setting rejection rate for contrasts selected post hoc when some nulls are false. *British Journal of Mathematical and Statistical Psychology*, 28, 214-232.
- Rodger, R. S. 1976. Tables on Stein's non-central Parameter $D\beta_j$, v_1 , v_2 required to set power for numerical alternatives to H_0 tested by two-stage sampling Anova. *Journal of Statistical Computation and Simulation*, 5, 1-22.
- Rodger, R. S. 1978. Two-stage sampling to set sample size for post hoc tests in ANOVA with decision-based error rates. *British Journal of Mathematical and Statistical Psychology*, 31, 153-178.
- Röhr, M. 1975. Theorie und Methodik der Kovarianzanalyse. *Probleme und Ergebnisse der Psychologie*, 52, 19-45.
- Rogan, J. C. & Keselman, H. J. 1977. Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of Variation. *American Educational Research Journal*, 14, 493-498.
- Rogan, J. C., Keselman, H. J. & Breen, L.J. 1977. Assumption violations and rates of type I error for the Tukey multiple comparison test: A review and empirical investigation via a coefficient of Variation. *Journal of Experimental Education*, 46, 20-25.
- Rogan, J. C., Keselman, H. J. & Mendoza, J. L. 1979. Analysis of repeated measures. *British Journal of Mathematical and Statistical Psychology*, 32, 269-286.
- Romaniuk, J. G., Levin, J. R. & Hubert, L.J. 1977. Hypothesis testing procedures in repeated measures designs: On the road map not taken. *Child Development*, 48, 1757-1760.
- Rosenthal, R. 1969(a). Interpersonal expectations: Effects of the experimenter's hypothesis. In: Rosenthal, R. & Rosnow, R. C. (Eds): *Artifact in behavioral research*. New York: Academic Press, 181-277.
- Rosenthal, R. 1977. Biasing effects of experimenters. *ETC.*, 34, 253-264.
- Rosenthal, R. 1979. The „file drawer problem“ and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- Rosenthal, R. 1980. On telling tails when combining results of independent studies. *Psychological Bulletin*, 88, 496-497.
- Rosenthal, R. & Gaito, J. 1963. The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, 55, 33-38.
- Rosenthal, R. & Gaito, J. 1964. Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, 15, 570.
- Rosenthal, R. & Rosnow, R. L. (Eds). 1969(a). *Artifact in behavioral research*. New York: Academic Press.

- Rosenthal, R. & Rosnow, R. C. 1969(b). The volunteer subject. In: Rosenthal, R. & Rosnow, R. C. (Eds). 1969(a): Artifact in behavioral research. New York: Academic Press, 61-118.
- Rosenthal, R. & Rubin, D. B. 1978. Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 3, 377-415.
- Rosenthal, R. & Rubin, D. B. 1979. Comparing significance levels of independent studies. *Psychological Bulletin*, 86, 1165-1168.
- Rosnow, R. & Davis, D. J. 1977. Demand characteristics and the psychological experiment. *ETC.*, 34, 301-313.
- Rotton, J. & Schönemann, P. H. 1978. Power tables for analysis of variance. *Educational and Psychological Measurement*, 38, 213-229.
- Rouanet, H. & Lépine, D. 1970. Comparison between treatments in a repeated measures design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, 23, 147-163.
- Rüppell, H. 1977. Bayes-Statistik. Eine Alternative zur klassischen Statistik. *Archiv für Psychologie*, 129, 175-186.
- Rützel, E. 1979. Bayessches Hypothesentesten und warum die Bayesianer Bias-ianer heißen sollten. *Archiv für Psychologie*, 131, 211-232.
- Rützel, E. 1980. Korrektur zu E. Rützel: Bayessches Hypothesentesten und warum die Bayesianer Bias-ianer heißen sollten. *Archiv für Psychologie*, 132, 187-188.
