Kurt Heller/Bernhard Rosemann
Unter Mitarbeit von Anne-Katrin Gaedike

Planung und Auswertung empirischer Untersuchungen

Eine Einführung für Pädagogen, Psychologen und Soziologen

Ernst Klett Verlag Stuttgart

1. Auflage 1974

Printed in Germany
Satz: Hagedorn, Berlin
Druck: Röck, Weinsberg
Einbandgestaltung: Werbeagentur Lorenz, München
ISBN 3-12-923400-4

Vorwort

Die Fortschritte in den heutigen Sozialwissenschaften (Pädagogik, Psychologie, Soziologie) sind ohne empirische Fundierung ihrer Erkenntnisse kaum denkbar. Wissenschaftliche Ergebnisse können freilich nicht mehr wert sein als die Methoden, mit deren Hilfe sie gewonnen wurden. Sicherlich hat man den Methoden wissenschaftlichen Tuns nicht erst seit heute Bedeutung beigemessen, doch ist ein breiteres Methodenbewußtsein – zumindest im deutschsprachigen Raum – verhältnismäßig spät (bezogen auf die hier angesprochenen Disziplinen eigentlich erst innerhalb der letzten 10 bis 15 Jahre) zu beobachten. Hierfür lassen sich u. E. mehrere Motive anführen. Neben der stärkeren Zuwendung zur außereuropäischen Forschung, insbesondere angloamerikanischer Provenienz, messen wir der zunehmenden Einsicht in die Notwendigkeit, wissenschaftliche Erkenntnisse stärker als bisher der Praxis zugänglich zu machen, hierbei die größte Bedeutung zu. Die Verbindlichkeit wissenschaftlicher Aussagen im Hinblick auf die Anwendungspraxis hängt ja zuallererst von der Redlichkeit der *Methoden* ab, d.h. ihrer Objektivität, Zuverlässigkeit und Gültigkeit. Somit kann sich weder der Theoretiker (Wissenschaftler) noch der Praktiker (Lehrer, Schulpsychologe, Bildungsberater u.a.) der Forderung nach eingehender Information über die grundlegenden Forschungsmethoden seines Faches heute mehr entziehen. Erst recht gilt diese Forderung für Studierende der Sozialwissenschaften, die ja eines Tages selbst entscheidend mit dazu beitragen sollen, wissenschaftliche Erkenntnisse auf die Praxis zu übertragen. Fundierte Methodenkenntnisse sind also für alle, die sich mit Psychologie, Pädagogik bzw. Erziehungswissenschaft(en) und Soziologie befassen, unabdingbare Voraussetzung ihrer Beschäftigung. Dies gilt nicht allein für das Verständnis einschlägiger Publikationen, sondern für eine angemessene Bewertung wissenschaftlicher Ergebnisse überhaupt.

Gehören Methodikveranstaltungen somit unbestreitbar zum Fundamentum (sozial)wissenschaftlicher Studienpläne, so bedeutet andererseits die Auseinandersetzung mit diesem Gegenstand erfahrungsgemäß für sehr viele – nicht nur für Studierende – eine unverhältnismäßig große Kraftanstrengung, worunter nicht selten (durchaus vorhandene) Interessen und (berechtigte) individuelle Ansprüche leiden. Die Verfasser dieses Lehrbuches haben sich deshalb zum Ziel gesetzt, den oft als schwierig eingestuften Lehrstoff didaktisch so aufzubereiten, daß er von jedem Leser mit „Hochschulreife", also ohne spezifische Vorkenntnisse (etwa wissenschaftstheoretischer oder mathematischer Art), verstanden werden kann. Darüber hinaus sollte der Leser möglichst umfassend über die einschlägigen Methodenprobleme informiert werden.

Die ausführliche Behandlung von wissenschaftstheoretischen Grundlagen und Methoden der Versuchsplanung vs. Versuchsausführung und statistischen Auswertung in *einem* Lehrbuch ist nach unserer Erfahrung vorteilhaft, wenngleich bisher (zumindest im deutschen Sprachraum) ohne Vorbild. Das Verständnis für die Zusammenhänge wissenschaftstheoretischer Voraussetzungen im Hinblick auf empirisch-operationale und statistische Vorgehensweisen dürfte dadurch wesentlich erleichtert werden. Zugleich wird hiermit dem Vorwurf, empirische

Forschung entbehre allzuoft einer ausreichenden theoretischen Reflexion, wirksam begegnet.
Unserem didaktischen Grundanliegen entsprechend waren wir um eine verständliche Diktion bemüht und verzichteten weitgehend auf die theoretische Herleitung einzelner Verfahren (Formelableitungen). Dem gleichen Zweck dienen die im Anhang 3 beigefügten Übersichtstafeln, die auch dem Anfänger eine rasche Orientierung und fehlerfreie Anwendung einschlägiger Verfahren der Korrelationsrechnung und Inferenzstatistik ermöglichen sollen. Zahlreiche Rechen- und Übungsbeispiele (mit den Lösungen im Anhang 1) erleichtern die selbständige, kontrollierte Erarbeitung der besprochenen Verfahren. Darüber hinaus gewährt der umfangreiche Tabellenteil im Anhang 2 dieses Buches, der auch weithin unzugängliche Tabellen (z. B. für den $2\hat{I}$-Test von Kullback) enthält, die erforderlichen Voraussetzungen für die praktische empirische Arbeit, ohne daß man im Regelfalle auf weitere Quellen angewiesen ist. Im detaillierten Stichwortverzeichnis am Ende des Bandes findet der Leser noch einmal die wichtigsten im Text erläuterten Begriffe mit den entsprechenden Seitenangaben. Insgesamt repräsentiert das Buch einen Stoffumfang, der etwa in einer zweisemestrigen Folge von jeweils zwei bis drei Wochenstunden zu bewältigen ist.
Der Leser mag nun selbst urteilen, ob die Verfasser seinen – und ihren – Ansprüchen gerecht wurden. Wenngleich mehrjährige Lehrerfahrungen und damit verbunden wertvolle Anregungen seitens der Übungsteilnehmer auf die Gestaltung des Buches entscheidenden Einfluß hatten, so ist das Autorenteam am Ende keineswegs mit sich und seinem Werk so zufrieden, daß es nicht für kritische Hinweise jedweder Art dankbar wäre. Vor allem der unvermeidliche Zeitdruck, unter dem angesichts der wachsenden Arbeitsbelastung an der Hochschule Vorhaben wie dieses realisiert werden müssen, ließ den Verfassern nicht die Muße, die sie sich selbst für die Niederschrift gerne gewünscht hätten. Andererseits bedeutete die drängende Nachfrage vieler Studenten nach diesem Buch Anregung und Ermutigung zugleich.
Daß das seit über drei Jahren geplante Vorhaben in der vorliegenden Form verwirklicht werden konnte, ist zuallererst der bereitwilligen und sachkundigen Mitarbeit der beiden Co-Autoren zu danken. Herr Dr. B. Rosemann übernahm die umfangreichen Kapitel 4.1 (Korrelationsrechnung) und 5.4 (Nonparametrische Verfahren), Frau Dipl.-Psych. A.-K. Gaedike die Kapitel 4.2 (Regression) und 5.3 (Parametrische Verfahren), die übrigen Teile wurden vom ersten Autor bearbeitet, der die Verantwortung für das Gesamtkonzept und etwaige Fehler allein trägt. Das Stichwortverzeichnis wurde von Fräulein Dipl.-Psych. R. Bonn und Herrn Dipl.-Psych. M. Schneider erstellt, die auch beim Korrekturlesen mithalfen. Die technische Herstellung des schwierigen Manuskripts lag in den Händen von Fräulein G. Klein und Herrn cand. päd. W. Büttner. Ihnen allen sei an dieser Stelle herzlich gedankt.

Bonn, im August 1973 Die Verfasser

Inhaltsverzeichnis

Einführung in die Forschungsstatistik 91

Einführung in die Wissenschaftsmethodik

1. Wissenschaftstheoretische Voraussetzungen

1.1. Zum Wissenschaftsbegriff

Nach A.Comte (1798–1857) vollzog sich die historische Entwicklung wissenschaftlichen Denkens in drei Stadien: vom mythologischen (theologischen) über das metaphysische (philosophisch-spekulative) zum positivistischen (empiriewissenschaftlichen) Stadium. Ohne hier in eine wissenschaftskritische Betrachtung dieser Stufenlehre einzutreten, sei doch festgehalten, daß die – vermeintliche – Antinomie von Spekulation und Empirismus bis auf den heutigen Tag in der wissenschaftstheoretischen Diskussion spürbar ist. Besonders augenfällig werden entsprechende Kontroversen in der Pädagogik bzw. den Sozialwissenschaften (Erziehungswissenschaft, Psychologie, Soziologie) ausgetragen, also in jenen Disziplinen, die sich von Fall zu Fall eher „geisteswissenschaftlich" versus „naturwissenschaftlich" verstehen oder wenigstens so gebärden. Es ist das Verdienst von Immanuel Kant (1724–1804), als einer der ersten die Notwendigkeit der Balance zwischen spekulativem (rationalem) und empirischem (erfahrungswissenschaftlichem bzw. operationalem) Ansatz erkannt und in seinen Schriften (Kritik der reinen Vernunft) vertreten zu haben. Seine Grundideen tauchen in den neueren Wissenschaftsdefinitionen fast immer auf, so auch in der Definition von C. F. Graumann (1965): „*Wissenschaft* ist ein System eindeutig formulierter Erkenntnisse, die so begründet sind, daß jeder Schritt – jedenfalls prinzipiell – nachvollziehbar ist."

Für das, was wir fortan als „Wissenschaft" bezeichnen, sind demnach folgende Bestimmungsstücke notwendig: 1. Wissenschaft eignet *System*charakter; 2. wissenschaftliche Erkenntnisse, Aussagen, Sätze usw. müssen *eindeutig formuliert* (die Kriterien einer hierzu notwendigen Wissenschaftssprache werden uns noch eingehend beschäftigen) und somit *mitteilbar* sein; 3. die *Ergebnisse* wissenschaftlichen Tuns müssen stets *kontrollierbar* sein, womit die *Angebbarkeit der Methode,* mit deren Hilfe wissenschaftliche Aussagen gewonnen wurden, als vielleicht wichtigstes Postulat im modernen Wissenschaftsverständnis angesprochen wäre (die Erörterung des Operationismus wird uns die Bedeutung dieses Kriteriums noch näherbringen). Damit wären unserer Meinung nach die Hauptkennfunktionen von „Wissenschaft" umrissen, d.h. die Kriterien benannt, die eine Unterscheidung zwischen wissenschaftlichem und nichtwissenschaftlichem Tun prinzipiell erlauben.

Um Mißverständnissen vorzubeugen, sei hier angemerkt, daß mit einer solchen Bestimmung nicht notwendig auch eine Wertung verknüpft ist, etwa darüber, ob wissenschaftliches Handeln „wertvoller" sei als nichtwissenschaftliches. Die genannten Kriterien dienen zunächst lediglich der Unterscheidung von „wissenschaftlichem" und „nichtwissenschaftlichem" Tun oder „wissenschaftlichen" und „nichtwissenschaftlichen" Aussagen. Daß sich wissenschaftliche Erkenntnisse, Feststellungen, Sätze usw. praktisch in vielen Fällen nichtwissenschaftlichen

Erkenntnissen, „Behauptungen“ usw. als überlegen erweisen, entspricht zwar allgemeinen Erfahrungen (wiewohl prinzipiell auch der umgekehrte Fall denkbar und – gelegentlich – zu beobachten ist), impliziert jedoch eine andere Fragestellung. Letztere Überlegungen hängen eng mit dem sog. Relevanzproblem wissenschaftlicher Forschung zusammen, das uns in einem anderen Zusammenhang noch interessieren wird (s. S. 64f.).
Nach der mehr oder weniger allgemein verbindlichen Definition von Wissenschaft beschäftigt uns im Hinblick auf die in diesem Buch angesprochenen Wissenschaftsdisziplinen jetzt die Frage, ob und gegebenenfalls wodurch sich die einzelnen Wissenschaften voneinander unterscheiden. Zu den aufgeführten – ubiquitären – Wissenschaftskriterien müßten dann weitere Kriterien zur Unterscheidung der Wissenschaftsdisziplinen selbst benannt werden. Diese liegen nach allgemeiner Auffassung im *Gegenstand* und in der *Methode* jeder Einzelwissenschaft begründet. Am Beispielfall der Psychologie sei dies kurz erläutert.

1.2. Gegenstand und Methode in der Wissenschaft

Solange die Psychologie Anhängsel der Philosophie war, konnte man als Gegenstand der „Psychologie“ qua „Seelenkunde“ die Psyche oder Seele (*psyche* heißt ursprünglich „Hauch“, „Atem“ und symbolisiert als Prinzip des Lebens die „Seele“) benennen. Der ideengeschichtliche Hintergrund der Psychologie wurde in dem lesenswerten Buch von Pongratz (1967) dargestellt. Demnach hat sich die Psychologie erst sehr spät emanzipiert und als selbständige Wissenschaftsdisziplin ausgewiesen; gewöhnlich wird als Datum hierfür das Jahr 1879 angeführt, in dem der aus Mannheim stammende Mediziner und Philosoph Wilhelm Wundt das erste Psychologische Institut (Laboratorium) in Leipzig begründete. Fortan fungieren Wahrnehmung, Denken, Intelligenz, Wille oder bedingte Reaktion, Lernen, Gedächtnis u.ä. Funktionen als Inhalte einer zunehmend empirisch und experimentell ausgerichteten Psychologie. Im Grunde werden damit freilich nur verschiedene Aspekte des Gegenstandes der Psychologie akzentuiert, den wir heute mit „Erleben“ und „Verhalten“ zu umschreiben gewohnt sind.
Freilich verliert mit dieser „Klammer“ der Dualismus, der sich schon bei Plato (427–347 v. Chr.) im sog. Leib-Seele-Problem ankündigte und durch Descartes' (1596–1650) Unterscheidung von „res cogitans“ und „res extensa“ (Trennung nach Bewußtseinsprozessen und leiblichen Vorgängen, die in Wechselwirkung zueinander stehend gedeutet werden) sowie die Leibnizsche (1646–1716) Prädestinationslehre (Lehre von der prästabilierten Harmonie, d.h. einer von Anfang an gegebenen Leib-Seele-Parallelität) nachhaltigen Einfluß auf die neuere Psychologie nahm, kaum von seiner Widersprüchlichkeit. Die Gegenstandsdefinition der heutigen Psychologie umreißt nach wie vor ein höchst komplexes Geschehen. In diesem Sinne schreiben etwa Thomae & Feger (1969) in ihrer Einführung in die Psychologie:

„Die Psychologie als Wissenschaft … *möchte menschliches Verhalten und Erleben möglichst angemessen erfassen, d.h. es nach Konstanz und Veränderlichkeit beschreiben und wenn möglich messen, die Bedingungen von Konstanz und Veränderlichkeit feststellen und den künftigen Verlauf, soweit es geht, vorhersagen.*

Diese Umschreibung kennzeichnet zunächst die Psychologie als eine ‚empirische' Wissenschaft, d.h. eine solche, die ihre Aussagen auf methodisch abgesicherte Erfahrung stützen möchte. Die Kennzeichnung des Gegenstandes nimmt auf jene Argumente Rücksicht, denen zufolge die ‚Seele' als ein metaphysischer Begriff nicht Gegenstand empirisch-wissenschaftlicher Erkenntnis sein könne. Sie sucht die Standpunkte jener miteinander zu verbinden, welche entweder das ‚bewußte Erleben' (vgl. Descartes; d. Verf.) oder das ‚Verhalten' (vgl. Behaviorismus; d. Verf.) zum Gegenstand der Psychologie erklären. Die gegebene Umschreibung vereinigt in sich die Zielsetzung einer deskriptiven Erfassung mit denen einer naturwissenschaftlichen Erklärung" (S. 1f.).

Der Wissenschaftsbestimmung eröffnen sich grundsätzlich zwei Möglichkeiten. 1. Die Definition der Wissenschaft kann vom *Gegenstand* her erfolgen (*problem-centered* im Sinne Maslows). Demnach resultieren die Unterschiede der Einzelwissenschaften aus der Unterschiedlichkeit ihrer Gegenstände. Eine solche Bestimmung wird immer dann problematisch, wenn keine Einmütigkeit über den betreffenden Gegenstand zu erzielen ist oder mehrere Disziplinen denselben Gegenstand für sich beanspruchen (vgl. ausführlicher Traxel 1968). 2. Analog zur Auffassung der exakten Naturwissenschaften kann eine Wissenschaft auch von ihren *Methoden* her bestimmt werden (*method-* oder *technique-centered* im Sinne Maslows). Hier gilt der Satz: „Eine wissenschaftliche Erkenntnis ist soviel wert wie die Methode, durch die sie gewonnen wurde."
Die in diesem Zusammenhang oft gestellte Frage, ob die Psychologie eine Natur- oder eine Geisteswissenschaft sei, kann weder von der gegenstandsbestimmten noch von der methodenorientierten Definition her eindeutig und zufriedenstellend entschieden werden. Methodisch gesehen, ist die Psychologie eine empirische oder Erfahrungswissenschaft, insofern hier alle angewandten Methoden empirischen Kontrollen unterliegen, d.h. auf ein *empirisches* Objekt ausgerichtet sind. Gegenständlich betrachtet, stellt die Psychologie eine Realwissenschaft dar (Wahrnehmung, Lernen, Wollen, Denken usw. sind hier als *Realgegenstände* angesprochen), wenngleich uns im subjektiven Erleben (z.B. Werterleben oder Urteil) auch *Idealgegenstände* (Ästhetik, Logik, Mathematik) zugänglich sind; siehe ausführlicher Kap. 1.3 unten. Methodisch wie gegenständlich ist und bleibt die Psychologie somit „ein Bürger zweier Welten", was in besonderer Weise für die Pädagogik bzw. Erziehungswissenschaft(en) gelten dürfte. „Gerade in der Mannigfaltigkeit der in ihr repräsentierten Bemühungen und der Vielschichtigkeit der in ihr berücksichtigten Aspekte ist die Psychologie allein in der Lage, der Differenziertheit und Komplexität menschlichen Verhaltens und Erlebens gerecht zu werden" (Thomae & Feger 1969, S. 3). Daß diese Aufgabe nur unter Wahrung der eingangs zitierten – generellen – Wissenschaftskriterien geleistet werden kann, braucht hier kaum erwähnt zu werden. Bevor wir diese weiter ausführen, seien einige Reflexionen über den Verhaltensbegriff, der ja zentrale Bedeutung im Kontext empirischer Sozialwissenschaften erlangt, angestellt.

1.3. Exkurs über den Verhaltensbegriff

Man sollte annehmen, daß ein so häufig verwendeter Begriff wie der des Verhaltens – zumindest in der wissenschaftlichen Literatur – eindeutige Festlegungen erfahren hat. Dem ist leider nicht so. Entweder wird der Begriff „Verhalten" überhaupt nicht definiert und somit mehr oder weniger unkritisch gebraucht, oder es werden vordergründige bzw. einseitige Deutungen (z.B. in behavioristischer Reduktion) angeboten. Im Bemühen, die Struktur des Verhaltens vom phänomenologischen Ansatz her aufzuhellen und auf diese Weise zur Klärung des Begriffs „Verhalten" beizutragen, verdient besonders ein von Graumann unternommener Versuch unsere Beachtung.

Nach Graumann (1960) ist Verhalten, wenigstens soweit damit beobachtbares Verhalten angesprochen ist, immer irgendwie auf den Menschen bezogen. Erst unter dieser Perspektive eignet dem Begriff „Verhalten" Sinndimension, d.h. psychologische bzw. sozialwissenschaftliche Bedeutung. Verhalten meint demnach „Sich-in-ein-bestimmtes-Verhältnis-Setzen" oder „Sich-in-einem-bestimmten-Verhältnis-Finden". Diese Reflexivität des Verhaltens kann das Verhältnis (einer Person) zu *Personen* – anderen Personen oder der eigenen Person – versus zu *Gegenständen* ausdrücken, wobei hier „Verhältnis" immer ein bestimmtes Verhältnis meint. Strenggenommen kann sich danach nur ein personales Wesen – der Mensch – zu etwas verhalten, im weiteren Sinne und bedingt jedes Lebewesen. Dagegen stehen Objekte ausschließlich in *Relation* zueinander, sie können kein „Verhältnis" eingehen. Zwar impliziert jedes Verhalten auch eine Relation, doch ist der Verhaltensbegriff umfänglicher und beinhaltet stets eine „personale Wendung", wie Graumann im Anschluß an Löwith (1962) formuliert. Zum Verhältnis i.e.S. gehört die Begegnung und somit ein soziales Moment. Hierin unterscheiden sich die Begriffe „Verhältnis" und „Relation". Die *Implikation des Personalen* (Humanen und Sozialen) im Verhaltensbegriff kommt bereits in zahlreichen Redewendungen der Alltagssprache zum Ausdruck, z.B.: „Er hat ein Verhältnis" – „Sie kommt aus guten Verhältnissen" – „Der Junge hat ein gutes (vs. schlechtes) Verhältnis zu den Eltern (Lehrern, Klassenkameraden, zur Schule, zur Arbeit usw.)".

Neben dem personalen Bezug wäre als zweites Charakteristikum die *Intentionalität* des Verhaltens zu nennen, d.h. die *Bezogenheit auf* andere oder sich selbst versus auf etwas. Wenn wir uns (in irgendeiner Weise) verhalten, sind wir immer auf jemanden/etwas gerichtet, den/das wir als unabhängig (von uns selber) existierend ansetzen. Diese Voraussetzung ermöglicht überhaupt erst „Verhalten" im bisher explizierten Strukturverständnis. Die Relate unseres intentionalen Bezogenseins können personaler, idealer oder materialer Art sein. Einige Beispiele mögen das Gesagte konkretisieren.

1. Beispiele mit *personalem* Relat: „Ein Lehrer ärgert sich über seine *Schüler*" – „Die Schüler ärgern sich über ihren *Lehrer*" – „Sie ärgerte sich über ihren *Freund*" – „Er beschwerte sich über seinen *Vorgesetzten*". – Die Präposition „über" deutet an, daß die Schüler, der Lehrer, der Freund oder der Vorgesetzte als ein vom momentanen, z.B. ärgerlichen, Verhalten (= Intentionalität) unabhängiges Seiendes begriffen werden. Nicht unabhängig ist dagegen das jeweilige Verhalten. Als Sonderfall eines personalen Relates wäre das Beispiel „Ich ärgere mich über mich selbst" anzusprechen.

2. Beispiele mit *idealem* Relat: „Ein Wissenschaftler ist darauf aus, eine bestimmte *Theorie* zu widerlegen" – „Ein Politiker fühlt sich an die Partei*doktrin* gebunden" – „Ein Pädagoge möchte ein bestimmtes Erziehungs*ziel (Ideal)* verwirklichen" – „Er hatte ein schlechtes Verhältnis zur *Wahrheit*" – „Ich erinnere mich an mein *Versprechen*". – Unabhängig von meinen Bemühungen, Erinnerungen, Erfolgen

oder Mißerfolgen existieren die Theorie, das Versprechen oder bestimmte Ideale, denen gegenüber ich mich so und so verhalte.
3. Beispiele mit *materialem* Relat: „Der Junge sparte für ein neues *Fahrrad*" – „Das Mädchen freute sich über das *Schmuckstück*" – „Sein größtes Hobby war ein *Segelboot*" – „Er zog die Kapuze über, als der *Schneesturm* (*Regen* usw.) einsetzte". – Auch die materialen Relate, auf die bezogen unser Verhalten so oder so aussieht, bestehen unabhängig von unserer Erfahrung, die hinwiederum in je spezifischer Weise durch unser Verhalten zu eben den genannten Relaten geprägt wird.
Die aufgewiesenen Merkmale (menschlichen) Verhaltens erlauben freilich noch keine vollständige Analyse, hierzu bedarf es eines weiteren – dritten – Kriteriums: der *Aspekthaftigkeit* oder *Hinsichtlichkeit*. Wir verhalten uns zu den materiellen, ideellen und personellen Relaten immer nur in einer bestimmten Hinsicht, diese interessieren jeweils nur aspekthaft. Nicht das Segelboot als solches interessiert, sondern sein Freizeitwert, vielleicht auch sein Prestigewert usw. Am Fahrrad interessiert beispielsweise die Möglichkeit, den Schulweg zu verkürzen, damit Sport zu treiben u. ä. Der Ärger über mich selbst bezieht sich wohl kaum auf das gesamte Ich, sondern auf meine Ungeschicklichkeit in einer konkreten Situation, meine Gedächtnishemmung bei der Prüfung oder eine verpaßte Gelegenheit. Ausdrücke wie „Ich könnte mich selber ohrfeigen!" oder „Ich Dummkopf!" lassen erkennen, daß damit nicht die gesamte Existenz, sondern die (eigene) Person in einer bestimmten Rolle, Funktion usw. gemeint ist. Auch der Ärger über den Schüler vs. Lehrer bezieht sich nicht auf die Gesamtpersönlichkeit, sondern nur auf ein bestimmtes Verhalten (Faulheit, Desinteresse, ungerechte Behandlung u.ä.).
Der Begriff des Verhaltens ist ein grundlegender Begriff der Psychologie bzw. der Sozialwissenschaften überhaupt. *Verhalten* meint immer Ein-sich-zu-jemandem-Verhalten oder Ein-sich-zu-etwas-Verhalten, und zwar stets ein Verhalten nur in einer bestimmten Hinsicht, die nach Graumann *Perspektivität* des Verhaltens meint. Alles, was wir erfahren und wozu wir uns verhalten, steht in einem bestimmten Horizont. Dem Begriff des Verhaltens eignet somit *Horizontstruktur*. Perspektive und Horizont als Korrelate des Verhaltens eröffnen erst seine volle Bedeutung und zugleich die Möglichkeit einer zureichenden Begriffsbestimmung dessen, was wir „Verhalten" nennen.

1.4. Kriterien einer Wissenschaftssprache

Die Graumannsche Wissenschaftsdefinition betont die Interdependenz von Begriffen und Methoden. Methoden sollen Begriffe klären, Begriffe sollen Erkenntnisse formulierbar und somit kommunizierbar machen. Für den Vollzug der Wissenschaft ist demnach die *Sprache* als Symbolsystem unentbehrlich, ja sie wird zum *wissenschaftlichen Grundwerkzeug* vor allen (anderen) Methoden wie Test, Experiment usw. Freilich: An eine solche (Wissenschafts-)Sprache sind grundsätzlich die gleichen Anforderungen zu richten, wie sie an jede Methode der Wissenschaft gestellt werden müssen. Es werden Objektivität (Intersubjektivität), Reliabilität (Verläßlichkeit) und Validität (Tauglichkeit oder Gültigkeit) gefordert. Objektivität und Reliabilität werden hier zum Hauptproblem. Die in einer Wissen-

schaftssprache verwendeten Begriffe oder Symbole müssen *eindeutig* definiert und *invariant* gebraucht werden.
Die Eindeutigkeit betrifft zuallererst ein Definitionsproblem. Am wichtigsten sind folgende *Definitionsarten:* die Realdefinition (Sacherklärung), die Nominaldefinition (Worterklärung) und die operationale Definition (Erklärung durch Angabe der Operation, d. h. des methodischen Zugangs zum Definitionsgegenstand).
Zur *Real*definition, der klassischen Form der Begriffsfestlegung, gehören *genus proximum* (Angabe der nächsthöheren Gattung = Gattungs- oder Oberbegriff) und *differentia specifica* (Angabe des artspezifischen Unterschiedes). Beispiel: „*Homo est animal rationale*" (Der Mensch ist ein vernunftbegabtes Lebewesen). Hierbei ist *animal* (Lebewesen) das *genus proximum*, d.h., der Mensch gehört zur Gattung der Lebewesen; *rationale* ist die *differentia specifica*, d.h., der Mensch unterscheidet sich von anderen Lebewesen (Tieren und Pflanzen) durch seine Vernunftausstattung. – Weitere Beispiele: „Leistungsmotivation ist das Bedürfnis nach Leistung" – „Intelligenz ist die Fähigkeit zum Denken" – „Ein Dreieck ist eine durch drei Geraden begrenzte, ebene Fläche". – Mit Hilfe von Realdefinitionen wird versucht, eine Wesensbestimmung zu geben, wobei alle notwendigen oder „wesentlichen" Merkmale genannt und alle überflüssigen oder „unwesentlichen" Bestimmungen vermieden werden sollen.
Die Aufgabe der *Nominal*definition liegt in der Beschreibung bzw. Umschreibung eines Sprachsymbols oder dem Angebot entsprechender Illustrationsbeispiele: „Intelligenz ist Begabung" – „Intelligenz meint die Fähigkeiten zum Denken, zum Erlernen von Sprache(n), zum Operieren mit Zahlen usw." – „Zum Bewußtsein gehören Wahrnehmen, Vorstellen, Gedächtnis usw." – „Motivieren ist: ‚Das habt ihr aber fein gemacht. Nun habe ich eine neue Aufgabe. Da bin ich gespannt, ob ihr die ebenso gut lösen könnt!'" – Bei der Nominaldefinition werden also (unbekannte) Begriffe durch andere, mehr oder weniger bekannte Begriffe oder auch Beispielbeschreibungen ersetzt. Sofern man keine Realdefinition aufstellen kann und eine operationale Definition (siehe unten) vermeiden möchte, ist man praktisch auf die Nominaldefinition bei der Begriffsbestimmung angewiesen. Für den Wissenschaftler bedeutet die Nominaldefinition freilich kaum mehr als eine – vorläufige – Plattform zur Verständigung mit anderen Kollegen versus Laien.
Die *operationale* Definition stellt einen radikalen Ansatz dar. Hier stehen die Verfahrensweisen im Mittelpunkt des Interesses, d.h., es wird der Versuch unternommen, Begriffe durch Operationen zu klären. Solche operationalen Begriffsbestimmungen sind uns bereits aus der Kinderpsychologie bekannt, sie sind aber auch für den Erfahrungswissenschaftler von höchstem Interesse. Häufig ist die Forderung nach semantischer Eindeutigkeit der Sprache qua Voraussetzungsbedingung der Kontrollierbarkeit von Forschungsergebnissen resp. wissenschaftlicher Aussagegültigkeit nur auf dem Wege operationaler Definitionen zu erfüllen. Im folgenden wiederum einige Beispiele operationaler Begriffsbestimmung: „Ein Auto ist, womit man fahren kann, womit man spielen kann, worin man sitzen kann" – „Ein Tisch ist, woran man essen kann, woran man Hausaufgaben machen kann, woran man Briefe schreiben kann" – „Eine Insel – da muß man mit dem Boot hinfahren" – „Fleiß ist, wenn man seine Hausaufgaben immer ordentlich macht, wenn man der Mutter beim Abtrocknen hilft, wenn man regelmäßig Klavier übt usw.". – Oder: „Hunger ist der Zustand nach 6stündigem Nahrungsentzug" – „Intelligenz ist das, was der Intelligenztest mißt" – „Pädagogik ist das,

was die Pädagogen tun". – Stärken und Schwächen des operationalen Ansatzes werden an diesen Beispielen deutlich. Sein unbestreitbarer Vorzug liegt in der Möglichkeit präziser Formulierungen konkreter Sachverhalte, wobei man oft mehrere Funktionseigentümlichkeiten zugleich benennt (s. Beisp. „Auto", „Tisch", „Fleiß"). Die Nachteile operationaler Definitionsversuche erweisen sich bei der Formulierung allgemeiner Begriffe; hierin ist die klassische Definition mit *genus proximum* und *differentia specifica* als ihren Konstituenten der operationalen Definition überlegen und durch nichts zu ersetzen. Andererseits: Wer ein Auto oder einen Tisch noch nicht gesehen hat, wird durch die klassische Definition des Autos oder des Tisches kaum gescheiter. Der Anschauung bzw. Vorstellung dienlicher ist die operationale Definition, diese Behauptung kann auch im Hinblick auf die Bestimmung abstrakter Begriffsinhalte wie „Fleiß", „Intelligenz", „Gerechtigkeit" u.ä. aufrechterhalten werden. Bei der operationalen Definition werden nicht so sehr Aussehen, Merkmale oder „Wesens"-Eigenschaften wichtig, sondern die Operation bzw. Funktion(en). Im Hinblick auf das Semantikproblem, das sich jeder Wissenschaft stellt, bleibt allerdings hier festzuhalten, daß in jedem Falle die generalisierende Abstraktion im Wissenschaftsvollzug irgendwo geleistet werden muß. Und dies kann nicht streng operational geschehen, sondern nur durch eine vorausgehende gründliche Phänomenanalyse und Begriffsbestimmung. Erst danach erfolgt – sinnvollerweise – die Übersetzung ins Operationale. „Nur auf diesem Wege ist dann eine klare Bestimmung dessen möglich, wovon ich vorher schon einen Begriff hatte. Haben wir aber vorher nur einen Zipfel des Begriffes, dann kann man zwar eine Hypothese verifizieren, jedoch kann man damit nicht viel anfangen, weil man nicht verallgemeinern kann, d.h. die Ergebnisse nicht einordnen kann in ein System, in einen wissenschaftlichen Kontext. Wissenschaft aber hat Systemcharakter, was beim rein operationalen Vorgehen leider oft ignoriert wird" (nach Graumann, a.a.O.).
Die für eine Wissenschaftssprache geforderte *Invarianz* bezieht sich sowohl auf die *inter*individuelle als auch auf die *intra*individuelle Konsistenz des Sprachgebrauchs. Verläßlich und tauglich ist eine Fach- oder Wissenschaftssprache ja nur, wenn ihre semantische Eindeutigkeit gesichert und ihre intersubjektive Verbindlichkeit (im Konsens der Experten) sowie ihre intrasubjektive Konsistenz (für den einzelnen Wissenschaftler muß ein und derselbe Begriff stets dieselbe Bedeutung involvieren) gegeben sind. Mit anderen Worten: Wer die Sprache als wissenschaftliches Werkzeug einsetzen will, muß sich den Regeln der Semiotik unterwerfen.
Die *Semiotik* (allgemeine Sprachtheorie oder Sprachanalyse) betrachtet die Sprache, d.h. ihre Zeichen oder Symbole (Wörter), in dreierlei Hinsicht: erstens unter *syntaktischem* Aspekt (die Regeln der Syntax sind für die Relationen der Zeichen untereinander verantwortlich), zweitens unter *semantischem* Aspekt (die Semantik regelt die Beziehung zwischen den Zeichen und dem jeweils Bezeichneten, also zwischen Symbol und Sache), drittens unter *pragmatischem* Aspekt (hiermit ist die Beziehung zwischen den Symbolen und denjenigen, die sich dieser Symbole bedienen, also die Anwendung der Sprache betroffen). Ein mit dem Anspruch der Brauchbarkeit auftretendes Symbolsystem qua Fachsprache muß allen drei Kriterien genügen. Es muß *semantische Eindeutigkeit* mit der Verbindlichkeit der Regeln der *Syntax* bzw. *Logik* (Stringenz) und den Anforderungskriterien sub specie *Pragmatik* (inter- und intraindividuelle Konsistenz) verknüpfen.

Da die Alltagssprache den diskutierten Anforderungen kaum jemals genügen kann, ist auch in den meisten Fällen eine schlichte Übernahme der Alltagssprache in die Wissenschaft(en) unmöglich. Hieraus sich ergebende Konsequenz ist der Rückgriff auf eine fachspezifische Symbolik oder *Fachsprache*, die den Kriterien der Eindeutigkeit (Präzisheit) und der Invarianz besser zu entsprechen vermag. Ihr obliegt als Aufgabe die „Symbolisierung der Erfahrung" (Graumann), d.h. die Versprachlichung (Symbolisierung) beobachtbaren Verhaltens und Erlebens. Bevor wir den Anforderungskanon einer Wissenschaftssprache weiter vervollständigen, sei noch kurz auf den Begriff der Erfahrung eingegangen.

Gadamer (1960) stellt hierzu fest, daß der Begriff der *Erfahrung* paradoxerweise zu den unaufgeklärtesten Begriffen gehört. In seiner Analyse arbeitet er folgende Wesenszüge heraus. „Daß Erfahrung gültig ist, solange sie nicht durch neue Erfahrung widerlegt wird *(ubi non reperitur instantia contradictoria)*, charakterisiert offenbar das allgemeine Wesen von Erfahrung, ganz gleich, ob es sich um ihre wissenschaftliche Veranstaltung im modernen Sinne handelt oder um die Erfahrung des täglichen Lebens, wie sie von jeher gemacht wurde" (S. 329). Dabei handelt es sich zunächst immer um eine „Erfahrung der Nichtigkeit". Denn Erfahrungen-Machen bedeutet, daß wir einen Sachverhalt seither nicht wirklichkeitsangemessen erfaßt haben; *nach* der Erfahrung wissen wir besser, wie es um die Sache steht. „Angesichts der Erfahrung, die man an einem anderen Gegenstande macht, ändert sich beides, unser Wissen und sein Gegenstand. Man weiß es nun anders und besser, und das heißt: Der Gegenstand selbst ‚hält nicht aus'. Der neue Gegenstand enthält die Wahrheit über den alten" (S. 337f.). Hierin erblickt Gadamer den tiefsten Bezug von Erfahrung und Einsicht. Beide müssen von uns erworben werden. „Einsicht ist mehr als die Erkenntnis dieser oder jener Sachlage. Sie enthält stets ein Zurückkommen von etwas, worin man verblendeterweise befangen war. Insofern enthält Einsicht immer ein Moment Selbsterkenntnis und stellt eine notwendige Seite dessen dar, was wir Erfahrung im eigentlichen Sinne nannten. Auch Einsicht ist etwas, wozu man kommt. Auch das ist am Ende eine Bestimmung des menschlichen Seins selbst, einsichtig und einsichtsvoll zu sein" (a.a.O.).

Im Anschluß an Buytendijk, der auf die Etymologie des Wortes „Erfahrung" hinweist (*Erfahrung* → erfahren = einen Weg immer wieder intensiv durchfahren), unterscheidet Graumann (1965) drei *Intensitätsstufen* menschlicher Erfahrung: 1. Bloßes Zur-Kenntnis-Nehmen oder Registrieren, ohne daß eine Verankerung im Kontext stattfindet. Rasches Vergessen läßt hier kaum eine Systematisierung gemachter Erfahrungen zu. 2. Auf der zweiten Stufe kommt es dann zum Wissen mit mehr oder minder ausgeprägtem Gewißheitsgrad. 3. Das Wissen um etwas geht schließlich über in die Stufe der eigentlichen Erkenntnis. Erfahrung in der Form von Gesehen-Haben (Wahrgenommen-Haben) fundiert alles Wissen und bildet die Grundlage für das Verständnis (= Einsicht).

Eine Art der Erfahrung stellt die bloße Kenntnisnahme dar, der aber ein sehr flüchtiger Charakter eignet. Erfahrungen können wir jedoch festhalten, indem wir sie symbolisieren. Dieser Tatbestand tritt am augenfälligsten dort in Erscheinung, wo dem Individuum nicht von vornherein ein solches Symbolsystem (Sprache) zur Verfügung steht. Die (künstliche) Sprachanbildung bei Taubstummen ist deshalb – jenseits sprachlicher Kommunikationsfunktion – gerade für die Erfahrungsintensivierung und damit die gesamte seelisch-geistige Entwicklung des gehörlosen Menschen so überaus wichtig. Die Korrespondenz zwischen Symbol und Erfahrung erweist sich freilich oft – nicht nur bei der Sprachgenese hörgeschädigter Kinder – als weit schwieriger, als man zunächst anzunehmen geneigt ist. Im Hinblick auf die Konstituierung einer Wissenschaftssprache anzustrebendes Ideal wäre hierbei, für jede Erfahrung ein eigenes Symbol (Schriftzeichen, mathematisches Symbol u.ä.) bereitzustellen. Praktisch würde dies eine Inventarisierung aller möglichen Erfahrungen (bzw. Erkenntnisse) voraussetzen und ein kaum realisierbares Unterfangen darstellen. Gelegentliche Versuche, diesen Schwierigkeiten durch Bildung von Neologismen zu begegnen, erscheinen nun nicht mehr ganz so „unsinnig", wie häufig von Laien angenommen wird. Freilich muß damit ein erheblicher Nachteil in Kauf genommen werden, insofern die Verständigung – die Mitteilbarkeit der Forschungsergebnisse ist ja unabdingbares Wissenschaftskriterium (s. S. 15 oben) – hierunter leidet. Sehr viel ökonomischer ist dann die Bildung von Symbol*klassen*. An ein solches Symbolsystem qua Wissenschaftssprache müssen nach Graumann (a.a.O.) vier Anforderungen gestellt werden:

1. Ein brauchbares Symbolsystem muß *unbegrenzte Möglichkeiten* (in semantischer, aber auch in syntaktischer Hinsicht) enthalten, denn der Entdeckung von Phänomenen (möglichen Erfahrungen) sind praktisch keine Grenzen gesetzt. Weiterhin sind zwischen einzelnen Phänomenen beliebig viele Verbindungsmöglichkeiten denkbar, so daß ein geeignetes Symbolsystem dieser Progression von Phänomenen (Erfahrungen) gewachsen sein muß. Hier ist die Alltagssprache überfordert, schon deshalb, weil der Wissenschaftler durch systematisches Suchen viel mehr Phänomene entdecken und Erfahrungen machen kann als der Alltagsmensch bzw. Laie in gleicher Lage.

2. Die zweite Forderung bezieht sich auf die *Flexibilität* des Symbolsystems, das nicht nur den unbegrenzten Möglichkeiten, sondern auch dem Wandel der Erfahrung Rechnung tragen muß. Aktuelle Erfahrungen bleiben ja nicht das, was sie im Augenblick früherer Erfahrung einmal waren, worauf besonders Gadamer hingewiesen hat. Auch unter diesem Aspekt ist die Alltagssprache in ihrer Leistungsfähigkeit begrenzt; die Superlativformen der Umgangssprache kennzeichnen nur grob Zu- und Abnahme, andere Veränderungsmöglichkeiten bleiben ihr weitgehend versagt. In der Wissenschaft benötigen wir hierzu sogenannte *dimensionierte* Kontinua. Der Begriff der Dimensionierung wird später noch zu klären sein (vgl. Kap. 2.2.6.3).

3. Schließlich muß ein brauchbares Symbolsystem für die Wissenschaft Art und Ausmaß der Beziehungen zwischen den Phänomenen und Erfahrungen sichtbar machen können. Der Forderung nach *Erfassung der Relationen* zwischen den Phänomenen kann die Alltagssprache infolge ihrer geringeren Kapazität ebenfalls nur unzureichend genügen; hierzu bedarf es oft einer Symbolisierung in Form von graphischen Darstellungen oder der Anwendung mathematischer Symbole (vgl. die nachfolgenden Ausführungen zur *deskriptiven* und *analytischen* Statistik bzw. *Korrelationsrechnung* in diesem Buch). In diese Formen der Darstellung auszuweichen bedeutet demnach keine Spielerei, sondern höchst zweckhafte Symbolisierung phänomenaler Bezüge und Erfahrungskomplexe in den empirischen oder Erfahrungswissenschaften.

4. Die vierte Anforderung betrifft die *Stufenfolge der Symbolisierung*, also den Prozeß im Wissenschaftsvollzug:

a) Die erste Stufe der Erfahrung ist die *Beobachtung*, die *beschrieben* (deskribiert) werden muß. Die *Beschreibung* oder *Deskription* stellt also die Grundstufe der Symbolisierung dar.
b) Der Deskription oder Beschreibung, manchmal auch Phänographie genannt, folgt sodann die *vereinfachende Erklärung* als zweite Stufe der Wissenschaft.
c) Schließlich folgt die *Theorienbildung* als dritte und letzte Stufe der Symbolisierung (vgl. Kap. 2.2.6.5). Auf allen drei Stufen kommen wir mit der Alltagssprache nicht oder nur unzulänglich aus.

Ein wichtiges Postulat für jeden wissenschaftlich Tätigen ist das *Voraussetzungsbewußtsein*. Früher sprach man in diesem Zusammenhang gern von „Voraussetzungslosigkeit"; heute wissen wir, daß es diese nirgends in der Wissenschaft (und erst recht nicht außerhalb) gibt. Sowohl die sprachlichen als auch die mathematischen Symbole bzw. Symbolsysteme (Fachsprachen) haben ihre Eigengesetzlichkeit. Diese wird vom Wissenschaftler, der diese Systeme benutzt, mit übernommen. Wo immer wir wissenschaftlich tätig werden, arbeiten wir mit Symbolen dieser oder jener Art und ihren Voraussetzungen. Wissenschaftlichkeit kann demnach nicht Voraussetzungslosigkeit bedeuten. Vielmehr wird der Wissenschaftler stets bemüht sein, sich dieser Voraussetzungen bewußt zu werden – im Gegensatz zum Laien, dem dieses Bewußtsein weithin fehlt. Wie oben schon aufgezeigt werden konnte, übernehmen wir in bezug auf eine Wissenschaftssprache drei Arten von Regeln als Voraussetzungsbedingungen jeglicher Symbolik: die Regeln der *Syntax*, der *Semantik* und der *Pragmatik*.
Psychologisch besonders wichtig ist in diesem Zusammenhang der dritte Aspekt, nämlich die Frage, wie sich ein Mensch zu den einzelnen semantischen Zeichen verhält. Glaube an die Statistik versus Horror vor mathematischen Symbolen und Formeln, Ressentiment gegen Zahlen überhaupt, Faszination durch bestimmte Begriffe wie „Frustration", „Gestalt", „Regelkreis", „Test" u. ä. oder auch politische Schlagwörter und Slogans bergen entsprechende Verhaltensmöglichkeiten genug. So können die Begriffe „Gesamtschule", „linksintellektuell", „Kapitalismus" u. dgl. m. bei manchen den Hautwiderstand meßbar erhöhen. Hier ist für jeden Wissenschaftler kritische Selbstprüfung resp. Kontrolle vonnöten.

Am Schluß unserer ziemlich umfangreich geratenen Ausführungen zu den Kriterien einer Wissenschaftssprache sei noch kurz auf das Verhältnis zwischen Syntaktik (Logik bzw. Mathematik) und Semantik versus das Verhältnis der Syntaktik/Semantik zur Empirie eingegangen. Dazu wäre folgendes festzustellen.
Syntaktisch gesehen, kann ein Satz wahr oder falsch sein, d. h., dem *deduktiven* Vorgehen, wie es z. B. in den axiomatischen Systemen der Logik oder Mathematik seinen Platz findet, eignet absolute Stringenz. Folgende Beispiele mögen das schlußfolgernde Denken oder logische Schließen via *Deduktion* illustrieren: „Alle Kinder mit einem IQ unter 85 sind lernbehindert. Otto und Roswitha haben einen IQ unter 85. Also sind Otto und Roswitha lernbehindert." – „Alle Menschen sind sterblich. Tarzan ist ein Mensch. Also ist Tarzan sterblich." – „Akademikerkinder sind grundsätzlich intelligent. Franz ist der Sohn eines Akademikers. Demnach ist Franz ein intelligenter Junge." – Bei der Deduktion erfolgt der Schluß vom Allgemeinen auf das Besondere, aus Gesetzmäßigkeiten werden Folgerungen für den Einzelfall abgeleitet, aus Theorien werden auf deduktivem Wege Hypothesen gewonnen (= *deduktiv-nomologische* Erklärung). Der Nachteil deduktiver Vor-

gehensweisen liegt im mangelnden Empiriebezug. Syntaktisch oder logisch sind Aussagen schon dann *wahr*, wenn sie (nur) die Regeln der Syntax, der Logik, der Mathematik einhalten, und immer dann *falsch*, wenn sie gegen diese Regeln verstoßen bzw. auf falschen Prämissen (Voraussetzungen) aufbauen – siehe Beispiel „Akademikerkinder".

Semantisch betrachtet, kann ein empirisch sinnvoller, etwa durch Messung oder eine andere empirische Operation gewonnener Satz niemals wahr vs. falsch, sondern immer nur *wahrscheinlich* sein. Einschlägig ist hier die Induktion, der Schluß vom Besonderen auf das Allgemeine, die Generalisierung (vgl. auch Kap. 2.2.6.4). Hierzu wiederum einige Beispiele: „Bärbel ist kleiner als der (gleich alte) Klassenkamerad Otto. Ute ist kleiner als der Klassenkamerad Bernd. Maria ist kleiner als der Klassenkamerad Ulf. Also sind Mädchen (einer bestimmten Alters- bzw. Klassenstufe) kleiner als Jungen." – „Der Mann A ist gestorben. Der Mann B ist gestorben. Der Mann C ist gestorben. Also sind alle Männer sterblich." – „Die Studierenden an den wissenschaftlichen Hochschulen der BRD sind zu 70% männlich (und zu 30% weiblich). Das Gymnasium besuchen in der BRD 60% männliche (und 40% weibliche) Schüler, die Realschule in der BRD 55% Knaben (und 45% Mädchen). Demnach sind in der BRD die Bildungschancen der Mädchen schlechter als die der Jungen." – Es leuchtet ohne weiteres ein, daß via Induktion gewonnene Aussagen dieser oder ähnlicher Art niemals letzte Sicherheit beanspruchen können. Beim induktiven Vorgehen wird stets von einigen Fällen, d.h. einer mehr oder weniger großen Zahl von Einzelfällen, auf die Gesamtheit aller Fälle geschlossen (= *induktiv-statistische* Erklärung). Das Risiko dieses Vorgehens hängt von verschiedenen Faktoren ab, wie die späteren Ausführungen zur Stichprobenbildung bzw. Stichprobenstatistik (Inferenzstatistik) noch zeigen werden. Der Vorteil induktiver Schlüsse liegt vor allem in der empirischen Fundierung, also der Möglichkeit, (deduktiv abgeleitete) Hypothesen empirisch zu überprüfen, d.h. zu veri- oder falsifizieren. Die Verallgemeinerung der so gewonnenen Ergebnisse erlaubt die Formulierung bestimmter Gesetzlichkeiten, die sich schließlich zu Theorien zusammenfügen, aus denen gegebenenfalls erneut Hypothesen abgeleitet werden. Induktion und Deduktion repräsentieren somit komplementäre (Denk-)Methoden und bilden die Grundlage für das *hypothetisch-deduktive Verfahren* in der empirischen Forschung, also auch in den modernen Sozialwissenschaften. Synonyme Begriffe hierfür sind die Termini „hypothetico-deduktive" oder „mathematico-deduktive" Methode.

Was endlich das Verhältnis von Syntaktik und Semantik zur Empirie betrifft, so gilt: Weder Syntax (Logik, Mathematik) noch Semantik können darüber entscheiden, was *empirisch* „richtig" oder „falsch" ist, sondern die Erfahrung selber, die syntaktisch (logisch usw.) und semantisch gefaßt ist. Freilich, vor jeder empirischen Untersuchung müssen Regeln (der Semantik, der Logik usw.) beachtet werden, die jedoch noch keine empirisch sinnvolle Beziehung garantieren. Diese Feststellung wird besonders dann zum Problem, wenn mathematische Modelle zur Anwendung kommen sollen. Diese sind zweifellos sehr wichtig und auch für die Sozialwissenschaften unentbehrlich, doch bedarf es stets kritischer Prüfung, ob eine sinnvolle, d.h. *isomorphe Relation* zwischen Modell oder Theorie auf der einen Seite und Empirie oder erfahrbarer Wirklichkeit auf der anderen Seite gegeben ist.

1.5. Die Bedeutung des Operationismus für die empirische Forschung

Beobachtung, Experiment, Interview u.ä. sind operationale Methoden, die in der Psychologie bzw. den Sozialwissenschaften ausgedehnte Anwendung finden (vgl. König 1956, 1962, Anger 1969, Bredenkamp 1969, Cranach & Frenz 1969, Landsheere 1969, Süllwold 1969 u.a.). Sie werden weiter unten noch eingehend erörtert. Operationalen Ansätzen kommt in der empirischen Forschung entscheidende Bedeutung zu. Zum wissenschaftstheoretischen Verständnis des Operationalismus oder Operationismus seien zunächst die von Stevens (1939) aufgestellten Grundsätze – wiederum in Anlehnung an Graumann (1965) – referiert:

1. „Wissenschaft ist eine Reihe von empirischen Sätzen, die die *Zustimmung* von anderen Gliedern der Gesellschaft findet." Der Konsens der Fachleute verhilft also einer Aussage dazu, in ein System (= Wissenschaft) aufgenommen zu werden. Eine solche Demokratisierung ist jedoch nicht ohne Problematik, die in folgenden zwei Fragen gipfelt:
a) Welche Personen bilden den Konsens bzw. sind dazu berechtigt?
b) Wie groß muß der Konsens sein?

2. „Zur Wissenschaft zugelassen sind nur Aussagen, die *öffentlich* und *wiederholbar* sind." Wiederholbarkeit setzt Öffentlichkeit voraus. Introspektion und subjektives (Evidenz-)Erleben sind danach keine Garanten der Wissenschaftlichkeit. Trotzdem können Aussagen, die auf innerer Erfahrung beruhen, ausnahmsweise durch andere bestätigt oder widerlegt werden, nämlich dann, wenn privat Erlebtes und Erfahrenes Öffentlichkeitscharakter erhält, d.h., wenn die genauen Bedingungen mitgeteilt werden, unter denen die betr. Erlebnisse und Erfahrungen gemacht wurden. Praktisch stehen dieser Forderung freilich erhebliche Schwierigkeiten entgegen (s. Kap. 2.1.3.3).

3. Sowohl für die Fremd- wie auch für die Selbstbeobachtung fordert Stevens eine strenge Trennung zwischen Beobachter und Beobachtetem. Subjekt- und Objektpol sind also strikt auseinanderzuhalten. Das Problem der Veränderung des Beobachteten durch den Beobachter wird vom Operationalisten zumeist übersehen.

4. Jeder Versuchsleiter kann prinzipiell zur Versuchsperson für einen anderen Versuchsleiter werden. Bei diesem Beobachtungsregreß muß es jedoch irgendwo einen unabhängigen Beobachter geben, wenn die relative Unabhängigkeit zwischen Beobachter und Beobachtetem nicht verlorengehen soll.

5. „Begriffe bedeuten nur dann etwas, wenn es *konkrete Kriterien* für ihre Anwendbarkeit gibt; eine These oder Aussage hat nur dann empirische Bedeutung, wenn die Kriterien ihrer Richtigkeit aus konkreten Operationen bestehen, die auf Verlangen ausgeführt werden können." Dieser Satz ist wohl der entscheidendste von allen im Hinblick auf die praktische empirische Arbeit (s.a. S. 20f.).

6. „Wenn komplexe Operationen auf einfachere zurückgeführt werden, dann zeigt sich als *die* fundamentale Operation schlechthin die *Unterscheidung (discrimination)*. Alle noch so raffinierten Verfahrensweisen gehen letztlich zurück auf die schlichte Unterscheidung: ‚etwas ist' oder ‚etwas ist nicht'. Nur wer Unterscheidungen – die intersubjektiv gültig sein müssen – treffen kann, ist Wissenschaftler. Die Abhebung muß eindeutig und vermittelbar sein." Hiermit ist die Übereinstimmung (Objektivität) mehrerer Beobachter gefordert, was schon relativ früh als Problem erkannt wurde. Am bekanntesten ist das einschlägige Beispiel von der

Entlassung des Assistenten an der Sternwarte von Greenwich, dessen (Zeiger-)Ablesungen von denen seines Chefs (Professors) abwichen. Dies hat später zur Entdeckung der „persönlichen Gleichung" *(personal equation)* geführt und ist im Prinzip Ausgangspunkt psychologischer Beschäftigung mit den sog. interindividuellen oder intersubjektiven Differenzen (= Gegenstand der differentiellen und diagnostischen Psychologie). Es ist also notwendig, daß die eigenen Diskriminationsleistungen – unter sonst gleichen Bedingungen – mit den Diskriminationsleistungen der anderen übereinstimmen, da sonst alle weiteren Operationen sinnlos werden. Der 6. Satz Stevens' gilt grundsätzlich auch für Kategorisierungen (s. Kap. 2.2.6.3).