- Rule, S. J. 1976. A general experimentwise error rate for multiple significance tests. *Perceptual and Motor Skills*, 43, 1263-1277.
- Rulon, P. J. & Brooks, W. D. 1968. On statistical tests of group differences. In: Whitla, D. K. (Ed.): *Handbook of measurement and assessment in behavioral sciences*. Reading, Mass.: Addison-Wesley, 60-99.
- Runkel, P. J. & McGrath, J. E. 1972. *Research on human behavior*. New York: Holt, Rinehart & Winston.
- Ryan, T. A. 1959. Multiple comparisons in psychological research. *Psychological Bulletin*, 56, 26-47.
- Ryan, T. A. 1962. The experiment as the unit for computing error rates. *Psychological Bulletin*, 59, 301-305.
- Ryan, T. A. 1980. Comment on „Protecting the Overall rate of type I errors for pairwise comparisons with an omnibus test statistic“. *Psychological Bulletin*, 88, 354-355.
- Sachdeva, D. 1973. Estimating strength of relationship in multivariate analysis of variance. *Educational and Psychological Measurement*, 33, 627-631.
- Sachs, L. 1968, 1974⁴. *Angewandte Statistik*. Berlin: Springer.
- Samiuddin, M., Hanif, M. & Asad, H. 1978. Some comparisons of the Bartlett and cube root tests of homogeneity of variance. *Biometrika*, 65, 218-221.
- Särndal, C. E. 1974. A comparative study of association measures. *Psychometrika*, 39, 165-187.

- Saniga, E. M. & Miles, J. A. 1979. Power of some goodness-of-fit tests of normality against asymmetric stable alternatives. *Journal of the American Statistical Association*, 74, 861-865.
- Sarris, V. 1968. Nicht-parametrische Trendanalysen in der klinisch-psychologischen Forschung. *Zeitschrift für experimentelle und angewandte Psychologie*, 15, 291-316.
- Schach, S. & Schäfer, T. 1978. *Regressions- und Varianzanalyse*. Berlin: Springer.
- Scheffé, H. 1959, 1961². *The analysis of variance*. New York: Wiley.
- Scheffé, H. 1970. Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association*, 65, 1501-1508.
- Scheffley, V. M. & Schmidt, W. H. 1978. Analysis of repeated measures data: A Simulation study. *Multivariate Behavioral Research*, 13, 347-362.
- Schlesselman, J. J. 1973. Data transformation in two-way analysis of variance. *Journal of the American Statistical Association*, 68, 369-378.
- Schmidt, F. L., Hunter, J. E. & Urry, V. W. 1976. Statistical power in criterion-related Validation studies. *Journal of Applied Psychology*, 61, 473-485.
- Schmidtke, A. & Jäger, R. 1976. Tabellen zur Überprüfung der Normalität von Schiefe und Exzeß. *Biometrische Zeitschrift*, 18, 413-418.
- Schrader, R. M. & McKean, J. W. 1977. Robust analysis of variance. *Communications in Statistics*, A6, 879-894.
- Schuler, H. 1980. *Ethische Probleme in der psychologischen Forschung*. Göttingen: Hogrefe.
- Schulman, J. L., Kupst, M. J. & Suran, B. G. 1976. The worship of „p“: Significant yet meaningless research results. *Bulletin of the Menninger Clinic*, 40, 134-143.
- Schwarz, H. 1975. *Stichprobenverfahren*. München: Oldenbourg.
- Schwarzer, R. & Steinhagen, K. (Hrsg.) 1975. *Adaptiver Unterricht. Zur Wechselwirkung von Schülermerkmalen und Unterrichtsmethoden*. München: Kösel.
- Scott, W. A. 1968. Attitude measurement. In: Lindzey, G. & Aronson, E. (Eds): *Handbook of Social Psychology*. Vol. II. Reading, Mass.: Addison-Wesley, 204-272.
- Searle, S. R. 1971(a). *Linear models*. New York: Wiley.