7. „Es gibt formale und empirische Sätze." Nach Stevens können nur solche Sätze Gegenstand einer empirischen Wissenschaft sein, bei denen eine semantische Identifikation (Abhebung) von Beobachtetem vorliegt (vgl. S. 24f.: Verhältnis von syntaktischen zu empirischen Aussagen). Freilich: Jede Wissenschaft wird auch syntaktische Aussagen beachten müssen!

Schließlich hebt Stevens den Operation(al)ismus gegenüber dem (schlichten) Positivismus einerseits und dem (orthodoxen) Behaviorismus andererseits ab. Er meint, der Operationismus sei kein „-ismus" im üblichen Sinne, sondern erhebe eine bestimmte Art wissenschaftlichen Vorgehens zum Kriterium der Wissenschaftlichkeit. Danach steht am Anfang jedes wissenschaftlichen Vorgehens eine Unterscheidungsleistung (Diskrimination), also eine Abhebung (Identifikation) der zu untersuchenden Phänomene. Dies kann grundsätzlich via Kategorisierung oder/ und via Dimensionierung geschehen (s. unten S. 84f.).

Während die Phänomenologie den Blick auf die Phänomene richtet, lenkt der Operationismus unsere Aufmerksamkeit unerbittlich auf die Vorgehensweisen, d.h., hier werden alle Aussagen von der Redlichkeit der Operationen abhängig gemacht. Beide Ansätze sind wichtig, wie unsere Ausführungen zu verdeutlichen suchten. Die operationalen Wissenschaftskriterien gelten für alle empirischen Verfahrensweisen (Beobachtung, Experiment, Test, Fragebogen, Interview u.ä.), soweit sie wissenschaftliche Untersuchungsmethoden darstellen. Für ein vertiefendes Studium der angeschnittenen Probleme empfehlen wir – neben den bereits erwähnten Veröffentlichungen – besonders noch die einschlägigen Publikationen von Carnap (1942, 1956, 1969), Carnap & Jeffry (1971), Stegmüller (1957, 1969, 1970), Stegmüller & Carnap (1972), Topitsch (1960), Marx (1963), Traxel (1964), Bochenski (1965), v. Cube (1965), Popper (1971), Selg & Bauer (1971), Westmeyer (1973), wo jeweils weitere Literaturquellen zu finden sind.

2. Grundmethoden der empirischen Wissenschaften

2.1. Beobachtung (Observational Techniques)

2.1.1. Begriff und Kriterien

Die (sozialwissenschaftliche) Methode der Beobachtung hat ihren Vorläufer in der Phänomenologie Husserlscher Prägung. Der Begriff „Phänomenologie" bezeichnet hier nicht die philosophische, sondern die empirische Phänomenologie qua *Methode* der Beschreibung oder Deskription. Edmund Husserl hatte ja zunächst (1900) die deskriptive Phänomenologie im Auge, wenn er schreibt: „Phänomenologie als Methode gefaßt ist die Methode der beschreibenden Fixierung von psychologischen Tatbeständen." Erst zehn Jahre später gab Husserl diese Gleichsetzung von Phänomenologie und deskriptiver Psychologie auf, indem er fortan den Aspekt der hintersinnigen (transzendentalen) Seinsschichten stärker in den Mittelpunkt der Betrachtung rückte. Seitdem existiert die philosophische Phänomenologie, deren Kern die „Wesensschau" oder „Ideation" (in mehreren Entwicklungsvarianten) bildet. Der theoretischen Ideation stellte später Max Scheler in seiner Wertphilosophie das „Wertfühlen", eine Art nichtverstandesmäßiger Werterfassung, gegenüber. Dabei ging es Scheler vor allem um den Aufweis moralischer und religiöser Wertschichten der Person. Bei Martin Heidegger schließlich wandelte sich die Phänomenologie zur Existenz- bzw. Seinsphilosophie mit dem Ziel der Existenz- bzw. Daseinsanalyse.

Die neueren Ansätze der phänomenologischen Methode in den empirischen Wissenschaften zielen jedoch nicht auf die Wesens-, Wert- oder Seinsschau der philosophischen Phänomenologie ab, sie knüpfen vielmehr an das ursprüngliche Konzept Husserls von der psychologischen Deskription an. Analog dazu wurde der Begriff „Deskriptive Pädagogik" – zum erstenmal wohl von Aloys Fischer konzipiert – geprägt und zu einem methodalen System der Erziehungswissenschaft ausgebaut (vgl. Röhrs 1968, S. 66ff.). Zur Vermeidung begrifflicher Fehldeutungen werden heute meistens Termini wie „Deskription", „Phänographie" oder „Beobachtung" bzw. „Beobachtungsmethode(n)" verwendet. Zwar hat sich international der Begriff „Observation" immer mehr durchgesetzt, doch eignet ihm keineswegs ohne weiteres semantische Eindeutigkeit. Bedeutsamer als die eigentliche Beobachtung ist nämlich bei den Observational Techniques die Beschreibung oder Deskription im ursprünglichen Sinne Husserls. Beobachtung als Methode impliziert also immer auch die *Beschreibung*. Diese hat nach der Forderung Husserls „in radikaler Vorurteilslosigkeit den wahren Gegebenheiten (zu) folgen" und „allen angelernten Theorien zum Trotz zu den Dingen selbst" vorzudringen (a.a.O.). Der hier geforderten Vorurteilslosigkeit in methodischer Hinsicht entspricht die Sachangemessenheit in objektiver Hinsicht. Daß diese Postulate in der Praxis der Deskription bzw. Beobachtungsmethode(n) oft nur unzulänglich berücksichtigt werden, bestreitet wohl niemand. So bestehen neben „sprachlichen Fallen" häufig auch unreflektierte theoretische Voraussetzungen, deren Bewußtmachung Aufgabe eines Wissenschaftlers sein müßte. Dieser sollte besonders kritisch Stil und Wortwahl der (verbalen) Beschreibung kontrollieren und sich in jedem Falle von vorschnellen Wertungen und (Vor-)Urteilen freihalten; dieser Punkt wird uns sogleich noch ausführlicher beschäftigen. Darüber hinaus sollte der wissenschaftlich Tätige seine „Lieblings"-Hypothesen weniger zu *beweisen* als vielmehr zu *überprüfen* suchen. In dieser Intention wird das Aufsuchen von Ausnahmen wichtiger als die Bestätigung. Durch die Formulierung von sog. Nullhypothesen (s. S. 66f.) kann solchen Forderungen praktisch Rechnung getragen werden. Die wissenschaftliche Grundgesinnung, die Husserls Postulate beinhalten, charakterisiert Scheler, wenn er vom Sichhingeben an den Anschauungsgehalt der Dinge, an die Unumstößlichkeit des Gegebenen und Evidenten spricht. „Loslassendes Anschauen" und „liebendes Vertrauen" sollen dies ermöglichen. Ähnlich betont Metzger (1954), man müsse der Wirklichkeit mit Ehrfurcht und Liebe gegenübertreten. Das Vorgefundene sei einfach hinzunehmen, so wie es ist, auch wenn es ungewohnt, unerwartet, unlogisch oder gar „widersinnig" sei und völlig den vertrauten Gedankengängen zu widersprechen scheine. Entschieden wendet sich Metzger damit gegen den eleatischen Grundsatz, daß alles Anschauliche nur (unwirklicher) Schein sei. Die Phänomene selber bestimmen also, was existiert, nicht primär unser Denken oder schlußfolgerndes Bemühen um restfreie und widerspruchslose Aussagen. Diese für jeden Wissenschaftler verpflichtende Hinwendung zur Wirklichkeit bedeutet freilich nicht, daß Gedankenkonstrukte (Hypothesen, Theorien, Modelle) im Wissenschaftsvollzug entbehrlich oder unwichtig seien. Doch besteht vielfach, besonders in der Pädagogik bzw. den Erziehungswissenschaften, ein deutlicher Überhang an solchen Angeboten unter Vernachlässigung des phänographischen bzw. empirischen Erkenntniszugangs (s. noch Roth 1958). In der gegenwärtigen Situation kann somit auf die Methode der Beobachtung qua Phänographie im Hinblick auf eine wirklichkeitsoffene und deutungsfreie Materialgewinnung weniger denn je verzichtet werden. Diese Feststellung gilt auch angesichts der vielen Möglichkeiten anderweitiger operationaler Vorgehensweisen, die von uns in keiner Weise unterschätzt werden sollen; eine sorgfältige Lektüre der folgenden Buchkapitel dürfte ohnehin kaum zu solchen Gedanken verleiten.

*Methoden*charakter, vorurteilsfreie und sachangemessene *Beschreibung* der Phänomene, die in den meisten Fällen erst eine wirklichkeitsoffene und deutungsfreie Materialsammlung erlaubt, bilden somit die Hauptkriterien deskriptiver Phänomenanalysen im Sinne wissenschaftlicher *Beobachtung*. Bevor wir auf die Grundtechniken der Beobachtungsmethode näher eingehen, seien die Hauptphasen dieses Vorgehens erläutert. Jede Form der Beobachtung beinhaltet prinzipiell drei mehr oder weniger voneinander abhebbare *Phasen*, nämlich a) die *Beobachtung* im engeren Sinne, b) die *Beschreibung* oder Deskription, c) die *Deutung* bzw. Interpretation oder Beurteilung.

Unter *Beobachtung* in diesem eingeschränkten Sinne ist nicht bloße Wahrnehmung gemeint. Beobachtung ist immer fixierend gerichtet, d.h. zentriert und selektierend zugleich, insofern man etwas Bestimmtes hierbei ausmachen will. Aus dieser Gesichtspunkthaftigkeit – der Mensch interessiert den Sozialwissenschaftler oder Pädagogen immer nur in einer bestimmten Hinsicht – resultiert eine bestimmte Absicht („absehen von"), worin nach Brown & Ghiselli, Graumann, Traxel u.a. ein wichtiges Strukturelement der Beobachtung zu sehen ist. Die Selektivität der

Beobachtung stellt jedoch kein grundsätzliches Problem der Observational Techniques dar; das wissenschaftsmethodische Problem liegt hier in der Deskription, d.h. in der schiedlichen Trennung von Beschreibung und Beurteilung (Deutung oder Interpretation).
Bei der *Beschreibung,* die der Beobachtung i.e.S. meist unmittelbar folgt, stehen uns, wie Graumann (1964, S. 90ff.) aufweisen konnte, vier Möglichkeiten der Verbalisierung offen: 1. das *verbale* Niveau, 2. das *adverbiale* Niveau, 3. das *adjektivische* Niveau, 4. das *substantivische* Niveau. Anhand einiger Protokollbeispiele wollen wir das Gesagte verdeutlichen:
1. Das *verbale* Niveau. Hierunter ist eine „reine" Beschreibung, d.h. die bloße Beschreibung prozessualer Abläufe, gemeint. Dieser Beschreibungsstil ist arm an Adverbien, Qualifikationen werden weitgehend vermieden. Beispiele: „Der Schüler bewegt den Oberkörper 3mal vor- und rückwärts." – „Er kaut an seinem Bleistift." – „Die Schülerin hält bei der Unterhaltung mit dem Lehrer stets ihren Kopf schief." – „Sie stößt ihre Tischnachbarin am Arm." – „Der Proband fragt den Testleiter nach der Uhrzeit." – usw.
2. Das *adverbiale* Niveau. Damit sind bereits erste Qualifikationen verknüpft, wie folgende Beispiele illustrieren: „Das blinde Mädchen greift *zitternd* nach der Hand der Erzieherin (beim Überqueren der Straße)." – „Die Lehrerin legt *tröstend* den Arm um die Schülerin." – Einen Schritt weiter geht etwa die folgende Notiz in einem Beobachtungs- bzw. Beschreibungsprotokoll: „Das blinde Mädchen greift *ängstlich* nach der Hand der Erzieherin." – Während in den ersten beiden Fallbeispielen leicht ein Konsens zwischen verschiedenen Beobachtern versus Beurteilern möglich sein dürfte, erscheint die dritte Fallskizze schon nicht mehr ganz unproblematisch, insofern hier in die „Beschreibung" eine erste Schlußfolgerung auf das der Handlung zugrunde liegende Motiv mit einfließt. Ob Angst oder Neugierde oder Gesprächsbereitschaft oder Kontaktstreben das blinde Mädchen veranlaßte, nach der Hand der Erzieherin zu greifen, wird aus der Zusammenschau aller Beobachtungsdaten, möglicherweise auch aus dem Kontext der Handlung, am Ende der Deskriptionsphase, also bei der eigentlichen *Interpretation* der beobachteten und deskribierten Befunde (vielleicht) *erschlossen* werden können; der Rückschluß auf die jeweiligen Motive, Interessen u.ä. gehört somit eindeutig zur Beurteilung und darf nicht schon in der Beschreibung dessen, was beobachtet wurde, vorweg- – als Deutung – genommen werden. In Protokollsätzen der vorgenannten Art wird jedoch die Plattform der reinen Beschreibung beobachteter Vorgänge teilweise bereits überschritten und damit die Kontrolle der Richtigkeit so gewonnener Aussagen erheblich erschwert oder gar unmöglich gemacht.
3. Das *adjektivische* Niveau. Mit „Beschreibungen" auf diesem Niveau setzt vollends die Beurteilung ein. Beispiele: „Sein Unterricht ist demokratisch." – „Der Knabe ist sehr schüchtern." – „Der Lehrer ist streng." – „Schüler A ist faul." – Was in diesen Fällen eigentlich geschehen ist, erfährt man erst gar nicht mehr. Beispiele dieser oder ähnlicher Art offenbaren bereits fertige (Vor-)Urteile und haben nichts mehr mit Deskription zu tun.
4. Das *substantivische* Niveau. Auf dieser Stufe münden Protokollsätze in bloße Klassifikationen, wie sie etwa für Typologien kennzeichnend sind. Beispiele: „Er ist ein Linksintellektueller." – „Die jugendliche Delinquentin tat ihm leid." – „Ich hatte einen Neurotiker vor mir." – „Die Rigidität ihres Verhaltens war die Ursache für ihre sozialen Schwierigkeiten." – Diese oder ähnliche Feststellungen

enthalten kaum wissenschaftlich verwertbare Informationen. Wegen ihrer Mehrdeutigkeit und der damit zusammenhängenden fehlenden Kontrollmöglichkeit sowie ihres – vorweggenommenen – Urteilscharakters haben Sätze dieser Art keinen legitimen Platz im Protokoll einer Verhaltensbeobachtung, also noch in der Phase der Deskription. Eine wissenschaftsmethodisch vertretbare *Beschreibung* beobachteter Phänomene (Sachverhalte, Erlebnisse, Verhaltensweisen, Interaktionen usw.) ist demnach nur auf den ersten beiden Beschreibungsstufen möglich, eine ‚Deskription' des dritten oder gar vierten Modus verbietet sich nach dem Grundsatz phänomengetreuer Beobachtung und Beschreibung des Beobachteten (Beobachtbaren). Geschieht dies doch, dann haben wir es nicht mehr mit Beobachtungs-, sondern mehr oder weniger mit reinen Beurteilungsmethoden zu tun. Siehe dazu die folgenden Ausführungen auf S. 32ff. u. 40ff.

Unsere Erörterungen sollten deutlich gemacht haben, daß zur Beobachtungsmethode prinzipiell drei – möglichst auch zeitlich zu diskriminierende – Schritte gehören: die Beobachtung i.e.S., die Beschreibung und die Beurteilung. Um nun in der dritten und letzten Phase aus den Deskriptionsunterlagen ein gesichertes *Urteil* ableiten zu können, ist zuvor ein Zwischenschritt unerläßlich: die Sichtung und Ordnung des Protokoll- bzw. Datenmaterials. Dabei sind folgende Fragen von Bedeutung: Was kommt immer wieder vor? Zeigen sich gleiche oder ähnliche Symptome? Was kommt selten oder überhaupt nicht zum Vorschein? Liegen diskrepante Aussagen vor? Sind originelle (mehr oder minder einmalige) Züge erkennbar, vs. überwiegen die gewohnten (durchschnittlichen) Verhaltensweisen? usw. Die eigentliche Beurteilung oder Interpretation des so geordneten Materials geschieht dann nach einschlägigen Prinzipien der Befunddeutung, nämlich der Bewertung von Einzelbefunden, der syndromatischen Ordnung und der abschließenden Kontextverankerung. Unter dem Aspekt diagnostischer Fragestellung haben wir dieses Vorgehen andernorts ausführlich dargestellt (vgl. Heller 1973, S. 94ff.). Die Deutung der Befunde erfordert immer einen Akt des Schließens, der seinen legitimen Platz am Ende des gesamten Beobachtungsprozesses – nicht früher – hat, und somit immer auch eine Abstraktionsleistung. Neben dem Rückschluß auf die Person, die bzw. deren Verhalten unter diesem oder jenem Frageaspekt beobachtet wurde, erfolgt jetzt abschließend die Einordnung des Gesamtbefundes in den jeweiligen Problemkontext, worin letzlich das Untersuchungsergebnis gipfelt. Werden die aufgewiesenen Schritte der Beobachtung, Beschreibung und Beurteilung strikt auseinandergehalten, dann sind die Ergebnisse der Beobachtungsmethoden, wissenschaftlich gesehen, brauchbare Aussagen, weil sie jederzeit – unter denselben Beobachtungs- und Deskriptionsbedingungen – repliziert und somit auf ihre Richtigkeit (Wahrscheinlichkeit) hin kontrolliert werden können. Andere Verfahrensansätze, die den genannten Kriterien nicht oder nur mangelhaft genügen, können keinen Anspruch auf wissenschaftliche Gültigkeit der Methode und der damit gewonnenen Erkenntnisaussagen erheben. Diese Einschränkung gilt praktisch auch gegenüber der sog. Kasuistik (Beobachtung und Beschreibung von Einzelfällen), deren Wert für die Veranschaulichung und möglicherweise für Erkundungszwecke unbestritten bleibt, die jedoch keine generalisierbaren Aussagen erlaubt.

2.1.2. Exkurs: Das Problem der Beschreibung in der Beobachtungsmethode

Die Phase der Protokollierung (Beschreibung des Beobachteten) ist das Kernstück der Beobachtungsmethode und zugleich ihr problematischster Teil. Wir wollen uns deshalb einschlägigen Fragen der Beschreibung oder Beobachtungsdeskription ausführlicher widmen.

Die Beschreibung kann grundsätzlich in freier oder gebundener Form vorgenommen werden. Bei der *freien* Form erfolgt die Deskription des Beobachteten nach freier, also nicht im voraus festgelegter Wortwahl bzw. Wahl irgendwelcher Symbole. Das Postulat der phänomengetreuen Beschreibung wird hier neben der Frage der Zuverlässigkeit so gewonnener Aussagen zum Hauptproblem. Fehlerquellen der freien Beschreibung liegen einmal im subjektiv begrenzten Wortschatz (des Beobachters) bzw. der aktuellen Verfügbarkeit relevanter Begriffe, zum andern aber auch in der Gefahr, daß „Wichtiges" übersehen versus „Unwichtiges" überbetont wird. So können Einstellungen (Voreingenommenheit, Erwartungshaltungen usw.) und daraus resultierende Selektionstendenzen der Wahrnehmung zu Beobachtungs-/Beschreibungsfehlern führen. Dem sucht die *gebundene* Form der Beschreibung zu begegnen, indem sie Eigenschaftslisten, verbale Beschreibungskategorien oder sog. *rating scales* als Objektivierungshilfen anbietet. Zwar lassen sich auch bei der freien Beschreibung Sicherungen einbauen, z. B. durch Beobachtertraining, Aufstellung mehrerer simultaner Beobachter bzw. Protokollanten (Konsens der Beobachter als Maßstab der Reliabilität) oder durch Verwendung technischer Hilfsmittel (Tonband, Film, Fernsehaufzeichnung), doch erweist sich in diesem Punkte die gebundene Form der Beobachtung fast allemal der freien Form überlegen. Dies gilt insbesondere dann, wenn zur Deskription Beobachtungssysteme (s. unten) Verwendung finden. Die einfachste Form stellt hierbei die Vorgabe *sprachlicher Kategorien,* die dichotom gewählt werden, dar. Zusätzlich kann dann eine Dimensionierung, d.h. eine Einstufung unter dem Aspekt des Mehr oder Weniger, vorgenommen werden. Häufig wird als die beste Form in diesem Zusammenhang die Verwendung einer *Rating*skala angesehen, die eine Zuordnung der Beobachtungseindrücke zu bestimmten Zahlenindizes erlaubt. Das zur Verfügung stehende Kontinuum kann dabei einen Nullpunkt haben oder einfach eine polare Ordnung nach Mehr vs. Minder aufweisen (s. dazu ausführlicher S. 34f. unten). Im Grunde erfolgt via *rating scales* wiederum eine Klassifikation nach Kategorien – analog zur Vorgabe verbaler Kategorien.

Auch von einem *Beobachtungssystem* wird verlangt, daß seine Kategorien deskriptiv und nicht deutend sind. Nur erfolgt hierbei eine Zuordnung des Beobachteten unter bestimmte Kategorien, d.h., die Beobachtung i.e.S. mündet hier in einen Zuordnungsversuch ein. Die Probleme der freien Phänomenbeschreibung, soweit sie die *sprachliche* Unvoreingenommenheit und Treffsicherheit angehen, fallen größtenteils hier weg. Der Vorteil geschlossener Beobachtungssysteme liegt in der größeren intersubjektiven Übereinstimmung der Beobachter und somit auch der Beurteiler. Beispielsweise erzielte man mit dem Kategoriensystem von Bales (1951), das zur Beschreibung von Interaktionen dient, bei geschulten Beobachtergruppen Reliabilitätskoeffizienten zwischen 0.75 und 0.95 (interindividuelle Reliabilitätskontrolle). Größter Nachteil aller gebundenen Formen der Beobachtung ist die Tatsache, daß den meisten Kategoriensystemen eine bestimmte Theorie mehr oder weniger explizit zugrunde liegt, wodurch unter Umständen die geforderte Vorurteilsfreiheit beeinträchtigt und somit wiederum die Objektivität der Beobachtung in Gefahr gebracht wird. Andererseits kann die Verwendung von Kategorienlisten die erforderliche und angestrebte Vollständigkeit der Erfassungsdimension(en) wesentlich verbessern und so verhindern, daß wichtige Aspekte unberücksichtigt bleiben. Im einzelnen ergeben sich bei der Benutzung von Beobachtungssystemen folgende Probleme (nach König 1962; Graumann 1965, 1966; v. Cranach & Frenz 1969 u.a.).

1. Problem des *Umfanges:* Manche Kategoriensysteme versuchen das Gesamtverhalten, andere nur bestimmte Verhaltensdimensionen zu erfassen. Bei einseitigen Konzepten besteht wiederum die Gefahr, daß bedeutsame Beobachtungen „vergessen" werden oder aber die Proportion zwischen Kategorisiertem und Nichtkategorisiertem (z.B. durch inadäquate Prozentangaben) undeutlich bleibt.
2. Problem der *Schlußfolgerung* (Inferenz): Je nach dem Grad der geforderten Inferenz eines Beobachtungssystems werden bei der Zuordnung Schlußfolgerungen überhaupt *nicht* (Bsp.: „liest viele Bücher" – „liest wenige Bücher" – „liest keine Bücher") versus mehr oder weniger *notwendig* (Bsp.: „geistig interessiert" – „geistig nicht interessiert"). Im letzten Fallbeispiel wird vom Bücherlesen, von bevorzugten vs. defizitären Beschäftigungsthemen, Arbeiten usw. *rückgeschlossen* auf die geistige Interessenstruktur, auf Motive, Einstellungen u.ä. Daß solche Inferenz-Kategorien die Kontrollierbarkeit der Ergebnisse erschweren, dürfte nach dem bisher Ausgeführten ohne weitere Erklärungen einleuchten; siehe auch S. 40ff.
3. Problem der *Aspektfülle:* Graumann meint hiermit die Entscheidung darüber, wieviel des Beobachtbaren der thematisierten Beobachtung anheimgegeben werden soll, z.B. nur die Mimik, die Pantomimik, Gebärden oder noch mehr. Diese Frage ist nicht identisch mit dem Problem des Umfanges (Nr. 1).
4. Problem der *diskreten* vs. *kontinuierlichen* Kategorien: Auf die Möglichkeit, qualitative Kategorien zu dimensionieren, werden wir noch zu sprechen kommen (vgl. Kap. 2.2.6.3). So kann man beispielsweise „Trauer" oder „Ängstlichkeit" beim Kinde in den verschiedenen Steigerungsstufen verbal beschreiben (dimensionieren): „N.N. runzelt die Stirn" – „läßt die Mundwinkel hängen" – „äußert Worte des Kummers" – „weint bitterlich" – „bebt am ganzen Körper". – Oder: „N.N. geht einen Schritt zurück" – „hebt schützend den Arm vors Gesicht" – „schüttelt sich" – „bleibt reglos stehen" – „starrt mit weit geöffneten Pupillen". – Solche Deskriptionen beinhalten diskret abgestufte Kategorien, wobei die Hauptproblematik im Abstand der gewählten Kategorien liegt. Gewöhnlich ist hier keine Äquidistanz gegeben. Weitere Beispiele finden sich im Kap. 2.1.3.2 unten.
5. Problem der Größe der zu beobachtenden *Einheiten (units):* Sollen nur ganze Sätze oder auch Satzteile, Wörter, Phoneme usw. versus Handlungsganzheiten oder auch Handlungsteile beobachtet und beschrieben werden? Die Größe der *Units* wird am besten *vor* der angesetzten Beobachtung von den einzelnen Beobachtern durch Konsensus festgelegt, wobei sich freilich oft eine gewisse Willkür kaum vermeiden läßt. So erhebt sich im konkreten Falle etwa die Frage, ob Vollzüge wie Essen, Geschenk-Auspacken, Tischtennisspiel, Unterrichtsstunde, Lösung eines Denkproblems u.ä. jeweils als Handlungsganzes oder selektiv nach ganz bestimmten Details beobachtet und beschrieben werden sollen. Wenn schon empirische Untersuchungen über den Beobachtungsgegenstand vorliegen, kann

man entsprechende *Units* operational definieren als das, was einer bestimmten Kategorie zugeordnet werden kann.

6. Problem der *Reichweite* von Beurteilungssystemen: Hiermit ist der Anwendungsbereich der Kategoriensysteme und somit die Generalität vs. Spezifität ihrer Gültigkeit angesprochen. So gibt es Systeme für sehr spezifische Zwecke, z.B. zur Erfassung von Interaktionen, Konferenzanalysen, Diskussionen, Unterrichtsstunden usw. Wieder andere Systeme treten mit einem sehr weiten Gültigkeitsanspruch auf. Prinzipiell gilt – analog zur Testdiagnostik – auch hier die Regel: Je weiter und allgemeiner der Anwendungsbereich definiert ist, desto ungesicherter ist der Informationswert des Kategoriensystems prinzipiell zu veranschlagen; je enger und spezifischer das Anwendungsgebiet ist, desto gegenstandsadäquatere und zuverlässigere Resultate sind zu erwarten.

7. Problem der *Aktivität* des Beobachters: Zwei Modi sind hier besonders relevant: a) die mittelbare oder indirekte Beobachtung, b) die unmittelbare oder direkte Beobachtung. *Indirekte* Beobachtungsverfahren sind nach König, Thomae u.a. Verfahren, in denen der Rückschluß auf das Verhalten von „Verhaltensniederschlägen" (Tagebüchern, Briefen, Reden, Kunstwerken usw.) aus erfolgt. Oft werden sämtliche Schätzverfahren bzw. Beurteilungstechniken, bei denen streng-genommen die Beschreibungsphase, oft auch die Beobachtung selbst, fehlt, hierunter subsumiert (vgl. Remmers 1963). Als *direkte* Verfahren werden Beobachtungstechniken im bisher definierten Sinne verstanden. Im Hinblick auf unsere Definitionskriterien der Beobachtungsmethode mit ihren drei Phasen der Beobachtung, Beschreibung und Beurteilung läßt Graumann obige Unterscheidung in direkte und indirekte Verfahren nicht gelten. Beobachtung qua Methode ist demnach immer direkte Beobachtung (unter Einschluß der Beobachtungs- und Beschreibungsphase). Die sog. indirekten Verfahrensmodi werden deshalb im Rahmen der *Beurteilungs*techniken besprochen (vgl. Kap. 2.1.3.4).

Der Vorteil der Verwendung geschlossener Kategoriensysteme liegt – neben der dadurch ermöglichten Vergleichbarkeit der Beobachtungsdaten – vorab in der Gewinnung quantitativer *Scores*, die statistisch weiterverarbeitet werden können. Voraussetzung hierfür ist freilich eine ausreichende Zuverlässigkeit des verwendeten Kategoriensystems. Die Reliabilität bei freier Beobachtungs- und Beschreibungsweise ist im allgemeinen schwerer nachzuweisen. Wenn beide Formen von den sonstigen Voraussetzungen her möglich erscheinen, wird man deshalb in der Regel dem geschlossenen Kategoriensystem den Vorzug geben, um die via Beobachtung gewonnenen Erkenntnisaussagen optimal abzusichern. Die größte Verläßlichkeit kann erfahrungsgemäß mit Hilfe von Skalierungen erzielt werden.

In der Literatur werden sowohl verbale Skalen als auch nonverbale (graphische und quantitative) Skalen aufgeführt (vgl. Guilford 1954, Ryans 1960, Medley & Mitzel 1963 bzw. Schulz u.a. 1970, Remmers 1963 bzw. Tent 1970, Hasemann 1964, v. Cranach & Frenz 1969 u.a.). Guilford (a.a.O.), der wohl die umfassendste Darstellung der Schätzskalen *(rating scales)* bietet, unterscheidet folgende Typen: *numerical scales, graphic scales, standard scales, rating by cumulated points* und *forced-choice technique*. Tent (a.a.O.) erwähnt noch die Skalierungsmethoden mit Hilfe von Strichlisten und Auswahlantworten. Die in einem Beobachtungssystem verwendeten *Rating*skalen enthalten zweckmäßig mindestens 3 und nicht mehr als 9 bis 11 Stufen. Am häufigsten werden 3er- und 5er- bzw. 7er-Skalen benutzt, z.B.: „N.N. braucht Hilfe beim Lernen: viel – durchschnittlich – wenig" – „Denkleistung: sehr hoch – hoch – mittel – niedrig – sehr schwach" – „Qualität des Unterrichtsversuchs von N.N.: ausgezeichnet – sehr gut – gut – befriedigend – ausreichend – mangelhaft – ungenügend". – Bei *numerischen* Skalen werden für die einzelnen Skalenbereiche nur Punkte angegeben, wofür die üblicherweise zur Einschätzung der Schülerleistung verwendete Notenskala ein klassisches Beispiel bietet. Freilich: Ob man die Skalenbereiche numerisch (Zensuren 1, 2, 3, 4, 5, 6 bzw. 1, 2, 3, 4, 5) oder verbal (sehr gut, gut, befriedigend usw.) definiert, ist hier im Grunde gleichgültig und ändert an der Skalenqualität kaum etwas – sofern das Problem der Äquidistanz nicht berührt oder einseitig betroffen wird. Diese Kautel gilt auch für die *graphischen* Skalen, z.B.:

Konzentration im Unterricht:

gut	durchschnittlich	gering

oder

Systematik des Vortrags:

sehr ausgeprägt	ziemlich stark	mittelmäßig	kaum erkennbar	überhaupt nicht vorh.

Die *Rating*skalen können unipolar oder bipolar angeordnet sein, z.B. wird der Grad der Ausdauer, der Mitarbeit, der Konzentration usw. den Intensitätsstufen 0 (extrem schwach) bis 5 (sehr stark ausgeprägt) zugeordnet = unipolare Anordnung; oder es erfolgt die Einschätzung der Lehrereigenschaften von 1 bis 7 in bezug auf folgende Merkmale: gerecht (7) vs. ungerecht (1), anregend (7) vs. langweilig (1), systematisch (7) vs. planlos (1), demokratisch (7) vs. autokratisch (1) usw. = bipolare Anordnung.
Die Zuordnung des beobachteten Verhaltens mit Hilfe solcher Skalen ist relativ einfach und dementsprechend die Übereinstimmung unter den Beurteilern im allgemeinen hoch. Prinzipiell ist freilich auch bei der Verwendung von *Rating*skalen darauf zu achten, daß die Beschreibungsphase nicht übergangen wird, d.h. vorschnell eine Beurteilung stattfindet. Diese Gefahr ist bei vielen *Rating*skalen konkret – nicht grundsätzlich – gegeben, wofür wir mit unseren letzten Beispielen schon erste Belege lieferten. Weitere Beispiele bringen wir in Kap. 2.1.3.4 unten.

2.1.3. Formen der Beobachtungsmethode

2.1.3.1. Allgemeine Einteilungskriterien

Die einzelnen Beobachtungstechniken lassen sich unter verschiedenen Gesichtspunkten gruppieren. Thomae (1954, 1970) unterscheidet beispielsweise die Gelegenheitsbeobachtung von der systematischen Beobachtung und der Dauerbeobachtung. Wird das Beobachtungsverfahren nur bei aktuellen Anlässen, z.B. während des Unterrichts (vom unterrichtenden Lehrer), im Rahmen einer Testerhebung (vom Testleiter), beim Spielen des Kleinkindes (von der Mutter) usw., also mehr oder weniger unsystematisch zum Festhalten besonderer Vorkommnisse, gelegentlicher Problemfragen, Verhaltensauffälligkeiten u.ä. eingesetzt, so spricht Thomae von *Gelegenheits*beobachtung. Diese Methode der Gelegenheits-oder Zufallsbeobachtung *(anecdotal record)* dürfte am häufigsten in der pädagogisch-psychologischen Praxis, etwa als Zusatzverfahren bei der Schülerbeurteilung, oder im Erkundungsansatz (wenn es sich um einen Vorstoß in wissenschaftliches Neuland handelt) zur Anwendung kommen. *Systematische* Beobachtung *(systematic observation)* hingegen bezieht sich auf den systematisch geplanten Verfahrensansatz der Beobachtung. Diese Form stellt vor allen anderen im Rahmen wissenschaftlicher Untersuchung die Methode der Wahl dar. Von *Dauer*beobachtung spricht Thomae, wenn die Beobachtung über einen längeren Zeitraum hinweg – mehr oder weniger kontinuierlich – angesetzt wird, z. B. über mehrere Tage (und Nächte) oder gar Wochen und Monate. In diesen Fällen sind natürlich immer mehrere Beobachter oder Beobachterteams, die sich ablösen können, erforderlich. Auch diese Form der (systematischen) Beobachtung wird fast ausschließlich in der wissenschaftlichen Forschung eingesetzt, z. B. zum Studium der Sprachentwicklung oder des Spielverhaltens bei Kleinkindern, zur Erfassung bestimmter Kognitionsvorgänge, gruppendynamischer Beziehungen, Interaktionsprozesse u. dgl. m. Sogenannte Längsschnittuntersuchungen (Longitudinalstudien) der Entwicklungspsychologie (vgl. Nickel 1972) sind fast immer auf diesen Verfahrensansatz angewiesen. Einen Sonderfall stellt in diesem Zusammenhang die *fraktionierte* (Dauer-)Beobachtung dar, wie sie beispielsweise von Thomae und Mitarbeitern praktiziert wurde (vgl. Coerper u. a. 1954 sowie Hagen u. a. 1962). Dabei wird die Beobachtung systematisch in bestimmten Abständen und zu genau festgelegten Zeiten (Tagen, Wochen, Monaten und Jahren) durchgeführt, so daß man mehr oder weniger repräsentative Verhaltensstichproben erhält. Intervallbeobachtungen dieser Art werden gewöhnlich ebenso zur Kategorie der „Dauerbeobachtung" gerechnet wie Veranstaltungen rund um die Uhr *(full time)*. Der größeren Eindeutigkeit wegen sollte man deshalb nur zwischen *unsystematischer* (Gelegenheits-)Beobachtung und *systematischer* Beobachtung unterscheiden. Letztere Form kann dann noch hinsichtlich der zeitlichen Erstreckung unterteilt werden in systematische *Kurzzeit*beobachtung, systematische *Langzeit*beobachtung und systematische *Dauer*beobachtung. Eine Variante stellt in diesem Zusammenhang das im angelsächsischen Raum sehr verbreitete Verfahren des *time sampling* dar. Damit ist die Aneinanderkettung zahlreicher, oft auf wenige Minuten beschränkter Kurzzeitbeobachtungen über einen längeren Zeitraum hinweg (z. B. 24mal 5 Min. innerhalb eines Tages) gemeint. Im Hinblick auf die Art der Protokollierung – ohne vs. mit Beobachtungs-/Beschreibungsschema – kann man schließlich von „freier" versus „gebundener" Beobachtungstechnik sprechen. Diese Klassifizierung entspräche in etwa der Einteilung Hofstätters oder Königs in unkontrollierte und kontrollierte Beobachtung, wenngleich die Beobachtung als Wissenschaftsmethode im eingangs definierten Wissenschaftsverständnis immer ein kontrollierbarer Verfahrensansatz sein muß.

Eine Ordnung der *Observational Techniques* kann auch vom Beobachtungs*gegenstand* her vorgenommen werden. Dann reden wir von *Verhaltens*beobachtung und *Erlebnis*beobachtung, denen gelegentlich als dritte Variante die *Ausdrucks*beobachtung hinzugesellt wird (vgl. Traxel 1964). Da die Unterscheidung in Erlebnis- und Verhaltensbeobachtung gerade für die Sozialwissenschaften, insbesondere die Psychologie, von fundamentaler Bedeutung ist, wird der Erörterung dieser Methoden jeweils ein eigener Abschnitt gewidmet (Kap. 2.1.3.2 u. 2.1.3.3). Ebenso werden wir die eigentlichen *Beurteilungs*verfahren, also jene Methoden,

bei denen die Phase der Beobachtung und/oder Beschreibung mehr oder weniger vernachlässigt wird bzw. fehlt, als gesonderte Gruppe weiter unten behandeln (Kap. 2.1.3.4) und im Zusammenhang damit auch auf die Beurteilungs*fehler* zu sprechen kommen.
Schließlich ist für die sozialwissenschaftliche Forschung noch eine Unterscheidung wichtig: Unter dem Gesichtspunkt der Aktivität des Beobachters (s. S. 34) kann die *teilnehmende* Beobachtung der *nichtteilnehmenden* Beobachtung gegenübergestellt werden. Die *teilnehmende Beobachtung (participant observation)* ist eine Komposition aus Selbst- und Fremdbeobachtung, d.h., der Beobachter versucht hierbei die Effizienz der Verhaltensbeobachtung durch das eigene Miterleben, Mitfühlen, Mithandeln usw. zu steigern. Diese Methode hat sich besonders in der Ethnologie bzw. bei der Erforschung geschlossener Sozialgruppen bewährt. In allerjüngster Zeit hat man dieses Verfahren auch in der Bildungsforschung eingesetzt, z.B. zur Aufdeckung von Motivationsbarrieren, Sozial- und Mentalitätssperren (gegen weiterführende gehobene oder höhere Schulbildung) usw. (vgl. Heller 1970). Das Gegenstück zur teilnehmenden Beobachtung bildet die *nichtteilnehmende Beobachtung (witnessing observation)* mit zwei Varianten: Der Beobachter ist entweder passiv neutral, jedoch unmittelbar dabei, oder nur Zeuge des Geschehens, d.h. nur mittelbar (z.B. mit Hilfe der Einwegscheibe) dabei. Die Entscheidung darüber, welche der genannten Methoden zum Einsatz kommen soll, hängt somit von der jeweiligen Situation, der Untersuchungspopulation und nicht zuletzt vom Untersuchungsziel ab.

2.1.3.2. Verhaltensbeobachtung

Zweifellos die wichtigste Form der Beobachtung in der pädagogischen Psychologie ist die Verhaltensbeobachtung. Hierbei wird das Verhalten *anderer* beobachtet und beschrieben, um daraus Schlußfolgerungen der verschiedensten Art und Zweckbestimmung zu gewinnen. Die Methode der Verhaltensbeobachtung ist vor allem dann indiziert, wenn die Probanden oder Versuchspersonen als Informationsquelle weitgehend oder ganz ausfallen und/oder der Einsatz anderer Methoden (Interview, Test, Film, Tonband) die zur Untersuchung anstehenden Phänomene beeinträchtigen bzw. verändern würde. In der Sozialpsychologie (zur Untersuchung gruppendynamischer Prozesse, Interaktionsformen u.ä.) und der pädagogischen Psychologie (zur Untersuchung des Erziehungsverhaltens, des Unterrichts, bei diagnostischen Untersuchungen u.ä.) bzw. Entwicklungspsychologie usw. ist man deshalb sehr oft auf Verhaltensbeobachtungen angewiesen. Für die Verhaltensbeobachtung gelten prinzipiell die gleichen Anforderungskriterien wie für alle anderen Formen der Beobachtung (s. oben Kap. 2.1.1). Auch hier wird sehr oft gegen das Postulat der phänomengetreuen (Verhaltens-)Beschreibung verstoßen. Die nachstehenden Protokollbeispiele mögen nicht nur zur Veranschaulichung dieses Problems beitragen, sondern zugleich (positive) Anregungen setzen. Das erste Beispiel, das wir Thomae (1970, S. 29f.) entnehmen, ist ein Auszug aus dem Protokoll einer Gelegenheitsbeobachtung und bezieht sich auf ein knapp 4jähriges Mädchen, das auf einer offenen Veranda in der Nähe eines großen Kinderheimes mit Spielen beschäftigt ist. Die Mutter sitzt in der Nähe und protokolliert ihre Beobachtungen.

Beginn: 16.22
Hat ihren Stoffhasen an den Ohren, dreht ihn herum. „Jetzt bin ich umgefallen!" Geht zu ihrem Stühlchen. „Jetzt mach ich ein Gespenst!" „Ist ein Mädel, der Stuhl!" Tanzt mit ihm herum, trägt ihn quer. „Schau, jetzt hat das ein Bauch!" „Jetzt bring ich das woanders hin!" „Mutti, wenn ich das draufstelle, was ist dann?" Setzt sich, zieht dann ihren Stuhl hinter sich her. Sieht Heuschrecke an der Wand. „Schau mal, was ist das? So ein grünes?" (Wird erklärt.) „Da hab ich Angst, Mutti, hast Du Angst?... Kann das auf meinen Stuhl raufkrabbeln? Kann er krabbeln? Mäuse können schon krabbeln! Mutti, weißt Du, was ich jetzt mach? In der Küche einen Purzelbaum!" (Gegenvorschlag) „Doch, das mach ich aber!"

16.25 (Verschwindet in der Küche.)
Rückkehr nach 3 Minuten. „War in der Küche drin, in Kinderzimmer, meine Kiste ist da, kann ich mir eine schöne Wohnung machen. Ball und alles. Ist so viel drin, schöne Kullern (Kugeln). Ja, das genügt auch." Hängt den Sack mit Spielzeug an den Stuhl, nimmt einen Stoffball, den sie in der Kiste fand, wirft ihn auf die Erde, steckt ihn in den Sack, holt ihn wieder heraus, steckt ihn hinter die Beobachterin auf deren Stuhl.
„Mach ich ein Gespenst, siehst Du, war ein Gespenst!" (Grimassiert dabei.) Wirft den Ball mehrmals auf die Erde, steckt ihn wieder in den Sack. Nimmt ihre Kasperpuppe, betrachtet sie. Faßt sie an die Nase, dann an die Mütze. Singt: „Seine Frau..." Setzt sich auf den Rand ihrer Spielkiste, rutscht schließlich ganz hinein. „Nun bin ich hineingefallen, nun kann ich nicht mehr aufstehen",

16.30 (singend:) „nun kann ich nicht mehr aufstehen, nun kann ich nicht mehr aufstehen, nun kann ich nicht mehr aufstehen."
Erhebt sich, wirbelt ihre Kasperpuppe herum. „Ist kein Gespenst nicht!" In der anderen Hand hält sie ihren Stoffhasen. Wirbelt jetzt Hase und Kasper gleichzeitig herum, legt sie dann in die Kiste. Nimmt sich ein Holzauto aus der Kiste. Schiebt das Auto auf dem Boden hin und her, dazu sehr angestrengt Laute ausstoßend: „brrrr, pffff, brrrrr." „Bernd hat gesagt, ich soll ihm das Auto schenken. Hab ich ihm aber nicht gegeben." Läßt das Auto stehen, nimmt aus der Kiste eine Stoffziege. „Warum kann die nicht laufen?"

16.33 Unvermittelt: „Ich mach mir eine Wohnung. Da brauch ich ein Brett!" Wird von einer Spielkameradin gerufen, läuft zu ihr hin, nach zwei Minuten zurück.
„Kennst Du die Hanna?" Nimmt ihr Stühlchen. „Jetzt mach ich mir mit Stuhl eine Wohnung!" Schleppt das Stühlchen in eine andere Ecke der Veranda. (Singend:) „Das Gespenst, das Gespenst!"

16.40 Trägt die Spielkiste ebenfalls in die andere Ecke. „Meine Wohnung mach!" usw.

Das nächste Beispiel entstammt ebenfalls einer Gelegenheitsbeobachtung, die im Rahmen einer Testuntersuchung protokolliert wurde (vgl. Heller 1973, S. 93).

K. zeigte sich sehr wissensdurstig und mitteilsam. Überhaupt wollte er lieber mit dem Tester plaudern als sich den geforderten Testaufgaben zuwenden. Während der eigentlichen Testdurchführung schweifte er ständig ab, so daß der Zeitaufwand (für die Testung) außergewöhnlich groß war. Andererseits erwies er sich als gutmütig und zeigte sich fortwährend behilflich; so wollte er ständig Bücher, die Tasche und ähnliche Gegenstände des Testleiters tragen.

Während das erste Beobachtungsprotokoll ein reiner Aktionsbericht ist, enthält das zweite Beispiel schon einige Qualifikationen und ist von Schlußfolgerungen nicht immer ganz frei (z. B. „gutmütig"). Gerade wenn die Verhaltensbeobachtung als Zusatzverfahren – etwa bei Testuntersuchungen oder während des Unterrichts

– eingesetzt wird, wo die Hauptaufmerksamkeit des Beobachters (Testers, Lehrers usw.) anderen Gegenständen gilt und eine sofortige Deskription des Beobachteten oft nicht möglich ist, schleichen sich allzuleicht „Vor-Urteile“ der genannten Art ein. In diesen Fällen empfiehlt sich deshalb die Verwendung eines Beobachtungs- bzw. Beschreibungsschemas, das nicht nur ökonomische Vorteile für die Protokollierung bietet, sondern zugleich als Objektivierungshilfe für die Verhaltensdeskription betrachtet werden kann. Am besten eignen sich hierzu operationalisierte Verhaltenskategorien. Beispielhaft bringen wir nachstehend die Methode von Puckett zur Deskription der „Schülerbeteiligung am Unterricht“ (zit. bei Medley & Mitzel 1963 bzw. Schulz u.a. 1970, S. 664f.).

Schüler meldete sich
Schüler meldete sich und wurde vom Lehrer drangenommen
Schüler meldete sich, wurde drangenommen und gab eine Einwortantwort
Schüler meldete sich, wurde drangenommen und gab eine befriedigende Antwort
Schüler meldete sich, wurde drangenommen und gab eine gute Antwort
Schüler meldete sich, wurde drangenommen und gab eine sehr gute Antwort
Schüler kam dran, ohne sich gemeldet zu haben
Schüler kam dran, ohne sich gemeldet zu haben, und gab eine Einwortantwort
Schüler kam dran, ohne sich gemeldet zu haben, und gab eine befriedigende Antwort
Schüler kam dran, ohne sich gemeldet zu haben, und gab eine gute Antwort
Schüler kam dran, ohne sich gemeldet zu haben, und gab eine sehr gute Antwort
Schüler kam dran, ohne sich gemeldet zu haben, und gab keine Antwort
Schüler stellte eine Frage
Schüler sprach ohne Aufforderung des Lehrers

Im Original findet sich noch ein Kodierungsvorschlag für die einzelnen (möglichen) Reaktionen, so daß die Arbeit des Protokollierens erheblich vereinfacht wird, z.B.: · ⊙ ⊙ ⊙– usw. In dieser Form hat sich die Erfassung der Schülermitarbeit im Unterricht bewährt. Auf ähnliche Weise kann man Verhaltensweisen des Lehrers und der Erzieher, das sozialpsychologische bzw. pädagogische Klima in der Schule und Familie bzw. im Internat, das Freizeitverhalten usw. via Verhaltensbeobachtung analysieren. Zahlreiche Beispiele hierfür finden sich bei Tausch & Tausch (1970), Schulz u.a. (1970), Walter (1973) u.a.

2.1.3.3. Erlebnisbeobachtung

Bei der Erlebnisbeobachtung ist die persönliche Art unseres *eigenen* Erlebens Gegenstand der Beobachtung und Beschreibung. In der Psychologie hat man lange Zeit die *Introspektion* (Selbstbeobachtung) als *via regia* angesehen, weil sie die einzige Methode darstellt, die die Psychologie exklusiv für sich beanspruchen kann, während sie die anderen Methoden mit einer Reihe fremder Wissenschaften teilen muß. In der heutigen Psychologie ist die Bedeutung der Erlebnisbeobachtung gegenüber der Verhaltensbeobachtung deutlich zurückgetreten. Wir möchten freilich die Einwände von James und anderen Kritikern der Introspektion oder, genauer: Retrospektion – alle Selbstbeobachtungen sind mehr oder weniger unmittelbar nach dem Erlebnisvorgang einsetzende *Rück*beobachtungen, eine synchrone Beobachtung ist hier im strengen Sinne nicht möglich –, wonach diese prinzipiell keine wissenschaftliche Methode sein könne, nur bedingt gelten lassen. Nach James, der die Selbstbeobachtung als reine Privatangelegenheit betrachtete, kommt dieser deshalb keine Wissenschaftlichkeit zu, weil die Mitteilbarkeit nicht objektiviert und somit die Kontrollierbarkeit der Aussagen nicht gewährleistet werden könne. Demgegenüber muß festgestellt werden, daß die Methode der Selbstbeobachtung ausnahmsweise dann zum wissenschaftlichen Verfahren avancieren kann, wenn mitgeteilt wird, was unter welchen Bedingungen wie erlebt wurde. Unter dieser Voraussetzung ist nicht nur die Publizität gewährleistet, sondern auch die Vergleichbarkeit der Untersuchungsergebnisse prinzipiell möglich. Die Übereinstimmung mit fremden Selbstbeobachtern wird hierbei zur entscheidenden Kontrollinstanz. Die auf dem Wege der Introspektion bzw. Retrospektion gewonnenen Untersuchungsergebnisse sind so – indirekt – kontrollierbar, womit das wichtigste Wissenschaftskriterium erfüllt wäre. Sofern man auf das Erleben als eigenen Gegenstand der Psychologie nicht verzichten möchte, wird man immer auch auf die Erlebnisbeobachtung qua (direkten) Zugang zu den Gefühlen, Stimmungen, Träumen, Vorstellungen usw. angewiesen sein. Deren Konstituierung als Wissenschaftsmethode ist also – wie wir zeigen konnten – grundsätzlich möglich, wenngleich die Gefahrenquellen, die in der konkreten Anwendung dieser Methode liegen, nicht unterschätzt werden dürfen.

2.1.3.4. Beurteilungstechniken

Von vielen Autoren werden die Beurteilungstechniken zu den – indirekten – Beobachtungsverfahren gerechnet, was jedoch nicht ganz konsequent ist. Nach unserer Kriterienbestimmung gehören ja zur Beobachtungsmethode stets drei voneinander abhebbare Phasen: Beobachtung (im engeren Sinne), Beschreibung und Beurteilung (interpretative Schlußfolgerung). Da bei den als Beurteilungsverfahren angesprochenen Techniken die ersten beiden Phasen meistens fehlen – zumindest fehlt durchgängig die Phase der Deskription –, sollte man diese besser zu einer eigenen Gruppe zusammenfassen.

Beurteilung via Schätzskalen (Rating)

Bei den Beurteilungstechniken lassen sich mehrere Arten unterscheiden. Zum ersten Typ wären jene Beurteilungsformen zu rechnen, bei denen – analog zu den Beobachtungstechniken – Verhaltens*abläufe* im Mittelpunkt der Gegenstandserfassung stehen, die eigentliche Beschreibungsphase jedoch – im Gegensatz zur Beobachtungsmethode – wegfällt oder wenigstens nicht manifest wird. Die Phase der direkten Beschreibung des Beobachteten wird hier übersprungen. Folgende Beispiele mögen das Vorgehen im einzelnen erläutern. Zunächst bringen wir ein Beurteilungsschema von Ryans, das dem Handbuch der Unterrichtsforschung (zit. nach Remmers 1963 bzw. Tent 1970, S. 866f.) entnommen ist. Zugleich steht dieses Konzept paradigmatisch für eine numerische Schätzskala (s. S. 35 oben).

Unterrichts-Beurteilungsbogen

zur Untersuchung von Lehrereigenschaften (nach Ryans 1960, S. 86f.)

Lehrkraft ______ Nr. ____ Geschl. ____ Klasse oder Fach ______ Datum ________

Ort ____________ Schule ____________ Zeit ____________ Beobachter ____________

Schülerverhalten:								
1. gleichgültig	1	2	3	4	5	6	7	lebhaft
2. störend	1	2	3	4	5	6	7	mitarbeitend
3. unsicher	1	2	3	4	5	6	7	selbstsicher
4. unselbständig	1	2	3	4	5	6	7	selbständig
Lehrerverhalten:								
5. ungerecht	1	2	3	4	5	6	7	gerecht
6. autokratisch	1	2	3	4	5	6	7	demokratisch
7. unzugänglich	1	2	3	4	5	6	7	zugänglich
8. verständnislos	1	2	3	4	5	6	7	verständnisvoll
9. schroff	1	2	3	4	5	6	7	freundlich
10. langweilig	1	2	3	4	5	6	7	anregend
11. stereotyp	1	2	3	4	5	6	7	originell
12. gleichgültig	1	2	3	4	5	6	7	lebhaft
13. farblos	1	2	3	4	5	6	7	anziehend
14. ausweichend	1	2	3	4	5	6	7	verantwortungsbewußt
15. unstetig	1	2	3	4	5	6	7	gleichbleibend
16. erregbar	1	2	3	4	5	6	7	beherrscht
17. unsicher	1	2	3	4	5	6	7	selbstsicher
18. planlos	1	2	3	4	5	6	7	systematisch
19. unbeweglich	1	2	3	4	5	6	7	anpassungsfähig
20. pessimistisch	1	2	3	4	5	6	7	optimistisch
21. unreif	1	2	3	4	5	6	7	ausgewogen
22. kleinlich	1	2	3	4	5	6	7	großzügig

Die skalierte Eigenschaftsliste von Ryans erinnert an das Polaritätsprofil *(semantic differential)*, eine erstmals von Osgood eingeführte Technik zur quantitativen Analyse von Begriffsinhalten. In der Quantifizierbarkeit solcher Beurteilungssysteme liegt denn auch ihr größter Vorteil, während die intersubjektive Gültigkeit und damit die Objektivität und Zuverlässigkeit solcher Schätzurteile – also ihre Kontrollierbarkeit – als Problem bestehenbleiben. „Die Konstruktion eines solchen Schätzskalensystems ist nicht weniger kompliziert als die Verfahren der direkten Registrierung. Was an einfacherer Anwendbarkeit gewonnen wird, geht durch schwierige Normierung und schwer erfüllbare Gütekriterien verloren. Wir sparen das direkte Protokollieren und Auszählen der Häufigkeiten der beobachteten Ereignisse, und wir handeln uns die Mühe ein, normier- und objektivierbare ‚persönliche Konstrukte' zu finden, die auch noch unsere Ansprüche an Reliabilität und Validität erfüllen" (Dieterich 1972, S. 97). Der Einsatz solcher oder ähnlicher Beurteilungssysteme erfordert – genauso wie die Verwendung der Beobachtungsmethode im freien oder gebundenen Verfahren – ein gründliches Training der Beurteiler *(Rater)*, und zwar sowohl hinsichtlich der Beobachtung als auch hinsichtlich der Beurteilung, d.h. der Interpretation des Beobachteten. Als sehr hilfreich erwiesen sich in diesem Zusammenhang operationale Erläuterungen der einzelnen Kategorien, wie sie Ryans seinen Beurteilern anbot. Auszugsweise seien hier zwei Beispiele wiedergegeben.