- Searle, S. R. 1971(b). Topics in variance component estimation. *Biometrics*, 27, 1-76.
- Selg, H. 1966, 1975⁴. *Einführung in die experimentelle Psychologie*. Stuttgart: Kohlhammer.
- Selg, H. & Bauer, W. 1971. *Forschungsmethoden der Psychologie*. Stuttgart: Kohlhammer.
- Shaffer, J. P. 1972. Directional statistical hypotheses and comparisons among means. *Psychological Bulletin*, 77, 195-197.
- Shaffer, J. P. 1973. Defining and testing hypotheses in multidimensional contingency tables. *Psychological Bulletin*, 79, 127-141.

- Shaffer, J. P. 1974(a). Multiple comparisons with unequal sample sizes. *Psychological Reports*, 35, 572-574.
- Shaffer, J. P. 1974(b). Bidirectional unbiased procedures. *Journal of the American Statistical Association*, 69, 437-439.
- Shaffer, J. P. & Gillo, M. W. 1974. A multivariate extension of the correlation ratio. *Educational and Psychological Measurement*, 34, 521-524.
- Shapiro, S. S., Wilk, M. B. & Chen, H. J. 1968. Comparative study of various tests for normality. *Journal of the American Statistical Association*, 63, 1343-1372.
- Sheridan, C. L. 1971. *Fundamentals of experimental psychology*. New York: Holt, Rinehart & Winston.
- Shuster, J. J. & Boyett, J. M. 1979. Nonparametric multiple comparison procedures. *Journal of the American Statistical Association*, 74, 379-382.
- Siebel, W. 1965. *Die Logik des Experimentierens in den Sozialwissenschaften*. Berlin: Duncker & Humboldt.
- Siegel, S. 1956. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill. Dt.: *Nichtparametrische statistische Methoden*. Frankfurt am Main: Fachbuchhandlung für Psychologie, 1976.
- Silverman, I. 1977. *The human subject in the psychological laboratory*. New York: Pergamon Press.
- Silverstein, A. B. 1974. Relations between analysis of variance and its nonparametric analogs. *Psychological Reports*, 34, 331-333.
- Silverstein, A. B. 1978. Critical values for nonparametric multiple comparisons. *Psychological Reports*, 43, 44-16.
- Singer, B. 1979. Distribution-free methods for non-parametric problems. A classified and selected bibliography. *British Journal of Mathematical and Statistical Psychology*, 32, 1-60.
- Skipper, J. K. Jr., Guenther, A. L. & Nass, G. 1967. The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist*, 2, 16-18.
- Smart, R. G. 1964. The importance of negative results in psychological research. *Canadian Psychologist*, 5, 225-232.
- Smart, R. G. 1966. Subject selection bias in psychological research. *Canadian Psychologist*, 7a, 115-121.
- Smith, J. E. K. 1976(a). Data transformations in analysis of variance. *Journal of Verbal Learning and Verbal Behavior*, 339-346.
- Smith, J. E. K. 1976(b). Analysis of qualitative data. *Annual Review of Psychology*, 27, 487-499.
- Smith, J. L. 1972. The eta coefficient in MANOVA. *Multivariate Behavioral Research*, 7, 361-372.
- Smith, R. A. 1971. The effect of unequal group size on Tukey's HSD procedure. *Psychometrika*, 36, 31-34.

- Snedecor, G. W. & Cochran, W. G. 1972. Statistical methods. Ames, Iowa: The Iowa State University Press.
- Soderquist, D. R. & Hussian, R. A. 1978. The Utility of Utility indices. *Bulletin of the Psychonomic Society*, 11, 136-138.
- Solomon, R. L. 1949. An extension of control group design. *Psychological Bulletin*, 46, 137-150.
- Som, R. K. 1973. A manual of sampling techniques. London: Heinemann.