1. Urteilsskala „gleichgültig – lebhaft“ in bezug auf das Schülerverhalten:
gleichgültig: teilnahmslos; spielte(n) den Gelangweilten; war(en) nur mit halbem Herzen bei der Sache; unruhig; schweifende Aufmerksamkeit; kam(en) nur langsam in Gang usw.
lebhaft: darauf bedacht, dranzukommen und teilzunehmen; beachtete(n) den Lehrer aufmerksam; arbeitete(n) konzentriert; schien(en) eifrig mitzumachen; sofort bereit, sich zu beteiligen usw.
2. Urteilsskala „störend – mitarbeitend“ in bezug auf das Schülerverhalten:
störend: ungezogen untereinander und/oder dem Lehrer gegenüber; unterbrach(en), störte(n), beanspruchte(n) Aufmerksamkeit; eigensinnig, widerspenstig; Weigerung mitzuarbeiten; streitsüchtig, reizbar; gebrauchte(n) Schimpfnamen und/oder klatschte(n); unvorbereitet usw.
mitarbeitend: höflich, kooperativ, freundlich untereinander und zum Lehrer; erledigte(n) Aufgaben ohne Klagen oder Anzeichen von Verdrossenheit; sprach(en) mit beherrschter Stimme; nahm(en) Hilfe und Kritik bereitwillig an; bat(en) um Hilfe, wenn erforderlich; ordentliches Betragen auch ohne Aufforderung durch den Lehrer, vorbereitet usw.

Mit Hilfe detaillierter Erläuterungen dieser Art konnte Ryans bei geschulten Beurteilern Reliabilitätskoeffizienten bis zu 0.8 und 0.9 erzielen, was eine relativ hohe intersubjektive Übereinstimmung andeutet. Diese ließ sich allerdings nur über kürzere Zeiträume bzw. durch permanentes Training der *Rater* aufrechterhalten. Diese Kautelen gelten praktisch für alle Skalentypen, soweit sie zum *Rating* benutzt werden, also auch für graphische Skalenformen (s. S. 35), Schätzskalen mit Punkthäufung und mit Zwangswahl. Die beiden letzten Skalentypen seien noch durch folgende Beispiele erklärt.

Beispiel einer *Schätzskala mit Punkthäufung* oder *Punktkumulierung (Punktsummenskala):*

1. Bevorzugt der Lehrer einzelne Schüler im Unterricht?	Ja	Nein	?
2. Erklärt der Lehrer den Unterrichtsstoff so, daß du ihn gut verstehen kannst?	Ja	Nein	?
3. Gibt der Lehrer zu viele Hausaufgaben auf?	Ja	Nein	?
4. Arbeitet die Klasse ruhig weiter, wenn der Lehrer unvorhergesehen das Klassenzimmer verläßt?	Ja	Nein	?
5. Gehst du gern zur Schule?	Ja	Nein	?

usw.

Auf Fragen dieser Art sind drei mögliche Reaktionen, die mit 0, 1 oder 2 Punkten bewertet werden, auszählbar. Diese Gewichtszahlen werden über sämtliche Aufgaben des Fragebogens (Beurteilungsschemas) aufsummiert. Die Punktsumme jedes einzelnen Beurteilers repräsentiert so das individuelle Schätzurteil in bezug auf dieses oder jenes Verhalten bei der Person X, z. B. Lehrerverhalten, Schülerverhalten usw. Diese Methode wird auch häufig zur Einstellungsmessung verwandt (vgl. Süllwold 1969, S. 497ff.).

Beispiel einer *Schätzskala mit Zwangswahl (forced-choice-technique):* Welche der Tätigkeiten würden Sie am liebsten wählen, wenn Sie sich für eine von jeweils vier Tätigkeiten entscheiden müßten? (Bitte ankreuzen!)

1. Sportberichte für eine Tageszeitung schreiben	—
Statistische Graphiken erstellen	—
Stromleitungen verlegen	—
Förster sein	—

2. Wissenschaftliche Untersuchungen durchführen —
Erwachsene unterrichten —
Stellenbewerber interviewen —
Diskutieren —
3. Golf spielen —
Fischen —
Reiten —
Klavier spielen —
usw.

Die Aufgabenbeispiele, die einem Interessentest entnommen sein könnten, verdeutlichen recht gut das Charakteristikum einer Schätzskala mit Zwangswahl: Aus einer Reihe vorgegebener Tätigkeiten (Interessen, Eigenschaften u.ä.) soll jeweils die bevorzugte Tätigkeit (Eigenschaft) benannt werden, was praktisch auf eine Zwangswahl zwischen mehreren Elementen – in unserem Beispiel zwischen vier Elementen – hinausläuft. Skalen dieser Art liefern „ipsative" Werte, d.h. Urteile über die intraindividuelle Merkmalsausprägung – im Gegensatz zu den „normativen" Werten, die den interindividuellen Vergleich (Aufschlüsse über zwischenmenschliche Unterschiede) gestatten. Viele Interessentests und Persönlichkeitsfragebogen basieren auf Schätzskalen mit Zwangswahl (vgl. noch Heller 1973, S. 154f.). Beispiele aus der Unterrichtsforschung finden sich wiederum bei Remmers (1963) bzw. Tent (1970, S. 874ff. u. 866ff.); siehe auch Walter (1973).

Exkurs über Beobachtungs- bzw. Beurteilungsfehler

Bevor wir uns den Beurteilungstechniken vom zweiten Typus zuwenden, sei kurz auf die wichtigsten Beobachtungs- bzw. Beurteilungs*fehler* eingegangen. Die Kenntnis dieser Fehlerquellen ist gerade für Lehrer und Erziehungswissenschaftler von Bedeutung, da diese wie kaum eine andere Berufsgruppe tagtäglich Eindrucks- und Schätzurteile mannigfacher Art abgeben (müssen). Nach Osgood u.a. erklärt sich die Gesamtvarianz solcher Urteile im wesentlichen aus drei Faktoren: *Evaluation* (allgemeiner Bewertungsfaktor hinsichtlich „sympathisch/angenehm vs. unsympathisch/unangenehm"), *Activity* (Aktivität, Fleiß usw. vs. Passivität, Trägheit usw.) und *Potency* (vorhandene vs. fehlende Fähigkeiten). Die wichtigsten subjektiven Fehlerquellen sind nun:

1. *Generosity error* (im Sinne Cronbachs) oder *Tendenz zur Gefälligkeitsnote.* Damit ist die Neigung mancher Lehrer oder Beurteiler gemeint, häufig zu gute Noten oder Beurteilungen zu erteilen (besonders gegenüber gut bekannten bzw. sympathischen Personen).

2. *Error of central tendency* oder *Tendenz zur Durchschnittsnote.* Diese Gefahr besteht meistens dann, wenn man sich seines Urteils nicht sicher ist, „ungerechte" Benotungen vermeiden will (dabei eher das Gegenteil erreicht) und/oder zu gutmütig, feige u.ä. ist, ein differenzierteres Urteil abzugeben.

3. *Contrast error* (nach Murray u.a.) oder *Tendenz zum konträren vs. analogen Merkmal.* Hiermit ist die Tendenz angesprochen, gegenteilige vs. ähnliche Eigenschaften (wie der Beurteiler sie an sich selbst zu erkennen glaubt) der beurteilten Person zuzuschreiben. Eine solche Einstellung der einen oder anderen Art verursacht bereits auf der ersten Stufe der Beobachtung (Wahrnehmung) gewisse Selektionsmechanismen, die dann das Urteil verfälschen können.

4. *Halo effect* (nach Thorndike) oder *Hofeffekt*. Er äußert sich darin, daß der Beurteiler aufgrund ungerechtfertigter Verallgemeinerungen, die auf Sympathie vs. Antipathie beruhen mögen, bevorzugt solche Eigenschaften des Beurteilten wahrnimmt versus unterdrückt, die mit dem Gesamteindruck übereinstimmen. Wer beispielsweise im Lateinunterricht sehr gute Leistungen zeigt(e), wird auch im Griechischunterricht (bzw. in Mathematik, Geschichte usw.) „wohlwollend" zensiert. Oder: Ein mittelmäßiger Schüler in den sog. Hauptfächern hat es oft schwerer als der Klassenprimus, in den sog. Nebenfächern – trotz vergleichbarer Leistungen – eine gute Note zu erzielen.
5. *Logical error* (nach Newcomb) oder *logischer Fehler*. Er funktioniert ähnlich wie der Hofeffekt. Einmal bzw. zu einer bestimmten Gelegenheit entdeckte Eigenschaften werden leicht zu „logischen" Eigenschaftsverbänden einer bestimmten Person oder Personengruppe erweitert. Solchen Beurteilungen liegt immer – mehr oder weniger explizit – eine bestimmte Persönlichkeitstheorie zugrunde: Von einem Naturwissenschaftler oder einem Mann erwartet man beispielsweise eher, daß er ein guter Techniker, Mathematiker, Autofahrer usw. ist, als von einem Geisteswissenschaftler oder einer Frau. Dem Künstler bzw. dem in einem musischen Bereich Tätigen werden leichter Gefühle, Phantasie, Produktivität u.ä. bescheinigt als etwa dem Juristen oder Steuerberater. Zur „impliziten Persönlichkeitstheorie" des Lehrers in bezug auf die Schülerbeurteilung liegt eine neuere Untersuchung von Hofer (1969) vor; siehe auch Erlemeier & Tismer (1973). Bei Verallgemeinerungen dieser Art kommt es dann leicht zu Fehlkombinationen. Ein Schüler, der sehr gute Leistungen im Lateinunterricht zeigt, *muß* eben *nicht* gleichzeitig ein ausgezeichneter Mathematiker sein. Nicht alle Mathematiker sind „kühle" Naturen, Verwaltungsfachleute amusische Menschen oder Lehrer notorische Besserwisser.

Vorstehende Liste enthält keine vollständige Fehlersammlung. Der Leser mag selbst – ergänzend durch eigene Erfahrungen oder vertiefende Lektüre (vgl. Hasemann 1964, Graumann 1966, v. Cranach & Frenz 1969, Donat 1970, Dieterich 1972, Fenner 1973 u.a.) – das eine oder andere hierzu beitragen. Das Wissen um die Gefahrenquellen ist sehr oft unentbehrliche Voraussetzung für eine objektive Beurteilung (und Beobachtung), es ist aber noch keine Garantie für entsprechendes Handeln. Dazu bedarf es fortlaufend kritischer Selbstprüfung des Beurteilers, die wenigstens von Zeit zu Zeit durch externe Kontrollinstanzen unterstützt werden muß.

Methoden der Content*analyse*

Der zweite Typ moderner Beurteilungstechniken findet seinen Vorläufer in der „Werkanalyse" der 20er und 30er Jahre. Hierbei kommt den *Verlaufsprodukten*, den Objektivationen des Verhaltens, entscheidende Bedeutung zu. Als Nachfolgerin der frühen Werkanalyse kann die heutige *Content*analyse (Inhaltsanalyse) mit ihren zahlreichen Methodenvarianten angesehen werden. Sie kommt bei den unterschiedlichsten Fragestellungen zur Anwendung, obgleich die Analyse des *sprachlichen* Gehaltes mitmenschlicher Kommunikation Hauptgegenstand dieses Verfahrensansatzes ist. So wird es auch verständlich, daß die *Content*analyse ihre größte Bedeutung in den jungen Wissenschaftsdisziplinen der Psycholinguistik bzw. Soziolinguistik erlangt hat. Das Erhebungsmaterial für eine *content*analyti-

sche Untersuchung reicht von persönlichen Verhaltensniederschlägen (Briefen, Tagebüchern, Autobiographien, Aufsätzen, Redeprotokollen usw.) bis hin zu amtlichen Dokumenten des Beurteilten (Urkunden, Auszeichnungen, Zeitungsberichten usw.). Zum Erhebungsmaterial zählen also praktisch alle sprachlichen Äußerungen – gelegentlich auch nonverbale Ausdruckserscheinungen –, die unter nichtstandardisierten oder (halb)standardisierten Bedingungen provoziert wurden. Von besonderem Interesse sind dabei *„expressive documents"* (vgl. Allport 1942, Gottschalk u.a. 1945, Berelson 1954, Hörmann 1967 u.a.).
Der *Content*analyse eignen drei wesentliche Merkmale. Das Verfahren ist *objektiv* (d.h. intersubjektiv vergleichbar und somit kontrollierbar), *systematisch* (ein verbindliches Kategoriensystem ermöglicht den Konsens) und *quantifizierbar* (die Möglichkeit der quantitativen Verarbeitung qualitativer Variablen auf dem Wege der Dimensionierung ist hier besonders wichtig und wertvoll). Ferner ist für den *content*analytischen Ansatz charakteristisch, daß das Vorgehen hier vollends auf eine einzige Phase beschränkt wird: die der eigentlichen Beurteilung. Die Phasen der Beobachtung und Beschreibung – durch den Beurteiler – fehlen bei der *Content*analyse.
Nach einer Übersicht von Graumann (1965) lassen sich an die *Content*analyse (CA) folgende Untersuchungs*funktionen* knüpfen (siehe auch Herrmann & Stäcker 1969):

1. Aus den sprachlichen Gehalten sind bestimmte *Mitteilungstendenzen* eruierbar. Die Analyse eines bestimmten Themas über Jahre hinweg kann so gewisse Veränderungen anzeigen, wie entsprechende Untersuchungen im historischen Raum veranschaulichen. Beispielsweise konnte Lasswell (1949) auf diese Weise einen „Wandel von weltrevolutionären zu nationalistischen Tendenzen" in der UdSSR zwischen 1918 und 1943 feststellen. Einschlägige erziehungswissenschaftliche Untersuchungsansätze wären etwa Analysen von Buchtiteln oder Zeitschriftenpublikationen pädagogischer Provenienz innerhalb einer bestimmten Epoche, Veröffentlichungsthemen eines bestimmten Erziehungswissenschaftlers, Parolen von Interessenverbänden usw. Auf diese Weise ließen sich Beschäftigungsschwerpunkte, Einstellungstrends u.ä. einzelner Personen oder ganzer Gruppen statistisch erfassen.

2. Ein weiteres Feld der CA ist die *Enthüllung internationaler bzw. interkultureller Differenzen*. So wurden etwa die beliebtesten Schauspiele in Deutschland und den USA vergleichend untersucht oder die HJ- und Boy-Scout-Literatur *content*-analysiert. Im pädagogischen Kontext bieten sich Schulbuchvergleiche, Analysen der Freizeitlektüre, des Filmbesuchs usw. von Kindern, Jugendlichen oder Erwachsenen an. Auf diese Weise lassen sich wissenschaftliche Erkenntnisse über die Wirkung unterschiedlicher Kulturmilieus, Schulsysteme, sozialer Determinanten wie Sprachvorbilder, Erziehungsmodi u.ä. gewinnen (vgl. z.B. Markefka & Nauck 1972).

3. Die CA erlaubt auch den wissenschaftlich begründeten *Vergleich verschiedener Kommunikationsmedien* (Presse, Funk, Film, Fernsehen), besonders hinsichtlich des *Kommunikationsniveaus* (Vorurteile, Parteilichkeit, Begünstigung usw.).

4. Via CA lassen sich genauere *Aufschlüsse über die eigenen Methoden* wie Interview, Fragebogen u.ä. gewinnen. Die CA ist somit ein Instrument, das für die verschiedensten Zwecke sprachlicher *Inhalts*analysen verwendbar ist. Andere Modi

der CA bezielen demgegenüber mehr die *formale* Struktur diverser Verhaltensniederschläge, wie die folgenden Punkte verdeutlichen sollen.
5. White (1949) definierte sog. *emphasis-units*. Diese spielen bei der Heldenverehrung, bei Persönlichkeitskulten u.ä. eine Rolle. So werden „Helden" durch häufige Erwähnung, durch starke Attribute, multiplizierende Adverbien usw. gekennzeichnet. White verglich mit Hilfe dieser Methode die Vorkriegsreden von Hitler und Roosevelt unter dem Gesichtspunkt der „Stärke" und fand, daß diese bei Hitler mit 35%, bei Roosevelt dagegen nur mit 15% betont war.
6. Die CA bietet ferner die Möglichkeit, verschiedene Kommunikationsmittel auf ihre *Verständlichkeit* hin zu untersuchen. Auf diese Weise erforschte u.a. Flesch (1951) die Leserlichkeit von Schulbüchern. Dabei stellte er besonders zwei Komponenten heraus: a) Leseleichtigkeit *(reading ease)* mit „durchschnittlicher Satzlänge" und „Silbenzahl der Wörter" als Kriterien; b) Prozentsatz persönlicher Wörter und Sätze *(human interest)*, z.B. Häufigkeit der persönlichen Fürwörter in der 1. und 2. Person.
7. Es wurde schon betont, daß die CA ein wichtiges Verfahren der modernen *Linguistik* darstellt. Die Methode ist hier gleichermaßen als Operationsansatz zur Aufdeckung *stilistischer* Eigentümlichkeiten (formale CA) sowie als Analysemodell zur Erhellung *inhaltlicher* Aspekte (Gehaltsanalyse i.e.S.) interessant. Häufig werden vom literarischen Verhalten Rückschlüsse auf den Urheber, vom *Wie* des sprachlichen Verhaltens auf den Sprecher oder *Wer* vollzogen.
8. In enger Beziehung dazu steht der Beitrag der CA zum Erkennen von *Intentionen* und überhaupt von *persönlichen Eigenschaften* des Kommunikators. Allerdings sind die Validierungserfolge dieses Vorgehens im allgemeinen schlechter als beim Einsatz bewährter diagnostischer Verfahren. Historisch interessant ist hier eine Untersuchung der Reden zu Stalins Geburtstag, um die Einstellung der beteiligten Sowjetführer gegenüber Stalin zu erkunden. Doch müssen auch in diesem Falle erhebliche Zweifel bezüglich der Aussagegültigkeit angemeldet werden; die untersuchten Reden dürften nicht ohne weiteres mit der (wahren) Einstellung ihrer Autoren identisch (gewesen) sein. Die Gefahr irriger Schlußfolgerungen war bzw. ist sehr leicht gegeben.
9. Die CA interessiert als Methode besonders auch im Hinblick auf die *persönlichkeitspsychologische* Forschung, etwa wenn bestimmte psychische Zustände von Einzelpersonen oder Gruppen erschlossen werden sollen. Analysen von Interviews, projektiven Testaussagen, Briefen, Tagebüchern u.ä. sind hier einschlägig. Allport hat die einzelnen Techniken dazu in seinem bereits zitierten Buch ausführlich dargestellt. Die Verwendungsmöglichkeiten der CA in diesem Sinne seien durch ein letztes Beispiel, das wiederum aus dem Zweiten Weltkrieg (durch den die *content*-analytische Forschung kräftige Anstöße erhielt) stammt, kurz illustriert: So versuchten amerikanische Psychologen die Moral der deutschen Heimatfront zu eruieren, indem sie erbeutete Post der deutschen Zivilbevölkerung *content*analysierten. Die Ergebnisse erbrachten seinerzeit eine merkliche Abstumpfung der Bevölkerung gegen nächtliche Luftangriffe. Der heutige Leser mag sich selbst gewisse Parallelen zum Vietnamkrieg unserer Tage ausmalen.

Allen geschilderten Vorgehensweisen *content*analytischer Forschung ist eines gemeinsam: Vom sprachlichen Gehalt dieses oder jenes Verhaltensproduktes bzw. dem *Wie*-Verhalten des Urhebers wird auf den Sender bzw. die Verfasserpersönlichkeit des betr. Schriftstücks geschlossen. Die Treffsicherheit eines solchen

Schlusses ist aber nur dann gewährleistet, wenn das Geschriebene oder Gesprochene der jeweiligen Einstellung, Motivation, Interessenlage usw. entspricht. In der Zukunft werden deshalb weitere Entwicklungsschritte zur Absicherung der CA, d.h. zur Verbesserung ihrer Validität und Reliabilität, notwendig.

Im folgenden sei die technische *Durchführung* und *Auswertung* der CA noch kurz angesprochen, im übrigen aber auf die oben erwähnte Literatur verwiesen. Zunächst wird das Material erhoben. Dem Kodierungsvorgang *(coding)*, d.h. der Transformation identifizierter Phänomene in ein wissenschaftliches System – analog dem Kategorisieren bei der wissenschaftlichen Beobachtungsmethode (etwa der Verhaltensbeobachtung in gebundener Form) –, fällt hier die entscheidende Rolle zu. Die dazu aufgestellten Übersichten oder Tabellen müssen ganz bestimmten Kriterien genügen. Sie sollen einer quantitativen Weiterverarbeitung zugänglich sein, sie müssen in Isomorphie zu den Zielfragen bzw. Aufgabenproblemen stehen, ein bestimmtes Item darf jeweils nur in *eine* Kategorie fallen, Einheiten der Protokollierung, des Kontextes sowie Zähleinheiten *(units)* sind genau festzulegen u.dgl.m. Je nach Vorgehensweise lassen sich so verschiedene *Techniken* der CA benennen:

1. *Frequenzanalyse.* Damit ist die klassische (zählende) CA gemeint. Die Häufigkeiten bestimmter Units lassen sich auszählen. So wird die Häufigkeit einer Vokabel, in einem Redetext z.B. „Friede“, „Macht“, „Eigentum“, „Bildung“, „Umwelt“, „sozial“, „ich“ usw., festgestellt und in Relevanz zum betr. Urheber (Redner, Autor) gesetzt. Mit dieser Technik lassen sich sehr gut Aufsätze, Vorträge u.ä. analysieren.
2. *Analyse wertender Einzelaussagen.* Diese von Osgood entwickelte Methode der CA untersucht vor allem Einstellungen, Gerichtetheiten oder Haltungen gegenüber sozialen und/oder politischen Problemen. Hierbei wird zunächst festgestellt, zu welchen Themen jemand Stellung genommen hat. Danach zählt man aus, wie oft Stellung genommen wurde und in welcher Form (wie) dies geschah; am häufigsten wird dafür folgendes Skalenmodell verwendet: befürwortend (+), indifferent (0), ablehnend (–). Die Anzahl der positiven und negativen Stellungnahmen im Verhältnis zu den indifferenten Meinungsäußerungen ergibt schließlich einen Zahlenindex, der statistisch weiterverarbeitet werden kann.
3. *Kontingenzanalyse.* Hier geht es um die Erhellung der sprachlichen bzw. der dahinter liegenden Assoziationsstruktur. Dabei interessiert vor allem die Frage, welcher Inhalt sich mit welchem anderen (häufig) verbindet versus welche Inhalte sich nie (selten) verbinden. Beispiele für eine solche thematische *Assoziation* liefert etwa Goebbels' Tagebuch, in dem im Zusammenhang mit den besetzten Ostgebieten oft von Nahrung und Rohstoffen die Rede ist. Daraus wurde geschlossen, daß die genannten Gebiete besondere Bedeutung im Hinblick auf die wirtschaftliche Versorgung Deutschlands erlangten. Ein Gegenbeispiel für *Dissoziation* wäre das Begriffspaar *„mother – sex“;* diese Begriffe verbinden sich sehr selten.
4. *Close-procedure analysis* sensu Taylor (1953, 1955) u.a. Dieser Verfahrensansatz basiert auf der Tatsache, daß Textlücken durch den Kontext inhaltlich determiniert werden. Den Untersuchungspersonen werden deshalb Lückentexte zum Ausfüllen gegeben, z.B.: „Hunde, die bellen, ... nicht“; „Ich war gespannt auf die ...“; „Im Dunkeln ging ich durch ..., als ich ... hörte“ usw. Man kann also mit Lücken verschiedener Art und Häufigkeit arbeiten. Für die Auswertung solcher

Lückentexte entwickelte Taylor ein Maß der *Verstehbarkeit (readability)*, mit dessen Hilfe interindividuelle Unterschiede erfaßt werden können. Diese Methode eignet sich auch für Schulbuchanalysen, Analysen der Unterrichtssprache u.ä.
5. Modi der *formalen* CA kamen in den letzten Jahren vor allem in der Linguistik und Begabungsforschung zur Anwendung. So wurden Wortschatzuntersuchungen mit Hilfe des *Diversifikationsquotienten* oder der Type-Token-Ratio (TTR), wobei das Verhältnis aus der Anzahl verschiedener Einzelwörter zur Gesamtzahl aller gebrauchten Wörter einer Sprachprobe ausgedrückt wird, vorgenommen. Auf diese Weise konnten unterschiedliche Formen des Sprachgebrauchs – Sprachkode der Mittelschicht (elaborierter Kode) vs. Unterschicht (restringierter Kode) – abgegrenzt werden. Ferner hat man mit Hilfe des *Subordinationsindexes* (Quotient aus abhängigen Nebensätzen und Anzahl finiter Verben) syntaktische Qualitäten des Sprachverhaltens erfaßt und gewisse Affinitäten zwischen sprachlicher Komplexität (Satzbauplänen) und Denkverhalten (Intelligenzstruktur) festgestellt; siehe ausführlicher Herrmann & Stäcker (1969, S. 418ff.) sowie Heller (1973, S. 38ff.).

Q-Techniken

Im Kontext der Beurteilungstechniken sind schließlich – als dritter Typ unserer Systematik – die sog. *Fragebogen- oder Q-Techniken* zu erwähnen. Diese finden besonders in der Persönlichkeits- bzw. diagnostischen und klinischen Psychologie Verwendung und gehen ursprünglich auf Stephenson (1953) zurück. Heute gibt es unter der Sammelbezeichnung *Q-Technik(en)* eine Reihe unterschiedlicher Verfahren. Die zwei wichtigsten Methodengruppen sind einmal jene Verfahren der Faktorenanalyse zum Vergleich *inter*personaler Eigenschaften (= *Q-technique*) und zum andern Verfahren zur Ermittlung *intra*individueller Erlebnis- und Verhaltensweisen (= *Q-sorting*). Eine ausführliche Beschreibung dieser Methodenansätze findet sich bei Mowrer (1953); Sheldon & Sorenson (1960) sowie Bennett (1964) berichten über pädagogische Anwendungsbeispiele.
Den Q-Techniken liegen *Rating*skalen zugrunde, die meistens drei Antwortkategorien enthalten: Zustimmend (+); Ablehnend (–); Unentschieden (0). Fragebogen dieser Art sind z.B. der MMQ und MPI von Eysenck, der MMPI-Saarbrücken oder der PIT von Mittenecker & Toman (vgl. Heller 1973, S. 144ff.). Hierbei werden den Probanden Fragen vorgelegt, zu denen sie Stellung nehmen müssen, etwa in der Form von „stimmt“, „stimmt nicht“ oder „weder noch“ bzw. „weiß nicht“. Unser früheres Beispiel einer Punktsummenskala (s. S. 42) wäre hier ebenfalls einschlägig. Es gibt auch 5- oder 7stufige – seltener 4stufige, 6stufige usw. – *Rating*skalen für die Beantwortung solcher Fragebogen, die von „stimmt in jedem Falle“ bis „trifft nie zu“ mehr oder weniger ein Kontinuum darstellen. Meistens sind jedoch nur drei Bereiche (+ vs. 0 vs. –) angegeben. Eine der zahlreichen Varianten der Q-sort-Methode, die bislang in der Literatur kaum erwähnt wird, ist das von A. Wagstaff anläßlich eines Vortrags in Heidelberg 1965 vorgestellte *Q-sorting experiential data*, ein Verfahren, das von *„naturalistic data“* ausgeht: Zunächst unterhält sich der Therapeut mit dem Probanden, wobei das Gespräch auf Tonband festgehalten wird. Das so gewonnene Material dient dann der Formulierung neuer Fragen an den Probanden, d.h., einzelne Aussagen des Probanden werden ihm in seiner eigenen Formulierung (beispielsweise auf Kärtchen geschrieben und in der folgenden Sitzung dargeboten) zur Stellungnahme vorgelegt. Der Proband wird also quasi mit seinen eigenen Aussagen getestet. Auf diese

Weise soll festgestellt werden, welche Relevanz die Aussagen des Probanden für ihn selbst bzw. im Kontext seiner Problematik haben. Auf andere Methodenvarianten kann hier ebensowenig eingegangen wie eine ausführliche Darstellung der angesprochenen Methoden geboten werden; siehe ergänzend noch Remmers 1963 bzw. Tent 1970, S. 946ff., G.G.Stern 1963 bzw. Parey & Ingenkamp 1970 u.a.