- Spector, P. E. 1977. What to do with significant multivariate effects in multivariate analyses of variance? *Journal of Applied Psychology*, 62, 158-163.
- Spielman, S. 1974. The logic of tests of significance. *Philosophy of Science*, 41, 211-225.
- Spielman, S. 1978. Statistical dogma and the logic of significance testing. *Philosophy of Science*, 45, 120-135.
- Spjotvoll, E. 1974. Multiple testing in the analysis of variance. *Scandinavian Journal of Statistics*, 1, 97-114.
- Spjotvoll, E. & Stoline, M. R. 1973. An extension of the T-method of multiple comparison to include the cases with unequal sample sizes. *Journal of the American Statistical Association*, 68, 975-978.
- Sprott, D. A. 1970. Note on Evans and Anastasio on the analysis of covariance. *Psychological Bulletin*, 73, 303-306.
- Stanley, J. C. 1973. Designing psychological experiments. In: Wolman, B. B. (Ed.): *Handbook of general psychology*. Englewood Cliffs, N. J.: Prentice Hall, 90-106.
- Stavig, G. R. & Acock, A. C. 1980. Coefficients of association analogous to Pearson's r for nonparametric statistics. *Educational and Psychological Measurement*, 40, 679-685.
- Steffens, F. E. 1970. Power of bivariate studentized maximum and minimum modulus tests. *Journal of the American Statistical Association*, 65, 1639-1644.
- Steger, J. A. (Ed.) 1971. *Readings in statistics*. New York: Holt, Rinehart & Winston.
- Stegmüller, W. 1973, 1974. *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie*. Band 1: Wissenschaftliche Erklärung und Begründung, 1969b, 1974(a). Band 2: Theorie und Erfahrung. 1. Halbband: Begriffsformen, Wissenschaftssprache, empirische Signifikanz und theoretische Begriffe, 1974(b). 2. Halbband: Theoriestrukturen und Theoriedynamik, 1973(c). Band 4: Personelle und statistische Wahrscheinlichkeit. 1. Halbband: Personelle Wahrscheinlichkeit und Rationale Entscheidung, 1973(a). 2. Halbband: Statistisches Schließen. Statistische Begründung. Statistische Analyse, 1973(b). Berlin: Springer.
- Stegmüller, W. Bd. 1: 1978⁶, Bd. 2: 19796(a). *Hauptströmungen der Gegenwartsphilosophie*. Bd. 1 und 2. Stuttgart: Kröner.
- Stegmüller, W. 1979(b). *The structuralist view of theories*. Berlin: Springer.
- Stegmüller, W. 1980. *Neue Wege der Wissenschaftsphilosophie*. Berlin: Springer.

- Stein, C. 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16, 243-258.
- Steinfatt, T. M. 1977. Measurement, transformations, and the real world: Do the numbers represent the concept? *ETC.*, 34, 277-289.
- Sterling, T. D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, 54, 30-34.
- Sterling, T. D. 1960. What is so peculiar about accepting the null-hypothesis? *Psychological Reports*, 7, 363-364.
- Stevens, J. P. 1972(a). Global measures of association in multivariate analysis of variance. *Multivariate Behavioral Research*, 7, 373-378.
- Stevens, J. P. 1972(b). Four methods of analyzing between Variation for the k group Manova Problem. *Multivariate Behavioral Research*, 7, 499-501.
- Stevens, J. P. 1979. Comment on Olson: Choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 86, 355-360.
- Stevens, J. P. 1980. Power of the multivariate analysis of variance tests. *Psychological Bulletin*, 88, 728-737.
- Stevens, S. S. 1951. Mathematics, measurement, and psychophysics. In: Stevens, S. S. (Ed.): *Handbook of experimental psychology*. New York: Wiley, 1-49.
- Steyer, R. 1977, 1979. Untersuchungen zur nonorthogonalen Varianzanalyse. Göttingen: Diplomarbeit. Weinheim: Beltz.
- Stilson, D. W. 1966. Probability and statistics in psychological research and theory. San Francisco: Holden-Ray.
- Student (W. S. Gosset). 1908. The probable error of the mean. *Biometrika*, 6, 1-25.
- Subkoviak, M. J. & Levin, J. R. 1977. Fallibility of measurement and the power of a statistical test. *Journal of Educational Psychology*, 14, 47-52.
- Summers, G. F. (Ed.) 1970. Attitude measurement. Chicago: Rand McNally.
- Suppe, F. 1974, 1977²(a). The search for philosophic understanding of scientific theories. In: Suppe, F. (Ed.): *The structure of scientific theories*. Urbana, Ill.: University of Illinois Press, 3-232.
- Suppe, F. 1977²(b). Afterword - 1977. In: Suppe, F. (Ed.): *The structure of scientific theories*. Urbana, Ill.: University of Illinois Press, 617-630.
- Suppes, P. 1970. A probabilistic theory of causality. Amsterdam: North-Holland.
- Suppes, P. & Zinnes, J. L. 1963. Basic measurement theory. In: Luce, R. D., Bush, R. R. & Galanter, E. (Eds): *Handbook of mathematical psychology*. Vol. 1. New York: Wiley, 1-76.
- Sutcliffe, J. P. 1958. Error of measurement and the sensitivity of a test of significance. *Psychometrika*, 23, 9-17.
- Sutcliffe, J. P. 1980. On the relationship of reliability to statistical power. *Psychological Bulletin*, 88, 509-515.

- Swaminathan, H. & DeFriesse, F. 1979. Detecting significant contrasts in analysis of variance. *Educational and Psychological Measurement*, 39, 39-44.
- Talwar, P. P. & Gentle, J. E. 1977. A robust test for the homogeneity of scales. *Communications in Statistics*, A6, 363-369.
- Tamhane, A. C. 1977. Multiple comparisons in model I one-way ANOVA with unequal variances. *Communications in Statistics*, A6, 15-32.
- Tamhane, A. C. 1979. A comparison of procedures for multiple comparisons of means with unequal variances. *Journal of the American Statistical Association*, 74, 471-480.
- Tang, P. C. 1938. The power function of the analysis of variance tests with tables and illustrations of their use. *Statistical Research Memoirs*, 2, 126-149.
- Tatsuoka, M. 1969. Multivariate analysis. *Review of Educational Research*, 39, 739-743.
- Tatsuoka, M. M. 1971. *Multivariate analysis*. New York: Wiley.
- Thomas, D. A. H. 1973. Multiple comparisons among means. A review. *The Statistician*, 22, 16-42.
- Tiku, M. L. 1971. Power function of the F-test under non-normal situations. *Journal of the American Statistical Association*, 66, 913-916.
- Tiku, M. L. 1975. A new statistic for testing suspected outliers. *Communications in Statistics*, 4, 737-752.
- Timaeus, E. 1974. *Experiment und Psychologie. Zur Sozialpsychologie psychologischen Experimentierens*. Göttingen: Hogrefe.
- Timaeus, E. 1975. Untersuchungen im Laboratorium. In: Koolwijk, J. v. & Wieken-Mayser, M. (Hrsg.): *Techniken der empirischen Sozialforschung*. Bd. 2: Untersuchungsformen. München: Oldenbourg, 195-229.
- Toothaker, L. E. 1971. The effect of the joint Violation of two assumptions on the size of the F-test under Permutation for the randomized block design. *Journal of Statistical Computation and Simulation*, 1, 55-64.
- Toothaker, L. E. 1972. An empirical investigation of the Permutation t-test. *British Journal of Mathematical and Statistical Psychology*, 25, 83-94.
- Trachtman, J. N., Giambalvo, V. & Dippner, R. S. 1978. On the assumptions concerning the assumptions of a t test. *The Journal of General Psychology*, 99, 107-116.