Zu den *Gütekriterien* (Objektivität, Reliabilität, Validität) solcher Einschätzverfahren werden in der Literatur recht unterschiedliche Ergebnisse berichtet. Die Verläßlichkeit (Reliabilität) und Gültigkeit (Validität) entsprechender Schätzurteile dürften einerseits von der Skalenqualität und deren empirischer Fundierung und andererseits von der jeweiligen Anwendungspopulation bzw. dem jeweiligen Untersuchungsziel abhängen. Die oben genannten Fragebogenbeispiele sind gut standardisiert und hinsichtlich der (Test-)Gütekriterien durchaus mit psychometrischen Verfahren (Intelligenz- oder Schulleistungstests) vergleichbar. Andere Verfahren genügen diesen Ansprüchen nicht oder nur sehr unvollkommen, so daß bei ihrer Verwendung größte Vorsicht geboten ist. Mit dem Prädikat „wissenschaftlich" können nur solche Operationsansätze (*Observation, Rating, Q-technique* usw.) ausgezeichnet werden, die den genannten Gütekriterien des Messens hinreichend genügen, d.h., deren Ergebnisse erwiesenermaßen objektiv (überprüfbar), reliabel (zuverlässig: konsistent vs. zeitstabil) und valide (aussagegültig) sind. Diese Forderung betrifft intermethodale Grundsätze, denen somit *alle* „wissenschaftlichen" Aussagen qua definitione principii Rechnung tragen müssen. Im besonderen Maße gilt dieser Anspruch gegenüber dem zweiten Methodenfundament empirisch-operationaler Forschung, dem Experiment.

2.2. Experiment (Versuch)

2.2.1. Begriff und Kriterien

Phänographie *(Observation)* und Experiment sind zwei sich ergänzende Methodenansätze, wobei die wissenschaftliche Beobachtungsmethode der umfänglichere und letzten Endes das Experiment fundierende Ansatz ist. Genaugenommen erweist sich nämlich das Experiment als Sonderform der Beobachtung, insofern hier eine Beobachtung unter besonderen Bedingungen stattfindet, weshalb man gelegentlich auch von „experimenteller Beobachtung" spricht (vgl. Traxel, Graumann u.a.). Sowohl die *Observational Techniques* als auch das Experiment zielen auf den *Verlauf* eines bestimmten Verhaltens oder Erlebens ab. Trotzdem besteht ein wichtiger Unterschied. Das zu beobachtende Verhalten bleibt bei den *Observational Techniques* relativ unangetastet; hingegen bewirkt das Experiment bereits seit Wilhelm Wundt eine „absichtliche, planmäßige Auslösung eines Vorgangs zum Zweck der Beobachtung" (Traxel 1964, S. 90). Zur Beobachtung kommt also beim Experiment die *willkürliche* Einwirkung in den Verlauf als Spezifikum hinzu.
In den bekannten Definitionen des Experiments tauchen die Kriterien „Herstellung", „systematische Variation" oder „Manipulation" und „Kontrolle" fast immer auf, wenngleich unterschiedliche Akzentuierungen eine einheitliche Begriffsfestlegung sehr erschweren. Ähnlich wie Traxel (a.a.O.) stellt Metzger (1952,

S. 143f.) den Herstellungsaspekt in den Mittelpunkt seiner Definition: „*Experiment* oder (wissenschaftlicher) *Versuch* heißt die Bemühung, Bedingungen *herzustellen* nur zu dem Zweck, ihren Einfluß auf einen fraglichen Gegenstand oder Sachverhalt zu beobachten und diesen dadurch besser kennenzulernen, gegebenenfalls schon bestehende Vermutungen über seine Natur (Deutungen seines Verhaltens) und ihre *faktische* Stichhaltigkeit zu prüfen.“ Bei Edwards (1954, 1960), Festinger (1953), Zimny (1961) u.a. liegt der Schwerpunkt auf den Kriterien der Manipulation und Kontrollierbarkeit der Versuchsbedingungen, die jedoch logisch dem erstgenannten Kriterium (vgl. die Metzgersche Definition) zuzuordnen sind. Im Anschluß an die Ausführungen Wundts (1874, 1896) wollen wir anhand eines 5-Punkte-Katalogs die wichtigsten *Kriterien* des Experiments kurz explizieren:

1. Die *Beobachtung* im engeren Sinne ist zweifellos ein essentielles Merkmal aller experimentellen Untersuchungsverfahren. Hiermit ist die Blickwendung auf bestimmte Situationsaspekte gemeint, also eine zentrierte und selektierende Wahrnehmung im früher aufgewiesenen Sinnzusammenhang (vgl. Kap. 2.1.1).
2. Die *willkürliche Einwirkung* als zweites Kriterium des Experiments zielt sowohl auf die Entstehung als auch auf den Verlauf. Der Experimentator oder Versuchsleiter (Vl) bewirkt, daß sich die Versuchsperson (Vp) bzw. Versuchspersonen (Vpn) an einem bestimmten Ort zu einem festgelegten Zeitpunkt unter genau definierten Bedingungen so und so verhält bzw. verhalten. Im Idealfalle ist das zu beobachtende bzw. beobachtete Verhalten so festgelegt, daß es ausschließlich und vollständig durch die Versuchsbedingungen (Instruktion des Vl) bestimmt wird. Dabei werden die Bedingungen, unter denen der Versuch steht bzw. abläuft, immer wieder umgestellt, um eine optimale Beobachtung des Verhaltens bzw. eindeutige Dependenzanalysen (Interpretation des Zusammenhangs zwischen unabhängiger und abhängiger Variablen) zu erzielen. In engem Zusammenhang damit muß der Aspekt der Variierbarkeit gesehen werden (s. Pkt. 3).
3. Die *Variierbarkeit* oder *systematische, isolierende Variation* beinhaltet ein zweifaches Problem: die Isolierung und die Manipulierbarkeit. Die *Isolierung*, die auf dem Wege der Dimensionierung (vgl. Kap. 2.2.6.3 unten) vollzogen wird, dient der Kontrolle, d.h. der Erhellung der jeweiligen Bedingungsstruktur beobachteter Phänomene bzw. Phänomenkomplexe. Im Experiment geht es immer um eine Isolierung von Aspekten einer bestimmten Situation, die hinsichtlich ihrer relationalen Funktion des Mehr oder Weniger betrachtet, also dimensioniert werden können. Unter Situation ist hier im Sinne Graumanns der Person-Welt-Bezug phänomenologisch zu verstehen. Daraus folgt, daß in jeder experimentellen Situation drei Glieder berücksichtigt werden müssen: Versuchsperson, Umwelt und momentanes Verhältnis zwischen Person und Welt. Nach neueren Erkenntnissen (z.B. Rosenthal 1966, 1967 u.a.) wäre quasi als viertes Glied der Versuchsleiter als konstitutive Variable des Experiments und damit als Determinante der Versuchsergebnisse mit zu berücksichtigen. Zum Variablenbegriff siehe nachfolgenden Exkurs. Die *Manipulierbarkeit* betrifft die Willkürlichkeit und die Variierbarkeit gleichermaßen. Ihre Grenzen liegen einmal in der Natur selbst (beispielsweise ist eine künstliche Variation der Sonnenbahn unmöglich), zum anderen bestehen ethische Schranken, worauf besonders Metzger (a.a.O.) hingewiesen hat. Menschen in einer existentiellen oder sozialen Krise in eine experimentelle Situation hineinzumanövrieren, etwa zur Untersuchung des Zusammenhangs von suizidalem

Verhalten und Persönlichkeitsstruktur resp. Familienmilieu, verbietet sich ebenso wie das Experimentieren aus bloßer, privater Neugier. Grundsätzlich gilt hier für den Experimentator, daß die menschliche Würde, Empfindsamkeit, Ehre, Gesundheit usw. unantastbar bleiben müssen. Die Versuchspersonen dürfen also keinen körperlichen, seelischen, charakterlichen oder sonstigen Schäden und Gefahren ausgesetzt werden.
4. Die *Wiederholbarkeit* des Experiments impliziert die Notwendigkeit, die Versuchsbedingungen (d.h. die Methode) genau mitzuteilen, so daß jeder Wissenschaftler, gleich welcher Schule er angehört, durch Wiederholung des Versuchs – prinzipiell – zu denselben Ergebnissen gelangen kann. Auf diese Weise werden die experimentellen Ergebnisse überprüfbar (s. Pkt. 5).
5. Die *Kontrollierbarkeit* ist eine unerläßliche Forderung in der Wissenschaft und somit auch ein Kriterium des Experiments. Variierbarkeit und Wiederholbarkeit sind praktisch Folgen der Willkürlichkeit und Steuerbarkeit (im Sinne Wundts). Wiederholung ist aber nur dann möglich, wenn die Öffentlichkeit des Experiments – und damit seine Kontrolle – gewährleistet ist. Für den Operationisten ist dieses Postulat eines der wichtigsten Wissenschaftskriterien überhaupt (s. S. 26f.). Dem Kriterium der Kontrolle kommt im Experiment allerdings exklusive Bedeutung zu (vgl. Kap. 2.2.4.3).

Zusammenfassend kann demnach festgehalten werden: Unverzichtbare Bestimmungsstücke des Experiments sind erstens die objektive *Beobachtung*, zweitens die strenge *Kontrolle* der Versuchssituation, drittens die *willkürliche Einwirkung* im Sinne einer systematischen *Isolierung* versus *Manipulation* der Versuchsbedingungen oder (unabhängigen) Variablen. „Wichtig an dieser Begriffsbestimmung ... ist die ausdrückliche Erwähnung der Variation eines Faktors, die der Vl vornimmt. Dadurch werden *zwei* veränderliche Größen eingeführt: 1. die sog. unabhängige Variable (UV), die vom Vl manipuliert wird; 2. die sog. abhängige Variable (AV), deren Kovariation mit der UV als Folge der vom Vl vorgenommenen Manipulation angesehen wird. Ziel eines so definierten Experiments ist es, herauszufinden, ob als Folge der Variierung der experimentellen Bedingungen (UV) die AV kovariiert" (Bredenkamp 1969, S. 332).

2.2.2. Exkurs über den Variablenbegriff

Wir erwähnten schon die Unterscheidung in *unabhängige (independent)* Variable (UV) und *abhängige (dependent)* Variable (AV). Wenn beispielsweise die Bedingungen, d.h. die psychologischen, pädagogischen oder sozio-kulturellen Voraussetzungen, der Schulleistung experimentell aufgeklärt werden sollen, dann wäre die Schulleistung selbst die AV, während Intelligenz, Geschlecht, Alter, Lernmotivation, Bildungsinteressen, sozio-ökonomischer Status des Elternhauses, pädagogisch-didaktisches Können des Lehrers usw. hier als UV fungierten. Im klassischen Experiment wird immer nur eine UV systematisch variiert, um so – bei gleichzeitiger Konstanthaltung aller übrigen Faktoren (= Störvariablen) – den Einfluß auf die beobachtete AV kontrollieren zu können. Die Variation in unserem fiktiven Beispiel wäre etwa durch Berücksichtigung der interindividuellen Differenzen bezüglich der genannten Merkmalsausprägungen (UV), die natürlich kon-

trolliert werden müßten (vgl. Kap. 2.2.4.3), möglich bzw. würde durch entsprechende Stichprobenorganisationen gewährleistet werden können. Sofern nur eine einzige UV untersucht – und variiert – wird, z.B. die Abhängigkeit der Schulleistung (AV) von der Intelligenz (UV) vs. die Abhängigkeit der Schulleistung (AV) vom Geschlecht (UV), erfordert dies einen sog. *univariaten* Versuchsplan. Werden mehrere UV (z.B. Intelligenz plus Geschlecht vs. Intelligenz plus Geschlecht plus Alter) eingeführt, dann sprechen wir von *multivariaten* Dependenzanalysen (s. dazu ausführlicher Kap. 2.2.4 unten). Unter relationalen Gesichtspunkten ist schließlich noch – neben der Unterscheidung von UV und AV – eine dritte Variablenkategorie von Bedeutung: die sogenannte intervenierende Variable (IV). „Als *intervenierende Variablen (intervening variables)* bezeichnet man ... solche, die sich mit ins Spiel setzen, jedoch hinsichtlich des Umfangs ihres Einflusses (noch) nicht kontrolliert werden konnten" (Drever & Fröhlich 1968, S. 244). Siehe ausführlicher unten S. 52f.
Die einzelnen Variablen (UV), die isoliert und manipuliert werden können, lassen sich in vier Klassen einteilen: a) die Klasse der *Reizvariablen* oder *Stimulus-Variation*. Im einzelnen fallen hierunter die Umweltvariablen *(environment-variables)*, die Aufgabe-Variation *(test-variables)* und die Instruktionsvariable, welche sowohl als *environment-* wie auch als *test-variable* definierbar ist. Die Instruktionsvariable ist besonders im pädagogisch-psychologischen Experiment von Bedeutung. Will man beispielsweise „Lernen" (AV) als Funktion der persönlichen Relevanz (UV) untersuchen, so ist eine Variation der Instruktionsvariable folgendermaßen denkbar: Bei der Aufforderung zur Aufgabenlösung würde man der Gruppe 1 nur sagen, daß es sich um Aufgaben für ein („irgendein") Experiment handelt; Gruppe 2 würde man aber vielleicht darüber informieren, daß die gestellten Aufgaben der Erfassung intellektueller Fähigkeiten dienen. Auf diese Weise werden unterschiedliche Einstellungen, Motivationen u.ä. – trotz gleicher Aufgabenstellung in beiden Gruppen – seitens der Versuchspersonen bewirkt, d.h. via Instruktionsmodifikation (mittelbar) unterschiedliche Versuchsergebnisse provoziert; b) die Klasse der *Reaktionsvariablen*, wobei Zeitdimensionierung und Fehlervariablen vs. Pluspunkte die Hauptrolle spielen; c) die Klasse der *Subjekt-* oder *Organismus-Variable*. Damit ist die Variable zwischen Reiz (*S*timulus) und Reaktion (*R*esponse) gemeint, die *O*rganismusvariable. Heute wird ja das ursprüngliche (behavioristische) S-R-Modell zum S-O-R-Modell erweitert gedacht. Einzelne O-Variablen sind z.B. Intelligenz, Geschlecht, Angstlichkeit bzw. alle Persönlichkeitsvariablen; d) die Klasse der *Versuchsleiter-Variablen*. Hier ist vor allem der *„experimenter bias"* angesprochen (vgl. Rosenthal 1966, 1967 u.a.). Jeder Vl geht mit bestimmten Vorstellungen, Erwartungen, Hypothesen usw. an den Versuch heran, was sich unter Umständen auf die Untersuchungsergebnisse bzw. die AV auswirkt. Durch Betonung bestimmter Wörter in der Instruktion, ein freundliches vs. mürrisches Gesicht, durch Ablese- oder Rechenfehler u.ä. kommt ein solcher *bias* zum Tragen. Dem kann man auf verschiedene Weise begegnen, wie noch zu zeigen sein wird (vgl. Kap. 2.2.4.3). Ferner wird durch die Aufstellung der sog. Nullhypothese und das Bemühen, diese zu widerlegen – nicht die eigene Hypothese zu verifizieren –, der *experimenter bias* in etwa neutralisiert.
Schließlich wäre noch auf die Unterscheidung zwischen „Intervenierenden Variablen" (IV) und „Hypothetischem Konstrukt" (HK) hinzuweisen. Sehr oft werden beide Begriffe mehr oder weniger unkritisch synonym verwendet. Während bei-

spielsweise Townsend (1953) alle O-Variablen als IV ansieht, sprechen McCorquodale & Meehl (1948) „von intervenierenden Variablen nur, wenn sie abstrakte empirische Beziehungen, hingegen von hypothetischen Konstrukten, wenn sie nichtbeobachtbare, erschlossene Größen und Prozesse bezeichnen wollen“ (zit. nach Selg 1966, S. 40). Termini wie „Gen“, „Intelligenz“, „Motivation“ u.ä. sind demnach HK, während funktionale Beziehungen i.e.S., z.B. Intelligenzquotient, Lernleistung, Fehlerprozent u.ä. oder auch „Hunger“, „Trieb“ usw., als operationale Begriffe IV darstellen. Diese begriffliche Differenzierung wird freilich in der Literatur – teils aus Unkenntnis, teils wegen Abgrenzungsproblemen – nicht immer durchgehalten, was verschiedentlich schon moniert wurde (Graumann, Selg u.a.). Das folgende Zitat, entnommen Selg (1966, S. 42), soll gleichermaßen zum besseren Verständnis der Schwierigkeiten und zur Begriffsfestigung beitragen.

In vielen experimentellen und nichtexperimentellen „Arbeiten der letzten Jahre werden Konstrukte wie ‚Intelligenz‘, ‚Motivation‘ usw. als intervenierende Variablen geführt – entweder weil die Begriffsbildung von McCorquodale und Meehl vernachlässigt wird oder weil man wie Townsend betonen will, die von den Konstrukten bezeichneten Entitäten seien *zwischen* einem ‚Agens‘ und einem ‚Verhalten oder Zustand‘ (zwischen ‚Reiz‘ und ‚Reaktion‘) wirksam (vgl. Graumann 1965).
Es ist ferner auch außerordentlich schwer, die Unterteilung immer streng beizubehalten, selbst wenn man sie für notwendig hält. Ehe man weiß, ob jemand einen Begriff nur als intervenierende Variable oder als hypothetisches Konstrukt gebraucht, muß ein weites Umfeld mitbekannt sein. In Hulls psychologischem System war der Antrieb *(drive)* ursprünglich eine intervenierende Variable, doch dürfte *drive* später weitergehende Bedeutungen *(surplus meanings)*, wie sie für hypothetische Konstrukte charakteristisch sind, angenommen haben, zumal in der Vorstellung der Hull zitierenden Autoren, die in anderen psychologischen Systemen beheimatet sind. Wer gewohnt ist, Triebe als große, geheimnisvolle, das Leben in Gang haltende Kräfte zu sehen, wird Hulls Terminologie leicht falsch verstehen.
Hypothetische Konstrukte sind nicht direkt manipulierbar; sie wären sonst lediglich als weitere unabhängige Variablen anzusehen und einzusetzen. Das schließt aber noch nicht aus, daß sich die Forschungsarbeit der Entitäten, die mit hypothetischen Konstrukten belegt sind, annehmen kann. Voraussetzungen solcher Approximationen sind jedoch *operationale* Definitionen.“ Siehe S. 20f. oben.

Der Begriff „Variable“ ist für die experimentelle Forschung konstitutiv. Beim Gebrauch dieses Begriffes sind freilich zwei Bedeutungsgehalte auseinanderzuhalten: der *engere* Begriff „Variable“ als manipulierbare, dimensionierbare Größe (UV vs. AV) und der *weitere* Begriff „Variable“, worunter dann alle definierten Kategorien und Größen – selbst dann, wenn sie nicht dimensionierbar eingeführt werden, also keine abgestufte Manipulation im strengen Sinne möglich ist (z.B. Geschlechts-„Variable“) – in einem Experiment verstanden werden. Nach Möglichkeit sollte man den Variablenbegriff nur im ersten (engeren) Sinnverständnis, d.h. im Sinne funktionaler Abhängigkeit, gebrauchen. Daneben sind die einzelnen – oben explizierten – Variablenarten zu beachten.

2.2.3. Formen des Experiments

Metzger (1952) unterscheidet zwischen *Erkundungsexperiment* und *Entscheidungsexperiment*. Diese Einteilung berücksichtigt das Ziel des Experiments. Während das Erkundungsexperiment die Seinsfrage (*Was* ist da?) interessiert, liegt dem Entscheidungsexperiment die Soseinsfrage (*Wie* ist das Phänomen (AV) beschaffen? In welchen Relationen des Soseins steht es?) zugrunde. Mit anderen Worten: Via Erkundungsexperiment soll Neues zutage gefördert werden. Via Entscheidungsexperiment soll eine Hypothese *verifiziert* (bestätigt) vs. *falsifiziert* (widerlegt) werden. Hierzu wird gewöhnlich die sog. Nullhypothese angesetzt, um widerlegt zu werden, d. h., es wird statistisch festgestellt, daß man sich bei der Aufstellung der Nullhypothese (s. noch Kap. 2.2.4.4) geirrt hat. Das ist besser, als wenn man etwas positiv behauptet, was man eventuell nachher zurücknehmen muß. Neben erkenntnistheoretischen Gründen sprechen auch Instruktions- bzw. Versuchsleitervariable als Einflußgrößen (vgl. *experimenter bias* resp. *experimenter effect*) für das indirekte Verfahren der Hypothesenprüfung, also die Verwendung der Nullhypothese.
Anhand der Literatur läßt sich feststellen, daß im anglo-amerikanischen Raum das Entscheidungsexperiment, auf kontinentaleuropäischem Boden das Erkundungsexperiment in der Vergangenheit bevorzugt worden ist. Dies hat erkenntnistheoretische Wurzeln. Während man im angelsächsischen Lager zumeist davon ausging, daß niemand a priori wissen könne, ob die Natur geordnet sei, zunächst also ein „Chaos" voraussetzte, war man im kontinentalen Raum von vornherein eher geneigt anzunehmen, daß Sein und Seiendes in einer bestimmten Ordnung gegeben seien. Man brauchte sich deshalb der Ordnung im „Kosmos" nur zu vergewissern; im Extrem genügte ein einziger untersuchter Fall, so bei K. Lewin. Beide Ansätze haben jeweils ihre Vor- und Nachteile.
Das *Erkundungsexperiment* geht die Phänomene direkt an. Die historisch bedeutsamen Experimente stammen aus dem kontinentalen Raum: Kepler, Galilei und Newton haben ihre epochemachenden Experimente unter der Annahme eines geordneten Kosmos durchgeführt. Diese Vorzüge werden allerdings durch eine Reihe von Gefahrenpunkten beeinträchtigt. Das Erkundungsexperiment versucht zu viel auf einmal zu erfassen. Es ist gegen Zufälle ungesichert. Auch ist es – wie schon angedeutet – schwerer, etwas zu beweisen als zu widerlegen. Vor allem aber besteht die *Gefahr unberechtigter Extrapolationen*. Das Problem der Generalisierung stellt sich allerdings auch beim Entscheidungsexperiment (vgl. noch Kap. 2.2.6.4).
Das *Entscheidungsexperiment* oder *„experimentum crucis"* hat den unbestreitbaren Vorteil, daß es gegen Zufälligkeiten gesichert ist. Es geht nicht von vorgefaßten Meinungen oder der Absicht aus, eine bestimmte These zu beweisen, sondern enthält die Intention, sich selbst zu widerlegen (vgl. Nullhypothese). Dem kommt eine logische Regel entgegen: Es ist einfacher, eine These zu widerlegen als zu beweisen. Genügt schon ein einziger Fall, um eine Annahme ad absurdum zu führen, so werden umgekehrt zur Erhärtung einer Hypothese zahlreiche Fälle benötigt. Nachteilig könnte sich hier das indirekte und zuweilen umständliche Herangehen an die Phänomene sowie ein übertriebenes Mißtrauen gegen die bestehende Ordnung im Kosmos auswirken. Ohne den Glauben an die (einmal erkannte) Ordnung der Natur, ihre Seinsbeständigkeit und Gesetzmäßigkeit wäre

freilich wissenschaftliche Forschung stets an den Anfang ihres Bemühens zurückgeworfen und somit praktisch ineffizient.
Eine weitere Unterscheidung experimenteller Methoden berücksichtigt das Milieu, in dem das Experiment stattfindet. Demzufolge spricht man von *Feldexperiment* versus *Laboratoriumsexperiment*. „Während das Feldexperiment im natürlichen Milieu abläuft, wird zur Durchführung eines Laboratoriumsexperiments erst ein Milieu hergestellt" (Bredenkamp 1969, S. 334). Gelegentlich spricht man in diesem Zusammenhang auch von *biotischen* (lebensnahen) vs. *künstlichen* Experimenten (= Laboratoriumsexperimenten).
Schließlich muß das *Feldexperiment* von der *Feldstudie* abgehoben werden. „Der Unterschied zwischen Feldexperiment und Feldstudie ... ist grundsätzlich: Während bei jenem der Vl die Bedingungen variiert, werden bei dieser *schon bestehende* Bedingungen, unter denen die AV beobachtet wird, selektiert (French 1953). Bei der Feldstudie besteht schon eine „natürliche" Variation, im Experiment wird die Variation erst „gemacht". So handelt es sich bei Untersuchungen zur Feststellung von Geschlechtsunterschieden nicht um Experimente, da der Faktor „Geschlecht" vom Vl nicht manipuliert werden kann. Andererseits sind nicht alle Untersuchungen, in denen der Vl die Bedingungen variiert, als Experimente zu klassifizieren. Dies trifft z.B. für den Fall zu, daß der Vl die Wirkung eines Films auf die Einstellung gegenüber einer Minorität prüfen will, ohne verhindern zu können, daß die Probanden sich quasi selbst den Bedingungen (Film/kein Film) zuteilen. In diesem Fall wäre das für das Experiment unerläßliche Kriterium der Kontrolle nicht erfüllt, da die Variation der AV zwischen den Bedingungen auch durch Persönlichkeitsunterschiede determiniert sein kann. (Diejenigen, die sich den Film ansehen, sind z.B. von vornherein positiver eingestellt.)" (Bredenkamp, a.a.O.).
In der Literatur werden mitunter noch andere Formen des Experiments erwähnt, etwa das sog. *ex post facto experiment*, dessen Bezeichnung jedoch irreführend ist. Im Grunde wird bei dieser Untersuchungsart überhaupt nicht manipuliert, womit ein wesentliches Kriterium des Experiments fehlt. „Bei dem sog. *ex post facto experiment* werden die Bedingungen nicht vom Vl variiert, sondern sie bestehen schon vor der Untersuchung. Da sich die Versuchspersonen, für die die Bedingungen zutreffen (z.B. Hochschulabschluß/kein Hochschulabschluß), bezüglich vieler anderer Faktoren unterscheiden, werden diese durch paarweises Parallelisieren (vgl. Kap. 2.2.4.3 bzw. 2.2.5.2; d. Verf.) zwischen den Bedingungen nachträglich homogenisiert. Diese Kontrolle führt (jedoch) ... zu einem statistischen Artefakt" (Bredenkamp 1969, S. 334f.). Damit wollen wir die Begriffsdiskussion bzw. Kriterienbestimmung des Experiments abschließen. Ergänzend zu der bereits angeführten Literatur verweisen wir besonders noch auf Festinger & Katz 1953, Lindquist 1953, Edwards 1954, 1960, Campbell 1957, Zimny 1961, Scott & Wertheimer 1962, Meili & Rohracher 1963, Holzkamp 1964, 1968 u.a.
Die experimentelle Forschung ist – besonders von geisteswissenschaftlicher Seite – schon viel geschmäht worden. Man warf dem Experiment Lebensferne, Künstlichkeit, Irrelevanz („Wissenschaft des Nichtwissenswerten") u.dgl.m. vor. In diesen vermeintlichen Nachteilen liegen jedoch wesentliche Vorzüge des Experimentierens beschlossen, wie der Aufweis der Kriterien eines wissenschaftlichen Versuchs verdeutlicht haben sollte. Dies heißt jedoch nicht, daß bereits sämtliche Probleme experimenteller Methodik gelöst seien oder dieser Verfahrensansatz universelle

Gültigkeit beanspruchen dürfte, also in jeder Situation und bei jedem Untersuchungsproblem die adäquate Methode darstellt. Andererseits begegnen wir experimentellen Theoremen und Prinzipien fast in allen Disziplinen und Teilbereichen der Sozialwissenschaften, z.B. der Lernpsychologie (vgl. Hilgard & Bower 1956), psychologischen oder pädagogischen Diagnostik (vgl. Cronbach 1960, Heller 1973), Unterrichtsforschung (vgl. Gage 1963, Campbell & Stanley 1963 bzw. Schwarz 1970, de Landsheere 1969), Sozialpsychologie (vgl. König 1956, 1962ff.; Graumann 1969, 1972) usw. Das Experiment in dieser oder jener Form gehört zum methodalen Grundkanon der empirischen Sozialwissenschaften und muß deshalb nicht nur den Wissenschaftlern, sondern allen in der einschlägigen Praxis Tätigen bzw. Studierenden der Pädagogik, Psychologie, Soziologie u.ä. zumindest in den Grundlagen vertraut sein. Mit den folgenden Ausführungen wenden wir uns nun der Versuchsplanung, der Versuchsdurchführung und der Auswertung experimenteller Ergebnisse zu.