- Traxel, W. 1974. *Grundlagen und Methoden der Psychologie*. Bern: Huber.
- Treiber, B. 1977. Untersuchungen zur Interaktion von Lehrmethode und Schülermerkmale: Reanalyse der Teststärke. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 90, 29-35.
- Treiber, B. & Petermann, F. 1976. Zur Interaktion von Lernermerkmalen und Lehrmethoden: Rekonstruktion und Normierung des ATI-Forschungsprogramms. Heidelberg: Bericht No. 4, aus dem Institut für Psychologie.
- Treinius, G. 1977. Teststärkeanalysen für 61 Untersuchungen zum Zusammenhang

- zwischen Lehrerverhalten und Schülerleistung. Zeitschrift für erziehungswissenschaftliche Forschung, 11, 50-63.
- Trickett, W. H. & Welch, B. L. 1954. On the comparison of two means, further discussion of iterative methods for calculation tables. *Biometrika*, 41, 361-374.
- Trickett, W. H., Welch, B. L. & James, G. S. 1956. Further critical values for the two-means problem. *Biometrika*, 43, 203-205.
- Tukey, J. W. 1949. One degree of freedom for nonadditivity. *Biometrics*, 5, 232-242.
- Tukey, J. W. 1957. On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, 28, 602-632.
- Tukey, J. W. 1977. *Exploratory data analysis*. Reading, Mass.: Addison-Wesley.
- Tukey, J. W. & McLaughlin, D. H. 1963. Less vulnerable confidence and significance procedures for location based on a single sample Trimming/Winsorization. 1. *Sankhyā*, A25, 331-352.
- Underwood, B. J. 1957. *Psychological research*. New York: Appleton-Century-Crofts.
- Underwood, B. J. 1966². *Experimental psychology*. New York: Appleton-Century-Crofts.
- Underwood, B. J. & Shaughnessy, J. J. 1975. *Experimentation in psychology*. New York: Wiley.
- Upton, G. J. G. 1978. *The analysis of cross-tabulated data*. Chichester: Wiley.
- Ury, H. K. 1976. A comparison of four procedures for multiple comparisons among means pairwise contrasts for arbitrary sample sizes. *Technometrics*, 18, 89-97.
- Ury, H. K. & Wiggins, A. D. 1971. Large sample and other multiple comparisons among means. *British Journal of Mathematical and Statistical Psychology*, 24, 174-194.
- Ury, H. K. & Wiggins, A. D. 1974. Use of the Bonferroni inequality for multiple comparisons among means with post hoc contrasts. *British Journal of Mathematical and Statistical Psychology*, 27, 176-178.
- Ury, K. H. & Wiggins, A. D. 1975. A comparison of three procedures for multiple comparisons among means. *British Journal of Mathematical and Statistical Psychology*, 28, 88-102.
- Vatza, E. J., Byatt, S. E., Kay, K. J., Kerchner, M., Richter, M. L. & Seay, M. B. 1980. Comment on „Combining results of independent studies“. *Psychological Bulletin*, 88, 494-495.
- Vaughan, G. M. & Corballis, M. C. 1969. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, 72, 204-213.
- Venables, W. 1975. Calculation of confidence intervals for noncentrality Parameters. *Journal of the Royal Statistical Society, Series B*, 37, 406-412.
- Verreck, W. A. (Ed.) 1974. *Methodological problems in research and development in higher education*. Amsterdam: Swets & Zeitlinger.

- Wainer, H. 1972. A practical note on one-tailed tests. *American Psychologists*, 27, 775-776.
- Wainer, H. 1973. The other tail. *British Journal of Mathematical and Statistical Psychology*, 26, 182-187.
- Wainer, H. 1976. Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics*, 1, 285-312.
- Wainer, H. & Thissen, D. 1976. Three steps toward robust regression. *Psychometrika*, 41, 9-34.
- Wainer, H. & Thissen, D. 1981. Graphical data analysis. *Annual Review of Psychology*, 32, 191-241.
- Wald, A. 1947, 1952³. *Sequential analysis*. New York: Wiley.
- Wald, A. 1950. *Statistical decision functions*. New York: Wiley.
- Wallenstein, S. & Fleiss, J. L. 1979. Repeated measurements analysis of variance when the correlations have a certain pattern. *Psychometrika*, 44, 229-233.
- Walsh, J. E. 1962, 1965, 1968. *Handbook of nonparametric statistics*. Vol. 1: 1962, Vol. II: 1965, Vol. III: 1968. Princeton, N. Y.: Van Nostrand.
- Wang, Y. Y. 1967. A comparison of several variance component estimators. *Biometrika*, 54, 301-305.
- Wang, Y. Y. 1971. Probabilities of the Type I errors of the Welch test for the Behrens-Fisher problem. *Journal of the American Statistical Association*, 66, 605-608.
- Webb, E. J., Campbell, D. T., Schwartz, R. D. & Sechrest, L. 1966. *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally. Dt.: *Nichtreaktive Meßverfahren*. Weinheim: Beltz, 1975.
- Weber, E. 1967⁶. *Grundriß der biologischen Statistik*. Stuttgart: Gustav Fischer Verlag.
- Weber, S. J. & Cook, T. D. 1972. Subjects effects in laboratory research: An examination on subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, 77, 273-295.
- Wegman, E. J. & Carroll, R. J. 1977. A Monte Carlo study of robust estimators of location. *Communications in Statistics*, A6, 795-812.
- Welch, B. L. 1947. The generalization of „Student's“ problem when several populations are involved. *Biometrika*, 34, 28-35.
- Welch, B. L. 1949. Further note on Mrs. Aspin's tables and on certain approximations to the tabled function. *Biometrika*, 36, 293-296.
- Welch, B. L. 1951. On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330-336.
- Werts, C. E. & Linn, R. L. 1971. Problems with inferring treatment effects from repeated measures. *Educational and Psychological Measurement*, 31, 857-866.
- Westermann, R. 1980. Die empirische Überprüfung des Niveaus psychologischer Skalen. *Zeitschrift für Psychologie*, 188, 450-468.

- Westermann, R. 1982. Zur Messung von Einstellungen auf Intervallskalenniveau. *Zeitschrift für Sozialpsychologie*, 13, 97-108.
- Westermann, R. Zur Wahl der Fehlerwahrscheinlichkeiten bei mehrfachen Signifikanztests. In: Lüer, G. (Hrsg.): Bericht über den 33. Kongreß der Deutschen Gesellschaft für Psychologie in Mainz 1982. Göttingen: Hogrefe (im Druck, a).
- Westermann, R. Zur empirischen Überprüfung des Skalenniveaus von individuellen Einschätzungen und Ratings. *Zeitschrift für Psychologie* (im Druck, b).
- Westermann, R. & Hager, W. 1982. Entscheidung über statistische und wissenschaftliche Hypothesen: Zur Differenzierung und Systematisierung der Beziehungen. *Zeitschrift für Sozialpsychologie*, 13, 13-21.
- Westmeyer, H. 1982. Wissenschaftstheoretische Aspekte der Feldforschung. In: Patry, J. L. (Hrsg.): Laborforschung/Feldforschung. Bern: Huber, 67-84.
- Wetherill, G. B. 1960. The Wilcoxon test and non-null hypotheses. *Journal of the Royal Statistical Society, Series B*, 22, 402-418.
- Wetherill, G. B. 1966, 1975². *Sequential methods in statistics*. London: Chapman and Hall.
- Wetzel, W. (Hrsg.) 1970. Neue Entwicklungen auf dem Gebiet der Zeitreihenanalyse. Göttingen: Vandenhoeck & Ruprecht.