2.2.4. Versuchsplanung

2.2.4.1. Allgemeine Probleme

Am Anfang jeder wissenschaftlichen Untersuchung, so auch beim Experiment, steht immer eine Problemfrage, z.B. die Frage, ob und ggf. in welchem Ausmaß bzw. wie die Schulleistung von der Intelligenz des Schülers abhängt. Zur Untersuchung dieser Frage wird der Vl einen *Versuchsplan (experimental design)* aufstellen, der etwa so aussieht.

Versuchsplan mit einer *variierten (gestuften) UV:*

Beispiel a)

Intelligenz (UV)	
über-durchschnittl.	unter-durchschnittl.
*)	

Beispiel b)

Intelligenz (UV)		
über-durchschnittl.	durch-schnittl.	unter-durchschnittl.
	*)	

*) Als AV fungiert hier jeweils die Schulleistung

Dependenzanalysen dieser Art erfordern einen *uni*variaten oder *uni*variablen Versuchsplan, d.h., *eine* UV wird in ihrem Einfluß auf eine (oder mehrere) AV untersucht. Eine solche (einfaktorielle) Plantafel kann eine unterschiedliche Anzahl von Zellen aufweisen, in unserem Beispiel a) $1 \times 2 = 2$ Zellen bzw. b) $1 \times 3 = 3$ Zellen.

Im klassischen Experiment haben wir es immer mit univariaten Versuchsplänen zu tun: *Eine* Bedingung (UV) wird systematisch variiert – bei gleichzeitiger Konstanthaltung (Kontrolle) aller übrigen Faktoren –, um so die Wirkung von UV

auf AV abschätzen zu können. Hierbei wird ausschließlich die UV vom Vl manipuliert, d.h. systematisch verändert (z.B. die Intelligenzvariable nach dem Grad der Ausprägung oder Höhe mehrstufig dimensioniert) und in ihrer Auswirkung auf die AV (z.B. Schulleistung) beobachtet. Die Veränderung der AV wird also vom Vl gleichfalls erwartet und beobachtet, jedoch nicht vom Vl manipuliert. UV sind alle Reizvariablen, AV meistens Reaktionsvariablen. Genausowenig wie die AV sind die IV (intervenierenden Variablen) vom Vl direkt manipulierbar; sein Einfluß hierauf ist immer nur mittelbar – über die systematische Variation der UV – möglich.
Die aufgewiesene Grundstruktur des Experiments mit der Unterscheidung in UV und AV gilt strenggenommen nur für den klassischen Untersuchungsansatz, das univariate Verfahren mit *einer* UV. Beim sog. *bi*variaten (zweifaktoriellen) versus *multi*variaten (mehrfaktoriellen) Verfahren werden zwei versus mehrere, selten „viele“, Variablen (UV) simultan vom Vl variiert. Versuchspläne mit zwei kontrollierten Variablen finden sich im 4. Kapitel dieses Buches (Korrelation). Beispiele multivariater Anordnungen repräsentieren die (multiple) Varianzanalyse versus die (multiple) Diskriminanzanalyse, die Faktorenanalyse oder die *Cluster*analyse (vgl. dazu Fröhlich 1971, Hofstätter & Wendt 1967, Überla 1968, Weber 1967 u.a.).
*Dependenz*analysen (Varianzanalysen) mit mehr als 3 UV sind sehr selten. Gewöhnlich wird in den sog. multivariaten Experimenten eine mehr oder minder große Zahl Variablen simultan verarbeitet und kontrolliert, ohne daß zwischen UV und AV unterschieden wird: So untersucht man in der Faktorenanalyse lediglich *inter*dependente (keine dependenten) Zusammenhänge einzelner Variablen, d.h., hier interessiert der Aufweis der Wechselbeziehungen zwischen den – nicht nach UV und AV unterschiedenen – Variablen, um komplexe (Variablen-)Strukturen mathematisch definieren zu können. Genaugenommen haben wir es hier also mit *Inter*dependenzanalysen, nicht mehr mit Dependenzanalysen zu tun. Die Besprechung dieser komplizierten Verfahren würde jedoch den Rahmen einer Einführung in die Methodenlehre sprengen; Interessierte verweisen wir auf die oben genannte Literatur. Zur Veranschaulichung eines „multivariaten“ (bivariaten) Versuchsplanes greifen wir erneut unser Beispiel auf, jetzt aber mit zwei Variablen (UV) und $2 \times 3 = 6$ Zellen.

Zwei*faktorieller Versuchsplan:*

Beispiel c)		Intelligenz (UV_1)		
		über durchschnittl.	durch-schnittlich	unter-durchschnittl.
Geschlecht (UV_2)	männlich			
			*)	
	weiblich			

*) AV ist wiederum die Schulleistung

Im letzten Beispiel einer Dependenzanalyse wird bereits deutlich, daß der Manipulierbarkeit i.e.S. Grenzen gesetzt sind. So kann der Vl nicht das Geschlecht (UV_2) seiner Vpn variieren, sondern muß die natürliche Variation berücksichtigen, d.h. in unserem Beispielfall selektieren (zwischen männlichen vs. weiblichen Vpn). Analog bedient sich der Vl bei *Inter*dependenzanalysen der vorgefundenen Variation und richtet sein Augenmerk auf die Zusammenhänge zwischen den Variablen, ohne im einzelnen Ursache und Wirkung zu beachten – oder zutreffender – beachten zu können.
Die Grundstruktur des Experimentierens wird also am besten beim oben beschriebenen klassischen Ansatz der Dependenzanalyse einsichtig. Der Vl variiert dabei systematisch die (eine) UV und hält alle übrigen Einflußgrößen auf die AV konstant, d.h. kontrolliert die sog. *Störvariablen* (z.B. interindividuellen Differenzen der Vpn). Darin liegt zugleich ein tiefgreifender Unterschied zum psychodiagnostischen Experiment oder Test. Während sich im klassischen Experiment der Blick auf die allen Individuen gemeinsamen Verhaltensweisen und somit die Aufdeckung von Gesetzmäßigkeiten richtet (*nomothetische* Zielsetzung des sog. *Forschungs*experiments), z.B. Gesetzlichkeit der Wahrnehmung, des Denkens, des Lernens usw., gilt das Hauptinteresse des diagnostischen Experiments versus der differentiellen Psychologie (Persönlichkeitspsychologie) den interindividuellen Unterschieden, also dem individuell Eigenartigen (*idiographische* Zielsetzung des sog. *Prüf*experiments oder Tests), z.B. den Intelligenz- und Leistungsdifferenzen oder ähnlichen intersubjektiven Merkmalsausprägungen bzw. Merkmalsvariationen. M.a.W.: Gegenstand der diagnostischen und Persönlichkeitspsychologie sind eben die sog. Störvariablen des Forschungsexperiments (vgl. dazu ausführlicher Heller 1973, S. 57ff., wo sich weitere Literaturhinweise, z.B. Cronbach 1960, finden).
Im Forschungsexperiment werden nun diese Störvariablen, d.h. alle Einflußvariablen auf die AV außer der/den variierten UV, immer konstant gehalten. Die *Kontrolle* im Experiment beinhaltet also zwei Aspekte: einmal die *Variation* der UV und zum andern die *Konstanthaltung* aller übrigen Einflußvariablen auf die AV. Diese Prinzipien gilt es in jedem Versuchsplan (dependenzanalytischer Art) zu berücksichtigen. Entsprechende Anwendungsbeispiele finden sich im mathematisch-statistischen Teil dieses Buches in größerer Zahl. Dadurch sollen dem Leser die verschiedenen Möglichkeiten der Versuchsplanung anhand konkreter Untersuchungsbeispiele – jeweils unter Berücksichtigung der relevanten Methoden – vorgestellt werden.
Bevor wir uns nun den einzelnen Kontrolltechniken, dem Herzstück des Experiments, zuwenden, sei kurz auf die wichtigsten Fehlerquellen versus Gütekriterien des Experiments eingegangen.

2.2.4.2. Experimentelle Fehler und Gütekriterien des wissenschaftlichen Versuchs

Neben der UV tragen weitere Faktoren zur Variation der AV bei, z.B. interindividuelle Differenzen (Alter, Geschlecht, Intelligenz u.ä.) der Vpn. Diese werden im Experiment – im Gegensatz zur klassischen Testtheorie – als *Fehler* angesehen, da sie die (Inter-)Dependenzanalyse mehr oder weniger empfindlich stören. Bredenkamp (1969, S. 337ff.), der sich auf Cox, Campbell, Campbell & Stanley u.a. be-

ruft, stellt drei Fehlerklassen heraus, die in Zusammenhang mit den Haupt*güte-kriterien* des Experiments, nämlich der Validität und Präzision, stehen. Demnach kann man *systematische* Fehler (*interne* vs. *externe* systematische Fehler) und *zufällige* Fehler unterscheiden, die wiederum die *Validität* (die *interne* vs. *externe* Validität) sowie die *Präzision* (im Sinne Lienerts) beeinflussen.
Von einem *systematischen internen Fehler* spricht man, wenn sich die Vpn hinsichtlich unbekannter bzw. unkontrollierter Variablen, die die Variation der AV mitbedingen, unterscheiden. Wird eine solche Konfundierung *(confounding)* vermieden, bezeichnet man das Experiment als *intern valide*. Der Typ des *systematischen externen* Fehlers betrifft ein Generalisationsproblem. Fehler dieser Art liegen vor, wenn die Bedingungen des Experiments nicht vergleichbar sind mit der (späteren) Situation, für die das experimentelle Ergebnis gültig sein soll. Kann dieser Fehler eliminiert werden, so ist das Experiment *extern valide*. Der dritte Fehlertyp tritt im Gefolge der Kontrolle des systematischen internen Fehlers in Erscheinung und stellt gewissermaßen ein Artefakt dar. Die Kontrolle systematischer interner Fehler, z.B. der O-Variablen, geschieht in der Regel mit Hilfe der sog. Randomisierungstechnik, d.h. der zufälligen Zuordnung der Vpn zu den unterschiedlichen Versuchsbedingungen (s. folg. Abschn.). „Da aber mehrere Vpn unter einer experimentellen Bedingung beobachtet werden, kann die AV zusätzlich auch *innerhalb* der Bedingungen variieren. Bei den Determinanten dieser Variation handelt es sich um die *zufälligen Fehler* .... Die Variation der zufälligen Fehler steht in direktem Zusammenhang mit der Präzision eines Experiments: Je kleiner die Variation dieser Fehler ist, desto größer ist die *Präzision*“ (a.a.O.). Nach Hays (1969, S. 381ff.) ist darüber hinaus die statistische Relation zwischen UV und AV (Omega^2-Wert) um so größer, je geringer eben diese Variation der zufälligen Fehler ist. Dies steht wiederum in Zusammenhang zur im Experiment intendierten Bestätigung der H_1-Hypothese (Alternativhypothese zur H_0), die durch die Erhöhung der Präzision größere Wahrscheinlichkeit erlangt.

2.2.4.3. *Methoden zur Kontrolle der Versuchsbedingungen*

Kennzeichnend für das Experiment sind vor allem zwei Prinzipien: die *systematische Variation* der in ihrem Einfluß auf die AV untersuchten UV und die *Kontrolle* aller übrigen Einflußgrößen (Störvariablen bzw. Fehler). Im folgenden sollen nun die wichtigsten Kontrolltechniken besprochen werden, wobei wir uns wiederum an den systematischen Aufriß Bredenkamps (a.a.O.) anlehnen. Die zu erörternden Techniken dienen vor allem der Sicherung der internen Validität sowie zur Erhöhung der Präzision. Probleme der externen Validität, das sog. Repräsentanzproblem im Sinne Holzkamps sowie einige kritische Anmerkungen zur Relevanz experimenteller Forschung überhaupt, sollen im nachfolgenden Abschnitt zur Sprache kommen. Gleichwohl müssen die interne und die externe Validität in einem gewissen Zusammenhang gesehen werden, weshalb einige Aspekte der externen Validität schon jetzt mit angesprochen werden.
1. Die Kontrolltechnik der *Konstanthaltung* bezieht sich auf alle vom Vl hergestellten Bedingungen bzw. möglichen Störfaktoren. Dadurch soll der Einfluß jener Faktoren, die außer der UV eine mögliche Variation der AV mit determinieren, neutralisiert werden. Dieses Verfahren stößt in der Praxis vielfach auf erhebliche

Schwierigkeiten; es steigert zudem nicht nur die interne Validität, sondern kann auch die externe Validität beeinträchtigen, worauf besonders Campbell, Campbell & Stanley u.a. aufmerksam gemacht haben. So sind die Ergebnisse u.U. Vl-abhängig, d.h. nur für den betr. Vl gültig.

2. Die Methode der *Elimination* geht auf die Beseitigung der Störfaktoren aus und ist vor allem beim Laboratoriumsexperiment einschlägig. Hier ist jedoch die Gefahr gegeben, daß „zuviel" vom natürlichen Milieu, d.h. wichtige Einflußgrößen der Alltagssituation eliminiert werden. Daraus kann eine gewisse Künstlichkeit der experimentellen Befunde resultieren, so daß die externe Validität in Frage gestellt wird.

3. Die Kontrolltechnik der *Randomisierung* wird oft als die beste Methode zur Sicherung der internen Validität des Experiments angesehen. Sie ist vor allem für die Kontrolle *unbekannter* Einflußgrößen bzw. Fehlervariablen wichtig. Dabei werden die Vpn oder Vpn-Gruppen rein zufällig den unterschiedlichen Versuchsbedingungen zugeteilt. So würde man etwa in einem Experiment, das die Effizienz einer neuen Unterrichtsmethode zum Untersuchungsziel hat, die Schüler (Vpn) bzw. Schulklassen (Vpn-Gruppen) per Losentscheid bzw. mit Hilfe sog. Zufallsziffern (vgl. Tab. I im Anhang dieses Buches) den zum Vergleich anstehenden Methodengruppen zuteilen. So entstünden zwei Vpn- bzw. Klassengruppen (Experimentalgruppe vs. Kontrollgruppe); Klassengruppe 1: Unterricht nach der *neuen* Methode; Klassengruppe 2: Unterricht nach der *alten* Methode. Ein solches Vorgehen erfordert unbedingt zwei Voraussetzungen: a) Jede Vp muß die *gleiche Chance* haben, in die Versuchs- vs. Kontrollgruppe zu kommen; b) genügend große Stichproben sind hier unerläßlich, damit das „Gesetz der großen Zahl" wirksam werden kann, d.h. die gegenseitigen Einflüsse der Störvariablen zu Null aufaddiert und damit neutralisiert werden.

4. In der experimentellen Praxis kann man sich nicht immer den für die Randomisierung erforderlichen Aufwand, dessen Hauptproblem der Stichprobenumfang darstellt, leisten. Deshalb wird häufig auf die Stichproben*parallelisierung (matching)* als Kontrollmethode ausgewichen. Einschlägig ist hier wiederum der Versuchs-Kontrollgruppen-Vergleich. Der Vl stellt zwei homogene Stichproben, deren Vpn sich hinsichtlich der relevanten Merkmale (z.B. Intelligenz, Geschlecht usw.) nicht unterscheiden dürfen, zusammen. Auf diese Weise entsteht die bekannte Anordnung der *Versuchsgruppe* vs. *Kontrollgruppe;* erstere wird den zu untersuchenden Bedingungen, d.h. der systematisch variierten UV, ausgesetzt, letztere nicht. Da dies der einzige Unterschied zwischen den beiden Vpn-Gruppen ist, kann somit der Einfluß der UV auf die AV hinreichend verläßlich beobachtet werden. (Zur Technik dieser speziellen Stichprobenbildung s. Kap. 2.2.5.2 unten.)

Zu den obligatorischen *Kontrollfaktoren* im pädagogisch-psychologischen Experiment gehören Alter, Geschlecht, Bildungsstufe, Intelligenz u.ä. Sofern die Korrelation zwischen der/den Kontrollvariablen und der AV mindestens 0.2 beträgt, ist mit einer Präzisionssteigerung des Experiments via Stichproben-Matching zu rechnen. Allerdings kann auch hier wieder die externe Validität eingeschränkt sein. Ferner kommen bei *nachträglicher* Homogenisierung der Vpn oder Vpn-Gruppen – analog zum sog. *ex post facto experiment* – leicht Regressionsartefakte ins Spiel (vgl. Campbell & Stanley 1963).

5. Quasi eine Variante der Parallelisierung ist die sog. *Cross*-Validierung als Kon-

trollmethode des Experiments. Hierbei wird der Versuch zweimal an denselben Vpn-Gruppen (Experimental- und Kontrollgruppe) durchgeführt, nur daß beim zweitenmal Versuchs- und Kontrollgruppe vertauscht werden. Es handelt sich also hier um eine gekreuzte Experimental-/Kontrollgruppen-Anordnung:

	1. Versuch:	*2. Versuch:*
Vpn-Gruppe 1:	Exper.-Gruppe	Exper.-Gruppe
Vpn-Gruppe 2:	Kontr.-Gruppe	Kontr.-Gruppe

Will man beispielsweise den Einfluß der positiven Bekräftigung auf die Schul- bzw. Lernleistung untersuchen, so würde man im ersten Versuch zunächst die Vpn-Gruppe 1 (= Experimentalgruppe) bekräftigen, die Vpn-Gruppe 2 (= Kontrollgruppe) jedoch nicht, während man im zweiten Versuchsdurchgang gerade umgekehrt vorgehen würde: Jetzt würde Vpn-Gruppe 2 zur Experimentalgruppe und Vpn-Gruppe 1 zur Kontrollgruppe bestimmt werden. Unerläßliche Voraussetzung für die *Cross*-Validierung ist natürlich wiederum die Parallelität der Vpn-Gruppen 1 und 2 hinsichtlich der Kontrollvariablen (Geschlecht, Alter, Intelligenz usw.). Daß dieses Verfahren bislang nicht häufiger zur Anwendung gelangte, dürfte weniger an der mangelnden Einsicht in die Effizienz dieses Ansatzes als am hierfür notwendigen Aufwand liegen.

6. Die Methode des sog. *Ausbalancierens* wird notwendig, wenn ein und dieselben Vpn mehrere Versuchsbedingungen – nacheinander – durchlaufen. Auf diese Weise sollen systematische interne Fehler, die auf serielle Einflüsse zurückzuführen sind, ausgeschaltet werden. Zimny (1961) nennt drei Methodenvarianten: das intraindividuelle Ausbalancieren, das interindividuelle vollständige Ausbalancieren und das interindividuelle unvollständige Ausbalancieren (s.a. Bredenkamp 1969, S. 346f.).

7. Lange Zeit unbemerkt blieb die Rolle des/der Vl als Störfaktor im Experiment. Während man die systematischen Fehler und direkten Einflußmöglichkeiten des Vl (z.B. Fehler bei der Zeigerablesung, der Verrechnung usw. – von der sog. persönlichen Gleichung einmal abgesehen – oder Betrug u.ä.) schon frühzeitig kontrollierte, machten auf die sog. indirekten Beobachtereffekte eigentlich erst Rosenthal und seine Mitarbeiter innerhalb des letzten Dezenniums nachdrücklich aufmerksam. Hiermit sind Beeinflussungen der Vpn gemeint, die aus bestimmten Erwartungshaltungen des Vl resultieren. So ließ Rosenthal verschiedene Vl-Gruppen mit gleich intelligenten Tieren (Ratten) Versuche durchführen, die sich nur bezüglich einer einzigen Bedingung unterschieden: Der einen Vl-Gruppe suggerierte er „intelligente" Tiere, der anderen Vl-Gruppe „unintelligente" Tiere. Obwohl alle übrigen Bedingungen konstant gehalten wurden, erzielten am Ende des Experiments die vermeintlich intelligenten Tiere bessere Resultate als die vermeintlich dummen.

Wenngleich die Rosenthalschen Befunde nicht unkritisiert blieben und der eine oder andere Aspekt experimenteller Anordnung und Verrechnungsmethodik selbst nicht völlig fehlerfrei sein mag, so bleibt doch aufgrund der Ergebnisse zahlreicher Nachuntersuchungen der sog. Rosenthaleffekt als Gefahrenquelle vieler experimentalpsychologischer bzw. sozialwissenschaftlicher Untersuchungen bestehen. Selbst Veranstaltungen im Doppelblindversuch (wo also weder die Vpn noch der/die Vl über Ziel und Zweck des Experiments informiert sind) bieten keine hinreichend sichere Gewähr für die Eliminierung solcher Erwartungseffekte, wie die Rosenthalschen Untersuchungen eindrucksvoll demonstrieren. Ähnliche Einflüsse sind durch diverse andere Persönlichkeitsmerkmale des Vl im Hinblick auf die Reaktionen der Vpn und damit die Versuchsergebnisse bei der Kontrolle im Experiment zu berücksichtigen. Geschlecht, Aussehen, Alter, Sprachweise u. ä. des Vl können in je spezifischer Weise das Verhalten der Vpn beeinflussen. Um Effekte dieser Art unter Kontrolle zu bringen, schlägt Rosenthal vor, der Experimental- und einfachen Kontrollgruppe noch eine dritte Vpn-Gruppe, die sog. *Erwartungskontrollgruppe,* hinzuzufügen:

Experimentalgruppe	Kontrollgruppe	Erwartungskontrollgruppe

Hierbei sollen die Vl der Erwartungskontrollgruppe dieselben Informationen erhalten wie die Vl der Experimentalgruppe, so daß sich bei beiden Vl-Gruppen identische Erwartungshaltungen einstellen. Entsprechen dann die Ergebnisse der Erwartungskontrollgruppe denen der Experimentalgruppe und weichen diese signifikant von denen der einfachen Kontrollgruppe ab, so kann daraus auf die Wirkung von Vl-Effekten geschlossen werden. Zeigen sich dagegen gleiche Ergebnisse bei der Kontrollgruppe und der Erwartungskontrollgruppe, die diesmal signifikant von den Resultaten der Experimentalgruppe abweichen müssen, so sieht Rosenthal die Annahme einer alleinigen Wirksamkeit der UV gestützt.
Auch gegen die Einführung einer zweiten Kontrollgruppe (Erwartungskontrollgruppe) läßt sich manche Kritik anbringen, z.B. die Frage, ob und *wie* man konkret die Forderung nach gleicher Erwartungshaltung und deren Induktion durch den Experimentator realisieren kann. Wie überall bei komplexeren Untersuchungsvorhaben in der Wissenschaft gibt es auch hier im Rahmen der Kontrolle experimenteller Determinationsvariablen keine universell gültigen Techniken. Die Kontrolle im Experiment hängt sehr eng mit der Versuchsplanung und -auswertung und somit dem konkreten Untersuchungsvorhaben einschließlich der intendierten Zielsetzung zusammen. Auf weitere Kontrolltechniken, z. B. die *Kovarianzanalyse,* können wir hier nicht mehr eingehen. Wer selber experimentell aktiv werden möchte, wird nicht um die Lektüre der Spezialliteratur, von der wir bereits eine stattliche Anzahl nannten, herumkommen. Über die neueste Entwicklung auf dem Gebiet der Versuchsplanung und Auswertung experimenteller Untersuchungen informieren laufend Beiträge in der Zeitschrift für experimentelle u. angewandte Psychologie, Annual Review of Psychology, Review of Educational Research u.a.

2.2.4.4. Probleme der externen Validität

Mit dem Problem der externen Validität ist das sog. *Repräsentanzproblem* im Sinne Holzkamps (1964, 1968) angesprochen. Hier ist nach der Generalisierbarkeit experimenteller Ergebnisse gefragt und damit in Zusammenhang stehend nach der *praktischen Relevanz* experimenteller Forschung überhaupt. Doch wenden wir uns zunächst dem ersten Aspekt, dem Problem der externen Validität i. e. S., zu.
Die Forderung nach Repräsentativität gilt für die experimentellen Variablen (UV vs. AV), die Stichproben (Vpn) und die experimentelle Situation (Vl) gleichermaßen. Die ausgewählten Variablen sind quasi als repräsentative Stichproben

einer zugehörigen Population von UV vs. AV zu verstehen. Die UV-Repräsentanz wird am ehesten durch eine Serie von Experimenten unter gleichen Bedingungen. jedoch „mit anderen Modalitäten des Bedingungsfaktors" kontrolliert werden können. Die AV-Repräsentanz betrifft z.B. die Wahl der Aufgaben und deren Generalisationsbasis im Hinblick auf andere, das „Gleiche" messende Aufgaben ähnlicher Art. Sofern externe Validität gegeben sein soll, darf die im Experiment aufgedeckte Beziehung nicht nur für die im konkreten Fall verwendete AV zutreffen, sondern muß einen gewissen Grad allgemeiner Gültigkeit repräsentieren. In engem Zusammenhang damit steht das Problem der Stichprobenrepräsentanz. Nicht selten stehen für Experimente im sozialwissenschaftlichen Bereich nur „Freiwillige" zur Verfügung. In diesen Fällen dürfte strenggenommen nur auf die Population der „Freiwilligen" generalisiert werden, wenigstens solange nicht sichergestellt ist, daß keine relevanten Unterschiede zwischen den zum Versuch bereiten und den die Teilnahme verweigernden Vpn bestehen. Auch die Anonymität der Vpn kann in bezug auf das experimentelle Ergebnis eine Rolle spielen, z.B. im Rahmen einer Fragebogenermittlung. Schließlich gibt es eine Reihe von Techniken, die durch Modifikation der „äußeren" Umstände die externe Validität zu erhöhen suchen: etwa wenn bei Untersuchungen im „unwissentlichen Verfahren" die Vp über den eigentlichen Zweck des angesetzten Experiments nicht nur im unklaren gelassen, sondern absichtlich getäuscht wird, z.B. eine projektive Persönlichkeitsdiagnostik als Intelligenztest ausgegeben wird. Neben der *Täuschungsmethode (deception)* hat in sozialwissenschaftlichen Experimenten vor allem noch die *Simulation* eine Bedeutung. Dadurch wird beispielsweise versucht, die Künstlichkeit des Laboratoriumsexperiments aufzufangen. Der Erfolg solcher Simulationsstudien im Hinblick auf die Erhöhung der externen Validität hängt nach Bredenkamp (a.a.O.) von zwei Voraussetzungen ab: „1. Im Kontext der simulierten Bedingungen muß... experimentiert werden können. Viele Faktoren, die in einem üblichen Laboratoriumsexperiment als Störfaktoren eliminiert werden, sind in einem Simulationsexperiment repräsentiert. 2. Die Validität eines Simulationsexperiments hängt von der Fähigkeit der Vpn ab, die geforderten Rollen spielen zu können."

Was endlich die Beziehung von Laboratoriums- und Feldexperiment zur internen vs. externen Validität betrifft, so gilt allgemein: Beim Laboratoriumsexperiment wird die interne Validität sehr oft auf Kosten der externen Validität optimiert, d.h., hier ist eine strengere Bedingungskontrolle möglich (im Idealfall wird die UV in ihrem Einfluß auf die AV vollständig aufgeklärt). Beim Feldexperiment wird weniger die externe als die interne Validität zum Problem, das Feldexperiment ist stärker effektzentriert (in der Regel werden hier mehrere UV bzw. ein ganzer Komplex von Variationsmöglichkeiten der AV untersucht), wodurch nicht selten die interne Validität in Mitleidenschaft gezogen wird. Beide Formen sind somit im Grunde keine konkurrierenden, sondern sich ergänzende Methodenansätze. Diese Feststellung gilt u.E. in besonderem Maße für die aktuelle experimentelle Forschungssituation im pädagogisch-psychologischen bzw. sozialwissenschaftlichen Bereich, z.B. der Unterrichtsforschung, Lernpsychologie, Erziehungspsychologie u.ä., wenngleich gerade hier bislang ein Überhang an Laboratoriumsexperimenten versus ein Nachholbedarf an Feldexperimenten (in der Schule, im Klassenzimmer, in der Familie, im Internat usw.) konstatiert werden muß – sofern man überhaupt von einem „Überhang" in diesem Zusammenhang sprechen darf.