- Whitney, D. R. & Feldt, L. S. 1973. Analyzing questionnaire results: Multiple tests of hypotheses and multivariate hypotheses. *Educational and Psychological Measurement*, 33, 365-380.
- Wike, E. L. & Church, J. D. 1977. Further comments on nonparametric multiple-comparison tests. *Perceptual and Motor Skills*, 45, 917-918.
- Wike, E. L. & Church, J. D. 1978. A Monte Carlo investigation of four nonparametric multiple-comparison tests for k independent groups. *Bulletin of the Psychonomic Society*, 11, 25-28.
- Wildt, A. R. & Ahtola, O. T. 1976. *Analysis of covariance*. Beverly Hills: Sage.
- Wilk, M. B. & Kempthorne, O. 1955. Fixed, mixed and random models. *Journal of the American Statistical Association*, 50, 1144-1167.
- Wilk, M. B. & Kempthorne, O. 1957. Nonadditives in a Latin square design. *Journal of the American Statistical Association*, 52, 218-236.
- Williams, J. D. 1974. A simplified regression formulation of Tukey's test. *Journal of Experimental Education*, 42, 4, 80-82.
- Wilson, K. 1975. The sampling distributions of conventional conservative and corrected F-ratios in repeated measurements designs with heterogeneity of covariance. *Journal of Statistical Computation and Simulation*, 3, 201-215.
- Wilson, R. S. 1975. Analysis of developmental data: Comparison among alternative models. *Developmental Psychology*, 11, 676-680.
- Winer, B. J. 1962, 1970, 1971². *Statistical principles in experimental design*. Tokyo: McGraw-Hill Kogakuska.

- Winne, D. 1968. Zur Planung von Versuchen: Wieviel Versuchseinheiten? *Arzneimittelforschung*, 18, 1611-1618.
- Wishart, J. 1932. A note on the distribution of the correlation ratio. *Biometrika*, 24, 441-456.
- Witte, E. H. 1977. Zur Logik und Anwendung der Inferenzstatistik. *Psychologische Beiträge*, 19, 290-303.
- Witte, E. H. 1978. Korrelationsstudien und Experimente: Ihre Kombination und ihre Aussagekraft. Hamburg: Arbeiten aus den Psychologischen Instituten der Universität Hamburg, Nr. 46.
- Witte, E. H. 1980. Signifikanztest und statistische Inferenz. Stuttgart: Enke.
- Wooding, W. M. 1969. The computation and use of residuals in the analysis of experimental data. *Journal of Quality Technology*, 1, 175-188, 299.
- Woodward, A. J. & Overall, J. E. 1975. Multivariate Analysis of variance by multiple regression methods. *Psychological Bulletin*, 82, 21-32.
- Woodward, J. A. & Overall, J. E. 1976. A Computer program for calculating power of the F-test. *Educational and Psychological Measurement*, 36, 165-168.
- Wormser, R. 1974. *Experimentelle Psychologie*. München: Reinhardt.
- Wottawa, H. 1974. Das „allgemeine lineare Modell“ - ein universelles Auswertungsverfahren. *EDV in Medizin und Biologie*, 3, 65-73.
- Wottawa, H. 1980. *Grundriß der Testtheorie*. München: Juventa.
- Wundt, W. 1896, 1913¹¹. *Grundriß der Psychologie*. Leipzig: Engelmann.
- Yates, F. 1933. The principles of orthogonality and confounding in replicated experiments. *Journal of Agricultural Science*, 23, 108-145.
- Young, R. K. & Veldman, D. J. 1963. Heterogeneity and skewness in analysis of variance. *Perceptual and Motor Skills*, 16, 588.
- Zimmermann, E. 1972. *Das Experiment in den Sozialwissenschaften*. Stuttgart: Teubner.
- Zimny, G. H. 1961. *Method in experimental psychology*. New York: Ronald Press.