Fortschritte sozialwissenschaftlicher Provenienz innerhalb der letzten Dezennien beruhen zu einem erheblichen Teil auf der methodologischen Entwicklung, d. h. der „Durchsetzung des Operationismus als methodischem Grundsatz", der „Verfeinerung und Präzisierung der Konzepte des experimentellen Designs und der statistischen Inferenz" sowie der „Ausarbeitung des hypothetico-deduktiven Verfahrens als eines formalen Kanons der Theorienbildung und Hypothesenherleitung" (Holzkamp 1970, S. 2). Hierin scheint die Psychologie gegenüber den übrigen Sozialwissenschaften eine führende Rolle übernommen zu haben (vgl. Heller 1972, S. 58). Auf der anderen Seite wird jedoch immer deutlicher, „daß weite Bereiche der psychologischen Forschung durch fortschreitende Desintegration und Banalisierung gekennzeichnet sind, ... daß die experimentelle Forschung in immer wachsendem Maße mit exakten Methoden Belanglosigkeiten und Trivialitäten zutage fördert", wie Holzkamp (a.a.O., S. 2ff. u. 7f.) etwas provozierend formuliert.
Ganz offenbar erfordert der heute vielfach „positivistisch eingeengte Methodenbegriff der Psychologie" eine Revision; ob der Ausweg allerdings im hermeneutischen Ansatz tradierter Prägung zu suchen ist, möchten wir doch bezweifeln. Sofern ein ausgewogenes Verhältnis zwischen theoretischem und empirischem Ansatz, d. h. die Konvergenz deskriptiv-phänomenologischer und operationaler Vorgehensweisen, als Anliegen nicht aus dem Auge verloren wird, sähen wir die größte Gefahr reduktionistischer Tendenzen, die zweifellos da und dort vorhanden sind, bereits gebannt. Sicherlich wird auch einer stärkeren Beachtung der „anthropologischen Relevanz" in der Psychologie wie überhaupt in den Sozialwissenschaften in Zukunft Bedeutung zukommen. Konkret bedeutet dies, daß stärker als bisher Anstrengungen unternommen werden müssen, um ein sachlich ausgewogenes Verhältnis zwischen *methodologischem Anspruch* und *empirischer Valenz des Zielgegenstandes* zu finden, was u. U. erfordert, „daß man auf weitere Präzisierung ... verzichten müßte, wenn die Relevanz durch diese Präzisierung reduziert werden würde" (a.a.O., S. 14). Anders gewendet: Die Repräsentanz des operational angegangenen Untersuchungsgegenstandes darf nicht auf Kosten methodischer Präzisionsansprüche vernachlässigt werden, obschon sich hier fast zwangsläufig gegenläufige Tendenzen offenbaren. Aber handelt es sich tatsächlich um inkompatible Forderungen? Trotz sehr ernstzunehmender Schwierigkeiten, die sich bei der Lösung dieses Ambivalenzkonfliktes jedem Wissenschaftler – sofern dieser sein Tun im eingangs definierten Sinne versteht – in den Weg stellen, möchten wir eine prinzipielle Unmöglichkeit im Zusammenhang unserer Fragestellung verneinen. Freilich bedarf in jedem Falle der operationistische Ansatz einer – vorausgehenden – gegenstandsadäquaten (semantisch und syntaktisch eindeutigen bzw. korrekten) Phänomenanalyse, wie umgekehrt diese des Präzisierungseffektes operationaler Bearbeitung nicht entraten kann. Allzuleicht könnte es sonst passieren, daß wir nur einen (wenn auch noch so sicher gefaßten) Zipfel der Wirklichkeit in Händen halten oder, im andern Falle, uns alles zwischen den Fingern zerrinnt. Die Betonung der *Notwendigkeit einer Konvergenz von phänomenologischem und operationalem Ansatz* ist eine fundamentale Prämisse, ohne die weder unser Wissenschaftsverständnis allgemein noch in bezug auf die Sozialwissenschaften richtig begriffen würde. Für den orthodoxen Operationalisten mag dies fast wie ein „Bekenntnis" anmuten, was es eigentlich nicht ist. Was bringt es, wenn man Hypothesen veri- oder falsifiziert, ohne einen genauen Begriff davon zu haben,

was man überhaupt untersuchen will, ganz abgesehen von der Unmöglichkeit in diesem Fall, die Befunde in den Kontext einzuordnen. Nicht der Verzicht auf den einen oder den anderen Ansatz bringt uns weiter, sondern das In-Fragestellen (erkenntnis)theoretischer Positionen und kritische Überprüfung der methodologischen Voraussetzungen empirisch-operational gewonnener Befunde. Daß eine fruchtbare Forschung im Bereich der Sozialwissenschaften zunächst vor allem *methodischer Innovationen*, vorab in den erziehungswissenschaftlichen Fächern, bedarf, steht u. E. dazu nicht im Gegensatz. Daraus folgt vielmehr, daß gründliche Methodenkenntnis heutzutage curricularer Bestandteil der Ausbildung angehender Erziehungswissenschaftler, Pädagogen, Psychologen oder Soziologen sein muß, will man sich nicht der Gefahr aussetzen, in nationalen oder internationalen Provinzialismus (zurück) zu fallen. Die Vermittlung grundlegender Methodenkenntnisse ist deshalb Hauptanliegen dieses Buches.

2.2.4.5. Hypothesenbildung

Wörtlich übersetzt bedeutet „Hypothese" eine „Unterstellung" oder unbewiesene Erklärung. Damit soll versuchsweise eine mögliche Antwort auf eine Frage bzw. einen problematischen Sachverhalt gegeben werden. Der erste Schritt einer wissenschaftlichen Untersuchung ist somit die Bildung einer Hypothese. Man kann verschiedene Arten von Hypothesen unterscheiden, z. B. Ausgangshypothese, Erklärungshypothese usw. oder theoretische, empirische bzw. experimentelle Hypothese. Hier ist letztere Form von Bedeutung.

Von einer *experimentellen Hypothese* sprechen wir, wenn die Hypothese bestimmten Kriterien genügt. Dabei werden allgemein die *operationale Definition* (s. S. 20f.) und die *Überprüfbarkeit* der Hypothese als wichtigste Anforderungen betrachtet. Townsend (1953, S. 45 f.) stellte folgende Kriterien zusammen:

1. Eine Hypothese muß eine *adäquate Antwort auf ein spezifisches Problem*, das *klar definiert* ist, sein. Eine solche Hypothese antwortet immer nur auf *einen* Aspekt des Problems, sie darf also nicht zwei Antworten zulassen – ausgenommen die Form der Alternativantwort (Ja – Nein). Wenn mehrere Antworten auf eine Hypothese möglich sind, ist die Hypothese zu locker formuliert.
2. Eine Hypothese sollte stets die *einfachste* Antwort auf ein Problem ermöglichen.
3. Eine Hypothese muß streng *verifizierbar* sein, was die Forderung experimenteller Überprüfung einschließt. Annahmen, die sich einer strengen Verifizierung entziehen, sollte man besser nicht als „Hypothesen" bezeichnen.
4. Eine Hypothese muß *falsifizierbar* sein. Sätze wie „Die Menschen bekriegen sich, weil sie einen Aggressionstrieb haben" oder „Wir essen, weil wir Hunger haben" können genaugenommen keine Hypothesen im explizierten Sinne sein. Aggressionstendenzen zeigen sich nicht nur im Krieg, bzw. man kann auch ohne Hunger essen. Besonders die Persönlichkeitspsychologie und Pädagogik, die Tiefenpsychologie, die Motivationspsychologie u.ä. enthalten solche Formulierungen im Übermaß. Daraus kann man jedoch keine echten Hypothesen gewinnen, d.h., solche Sätze sind nicht veri- oder falsifizierbar.

Einschlägig für die Hypothesenprüfung ist das *Entscheidungsexperiment* (s. S. 54f.), das eine *Alternativhypothese* erfordert, d.h., die Antwort wird hierbei alternativ vorformuliert. Die Alternativhypothese, die meistens die Erwartungen des Ex-

perimentators ausdrückt, wird abgekürzt H_1 bezeichnet. Davon abzuheben ist die sog. *Nullhypothese,* abgekürzt: H_0 bezeichnet (s. unten). So liest man in der Literatur häufig folgende oder ähnliche Formulierungen: „Die Hypothese X – gemeint ist dabei immer die sog. Alternativhypothese, nicht die Nullhypothese – konnte verifiziert werden" vs. „Die Hypothese X wurde falsifiziert" oder, exakter: „H_0 wird zugunsten von H_1 verworfen" vs. „H_0 muß beibehalten werden" usw. Die Nullhypothese (H_0) besagt nämlich nichts anderes, als daß die Differenz zwischen den zur Untersuchung anstehenden bzw. beobachteten Variablen gleich „null" ist. Genauer müßten wir formulieren: „... daß die Differenz ‚nur zufällig' ist", denn mehr wird eigentlich nicht ausgesagt. Mit anderen Worten: Die Nullhypothese behauptet, daß zwischen den verschiedenen Untersuchungsvariablen kein operational-statistisch aufweisbarer Unterschied besteht, d.h., daß die beobachteten Differenzen „statistisch *nicht* bedeutsam" bzw. „*nicht* signifikant" sind. Es ist wichtig, sich hierbei immer den Bezug der Nullhypothese zur „Nur-Zufälligkeit" vor Augen zu halten, denn vor dem Hintergrund statistischer Entscheidungsmodelle (vgl. Kap. 5.1) beinhalten Aussagen der vorgenannten Art immer auch eine gewisse Irrtumswahrscheinlichkeit, d.h. eine bestimmte Wahrscheinlichkeit dafür, daß man sich in der Beibehaltung der Nullhypothese geirrt hat. Entsprechend kann man sich – ebenfalls – irren, wenn man die Nullhypothese zurückweist, d.h. die beobachteten Differenzen als überzufällig interpretiert, das experimentelle Ergebnis demnach als „signifikant" oder „sehr signifikant" bzw. „hochsignifikant" ansieht (vgl. Kap. 5.2).
Daß man bei der Hypothesenbildung auf die Form der Nullhypothese zurückgreift, hat auch noch andere, bereits früher angedeutete Gründe. Die Nullhypothese negiert die Erwartungseinstellung (H_1) des Experimentators, der somit gezwungen ist, sich selbst zu widerlegen. Dadurch wird der sog. *experimenter bias* in etwa neutralisiert. Ferner ist es einfacher, eine These zu widerlegen als zu beweisen (s. S. 54).
Im erörterten Kontext sei schließlich noch auf die Unterscheidung von „einseitiger Fragestellung" versus „zweiseitiger Fragestellung" hingewiesen. Folgende Hypothesenbeispiele mögen die ein- versus zweiseitige Problemstellung erklären.

H_0 bei *einseitiger* Fragestellung: „Überdurchschnittlich intelligente Kinder zeigen keine besseren Schulleistungen als unterdurchschnittlich begabte. Eventuell zu beobachtende Besserleistungen der überdurchschnittlich begabten Schüler sind nicht signifikant."

H_0 bei *zweiseitiger* Fragestellung: „Überdurchschnittlich intelligente und unterdurchschnittlich intelligente Kinder unterscheiden sich nicht (nicht signifikant) bezüglich ihrer Schulleistungen."

Während bei der zweiseitigen Fragestellung nichts über die Richtung des möglichen – durch die Form der H_0 zunächst in Frage gestellten – Unterschiedes zwischen den über- und den unterdurchschnittlich begabten Schülern ausgesagt wird bzw. bekannt ist, drückt der Vl bei der Hypothesenformulierung nach dem Modus der einseitigen Fragestellung klar aus, zu wessen Gunsten er die Unterschiede erwartet. Bei der einseitigen Fragestellung müssen also schon Informationen über vorliegende Differenzen, d.h. empirische Belege für die Tatsache des Unterschiedes in der genannten Richtung, bekannt sein. Die Nullhypothese bei

einseitiger Fragestellung bezieht also ausdrücklich die Möglichkeit ein, daß H_1 – wenigstens teilweise – zutrifft.
Die Einteilung in ein- und zweiseitige Problemstellungen ist vorab im Hinblick auf die statistischen Prüftests (vgl. Kap. 5.3 u. 5.4) von Bedeutung. Bei einseitiger Fragestellung sind die betr. Stichprobentests „stärker", d.h. eine beobachtete Differenz ist mit Hilfe sog. einseitiger Tests leichter signifikant zu machen. Es kann vorkommen, daß der zweiseitige Test noch keine signifikanten Differenzen nachweisen kann, während dies unter Verwendung des einseitigen Tests sehr wohl der Fall ist (s. a. Kap. 5.2). Freilich, die einseitige Hypothesenprüfung erfordert stets verläßliche Vorinformationen über bereits bestehende Differenzen in einer bestimmten Richtung und kann nur darüber Auskunft geben, ob die Differenzen in der bei der Hypothesenformulierung ausgedrückten Tendenz zufälliger (nichtsignifikanter) Art versus überzufälliger (signifikanter bzw. hochsignifikanter) Art sind.

2.2.5. Durchführung des Experiments

Es ist eine Ermessensfrage, ob man die Stichprobenbildung, das sog. Sampling, noch zur Versuchsplanung oder schon zur Versuchsdurchführung rechnen will. Beide Schritte – Planung und Ausführung – sind in der experimentellen Praxis mehr oder weniger eng miteinander verzahnt. Eine getrennte Behandlung der Planungs- und Ausführungsprobleme erfolgt hier mehr aus didaktischen Gründen und darf keineswegs dahin gehend ausgelegt werden, daß Planungs- und Durchführungsphase beziehungslos aufeinanderfolgen. Vielmehr müssen alle Schritte auch der Versuchsdurchführung bis ins Detail genau geplant und – vorher – festgelegt werden. Die eigentliche Durchführung betrifft somit strenggenommen nur die Realisierung, d.h. den technischen Ablauf, des geplanten Vorhabens. Im weiteren Sinne gehören jedoch zur Durchführung des Experiments Überlegungen zum Design genauso wie das Sampling, die Datenerhebung und die Verrechnung bzw. Datenverarbeitung (vgl. Kap. 2.2.6 unten). Nach der Erörterung essentieller Probleme der Versuchsplanung in den vorausgegangenen Abschnitten wären nun die wichtigsten Stichprobentechniken sowie Methoden der Datenerfassung zu behandeln.

2.2.5.1. Stichprobenbildung (Sampling)

Stichprobenbildung *(Sampling)* und Stichprobenfehler sind entscheidende Voraussetzungen im Hinblick auf die Güte und Zuverlässigkeit statistischer Schlußfolgerungen (vgl. Kap. 5 in diesem Buch). Die meisten Stichprobentechniken lassen sich auf *zwei Grundtypen* zurückführen: die *Zufalls*stichprobe und die *repräsentative* Stichprobe. Aus diesen sind für die praktischen Bedürfnisse zahlreiche Varianten entwickelt worden, die man fast ausnahmslos als *Mischtypen*, d.h. zwischen echter Zufallsstichprobe und Repräsentativstichprobe angesiedelt, betrachten kann. Sämtliche Bezeichnungen beziehen sich auf die Art der Herstellung der Stichprobe, stellen also operationale Definitionskriterien dar.

Zufallsstichprobe
In der Forschungspraxis ist es selten möglich, *alle* Fälle der jeweils interessierenden Vpn-Gruppe *(Population)* zu untersuchen; gewöhnlich ist man hier auf eine relativ kleine Vpn-Auswahl *(Stichprobe)* angewiesen. Dabei stellen sich einige spezielle Probleme, die in der Frage gipfeln: Sind die anhand von Stichproben gewonnenen Untersuchungsergebnisse *(= Statistiken)* repräsentativ im Hinblick auf die Populationsdaten *(= Parameter)*? Die zur Prüfung dieser Frage angesetzten Stichproben- oder Signifikanztests (vgl. Kap. 5) kontrollieren, ob bzw. inwieweit die untersuchte Stichprobe tatsächlich der in Frage stehenden Population entstammt. Die Klärung kann auf zweierlei Wegen vorgenommen werden: Einmal wird mit Hilfe der sog. Nullhypothese eine Entscheidung darüber herbeigeführt, ob die gefundene Statistik zufällig (nicht signifikant) oder überzufällig (signifikant bzw. sehr signifikant) von dem bekannten bzw. theoretischen Parameter abweicht. Zum andern kann man – in Umkehrung des obigen Wahrscheinlichkeitsschlusses – die sog. Vertrauensintervalle bestimmen, d.h. die Bereiche in der Verteilung, in denen der (gesuchte) Parameter mit einer bestimmten Wahrscheinlichkeit liegen muß. Eine ausführlichere Darstellung der skizzierten Probleme findet sich in Kap. 5.2.
Dem geschilderten Vorgehen liegt prinzipiell die Forderung nach *Zufalls*stichproben zugrunde. Darunter versteht man das *zufällige* Herausgreifen von Populationsangehörigen zum Zwecke einer Untersuchung. Wichtigstes Kennzeichen der Zufallsstichprobe ist demnach die Voraussetzung, daß jedes Mitglied der Population die *gleiche Chance* erhält, in die Stichprobe zu gelangen. Diesem Kriterium entspräche etwa die Auswahl durch Losentscheid oder im Lotterieverfahren. Ökonomischer und zugleich objektiver wäre die Bildung einer Zufallsstichprobe mit Hilfe sog. *Zufallszahlen* (vgl. Tab. I im Anhang), weshalb das Ergebnis einer solchen Stichprobenbildung auch als „exakte" (Zufalls-)Stichprobe bezeichnet wird.
Soll beispielsweise eine Stichprobe von N = 90 Viertklaßkindern aus einer entsprechenden Schülerpopulation einer Mittelstadt (N = 467) gebildet werden, so könnte man hier folgendermaßen vorgehen: Zunächst werden alle 467 Schüler (Angehörige der Population) namentlich erfaßt und durchnumeriert. Sodann legt man sich auf irgendeine Stelle innerhalb von Tab. I (s. S. 253f.) als Anfangspunkt fest und bestimmt die Reihenfolge des Ablesens, die beliebig senkrecht (von oben nach unten oder von unten nach oben) versus waagerecht (nach rechts oder links) gewählt werden kann. Für unser Illustrationsbeispiel benötigen wir 90 Vpn aus einer Population von 467 Schülern. Gemäß der Forderung, daß bei einer Zufallsstichprobe jedes Mitglied der Population die gleiche Chance haben muß, in die Stichprobe zu gelangen, wären hier dreistellige Zahlen abzulesen. Nachdem wir den Anfangspunkt bei Reihe (Spalte) 5 und Zeile 10 fixiert und uns auf die senkrechte Reihenfolge von oben nach unten festgelegt haben, können wir mit der eigentlichen Stichprobenbildung beginnen.
Die erste so ermittelte (dreistellige) Zahl lautet 176, d.h., der Schüler mit der Nummer 176 käme als erster in die Stichprobe. Die nächste Zahl in Tab. I (Spalten 5/6/7 und Zeile 11) heißt 880 und wird übergangen, da sie unser Populations-N übersteigt. Die folgende Zahl (Spalten 5/6/7 und Zeile 12) heißt 374, d.h., der Schüler mit der Nummer 374 käme als zweiter in die Stichprobe. Entsprechend wären noch folgende Schüler mit der Nr. 92, 36, 124, 214, 129, 371, 438, 3, 11, 131, 428, 337, 276, 43, 134, 387, 386, 8, 51 (Spalten 8/9/10 und Zeile 1), 401 (Spal-

ten 8/9/10 und Zeile 3) usw. zu berücksichtigen. Ist man am Tabellenende – rechts unten – angelangt, ohne daß das Stichprobenkontingent aufgefüllt ist, so beginnt man wieder an einer anderen Stelle innerhalb der Tab. I – ggf. unter Variierung der Reihenfolge – und fährt so lange fort, bis die Stichprobe vollständig ist. Tritt eine Zufallszahl innerhalb derselben Stichprobenbildung zweimal auf, so wird sie nur das erste Mal berücksichtigt und im folgenden übergangen. Strenggenommen müßte man freilich unter strikter Wahrung der Chancengleichheit einmal gezogene Vpn-Nummern wieder in die Grundgesamtheit (Population) zurückgeben, doch verzichtet man in der Praxis der Stichprobenbildung auf eine entsprechende Rücklage. Eventuell dadurch verursachte Stichprobenfehler sind praktisch irrelevant.

Repräsentative Stichprobe

Häufig läßt sich eine echte Zufallsstichprobe nicht herstellen, etwa bei großem Populations-N oder wenn die Angehörigen der Population regional sehr weit – z. B. über einen Flächenstaat oder die gesamte BRD hin – streuen und somit nur schwer faßbar sind. In solchen Fällen *stratifiziert* oder *organisiert* man eine Stichprobe, d.h., man bildet eine sog. *repräsentative* Stichprobe. Unabdingbare Voraussetzung für die Organisation einer Stichprobe sind allerdings verläßliche Informationen über die *Strata* (Schichten), also die im Hinblick auf den Untersuchungsgegenstand relevanten Populationsmerkmale. Solche Merkmale können die Verteilungsparameter der Geschlechtsvariablen, der sozio-ökonomischen Statusvariablen (Prozentsatz der Akademiker, Beamten, Angestellten, Arbeiter, Bauern usw.), der Altersvariablen, des Bildungsganges bzw. der Schullaufbahn u. dgl. m. betreffen. Grundsätzlich soll die *organisierte Stichprobe ein getreues Abbild der Grundgesamtheit* darstellen, d.h. hinsichtlich der Kontrollmerkmale proportional zur Population zusammengesetzt sein. In diesem Zusammenhang ergibt sich eine Reihe von Problemen, von denen hier nur zwei angesprochen werden:

A	B	C	D	E	F

1. Die Verteilung der relevanten Merkmale (z.B. A bis F) in der Population ist unbekannt. In diesem Fall muß sich der Vl entsprechende Informationen selbst besorgen. Als Unterlagen dazu werden statistische Jahrbücher, Materialien zur Haushaltsberatung im Bereich der Kultusverwaltung, Schulstatistiken u.ä. herangezogen.
2. Oft stellt sich auch die Frage, ob man in bezug auf relevante Merkmale stratifiziert, d.h., ob die kontrollierten Variablen überhaupt Einflußgrößen darstellen. Im Zweifelsfalle sollte man deshalb – schon um unnötigen Aufwand bei der Stichprobenbildung zu vermeiden – diese Frage durch Voruntersuchungen klären. So könnte sich bei der Untersuchung der Schuleignung die Frage nach dem Einfluß der Konfessionsvariablen, etwa unter dem Eindruck disproportionierter Konfessionsanteile bei der Realschul- und Gymnasialpopulation, stellen. Zur Klärung würde man hier jeweils die gleiche Anzahl (z.B. 100) per Zufall ausgewählter evangelischer und katholischer Schüler der 4. Grundschulklasse einem Schuleignungstest unterziehen. Ergäben sich in diesem Vorversuch keine (signifikanten) Differenzen zwischen den Konfessionen bezüglich der Schuleignung, so wäre der

Schluß gerechtfertigt, daß die Konfessionsvariable *keine* Einflußgröße im Hinblick auf den Untersuchungsgegenstand darstellt. Die Verteilung der Konfessionsvariablen müßte demnach – in diesem Falle – nicht kontrolliert, d.h. bei der Stichprobenorganisation nicht beachtet werden. (Davon abgesehen ist die Konfessionszugehörigkeit relevant im Hinblick auf die Schullaufbahnentscheidung [vgl. Heller 1970], die jedoch bei unserem Illustrationsbeispiel nicht gefragt war.)
Innerhalb der einzelnen *Strata* (Schichten) werden dann wiederum die Vpn per Zufall ermittelt. Hinsichtlich des Ergebnisses müßten sich also eine echte Zufallsstichprobe und eine gut organisierte (stratifizierte oder repräsentative) Stichprobe gleichen: Beide Grundtypen repräsentieren – im verkleinerten Maßstab – die Verhältnisse der Population.
Mittenecker (1970, S. 28f.) weist in diesem Zusammenhang noch auf die Möglichkeit hin, nicht von vornherein bei der eigentlichen Stichprobenbildung (Vpn-Auswahl), sondern erst hinterher bei der *Auswertung* der Resultate auf entsprechende Repräsentativität zu achten, etwa durch die Berechnung eines „gewogenen" Gesamtpunktwertes. Dieses Vorgehen stellt freilich gewissermaßen eine Notlösung dar und ist nicht frei von Gefahrenquellen, auf die wir schon aufmerksam machten (s. S. 59f.).

Mischtypen

Wichtiger als die zuletzt angesprochene Variante sind folgende Methoden, die Bestandteile sowohl der Zufallsstichprobe als auch der repräsentativen Stichprobenbildung aufweisen und deshalb als Mischtypen bezeichnet werden können. U.a. sind hier die sog. geschichtete Zufallsauswahl, die Klumpenauswahl und die Quotenstichprobe von Bedeutung.
Bei der *geschichteten Zufallsauswahl* gliedert man die Gesamtpopulation in mehrere Teilpopulationen, aus denen jeweils Zufallsstichproben gezogen werden. Dieses Verfahren empfiehlt sich immer dann, wenn die Grundgesamtheit eine sehr heterogene Merkmalsstruktur aufweist (z.B. Schülerpopulation der Sekundarstufe I). Freilich besteht hier nicht mehr für jedes Mitglied der Gesamtpopulation im engeren Sinne Chancengleichheit; Zufälligkeit der Auswahl ist hierbei nur innerhalb der Teilpopulationen bzw. innerhalb der *Strata* (Schichten) gewährleistet.
Einen Schritt weiter geht man bei der sog. *Klumpenstichprobe*. Nach der Populationsteilung in eine Reihe von Untergruppen wählt man daraus – wiederum per Zufall – *einzelne* (nicht alle) Teilpopulationen aus und faßt diese als Stichproben auf. Diese zufällig ausgewählten „Klumpen" werden dann geschlossen untersucht (= eigentliche Klumpenauswahl) oder bilden erst die Grundlage für die endgültige Stichprobenauswahl, z.B. einzelne aus den Klumpen zufällig ausgewählte Individuen (= mehrstufige Klumpenauswahl).
Schließlich wäre noch die sog. *Quotenstichprobe* zu erwähnen, die auf den ersten Blick mit der oben beschriebenen repräsentativen Stichprobe sub specie Herstellung verwechselt werden könnte. Hierbei wird die Stichprobe nach einem proportional zur Verteilung der Kontrollmerkmale (in der Population) aufgestellten *Quotenplan* zusammengestellt. Der einzige, aber wesentliche Unterschied zwischen der besonders in der Meinungsforschung favorisierten Quotenstichprobe und der organisierten Repräsentativstichprobe besteht darin, daß der Vl (z.B. Interviewer) im ersten Falle sich selbst die Vpn aussuchen darf, während im zweiten Fall die Vpn-Auswahl innerhalb der *Strata* rein zufällig – also unabhängig vom Vl –

getroffen wird. Es liegt auf der Hand, daß bei der Quotenstichprobe leicht subjektive Momente ins Spiel kommen, und zwar auch dann, wenn sich der Vl bezüglich der Vpn-Auswahl strikt an die festgelegten Quoten hält. Praktisch werden solche Bedenken erst bei größeren Stichproben ($N > 1000$ bzw. $N > 2000$) gegenstandslos.

Experimental- und Kontrollgruppe

Stehen aus derselben Population zwei oder mehrere Stichproben unter variierten Versuchsbedingungen zum Vergleich an, so ist die Organisation sog. Experimental- und Kontrollgruppen notwendig. Dazu dienen vor allem folgende Techniken:
1. *Die Matched-Pair Methode.* Hierbei wird jedes Individuum einzeln im Hinblick auf die relevanten Versuchsvariablen bzw. Merkmale untersucht. Anschließend werden immer zwei gleichwertige Vpn („Zwillingspaare") zusammengestellt und auf die Experimental- vs. Kontrollgruppe verteilt. Dies wäre die individuelle Form der Stichprobenparallelisierung.
2. *Die Matched-Group Technique.* Hierbei achtet man nur noch darauf, daß die Gruppen insgesamt gleichwertig sind, d.h. sich hinsichtlich Mittelwert und Streuung nicht unterscheiden. Mit dieser Methode lassen sich äquivalente Stichproben leichter *matchen* als nach der ersten (individuellen) Methode.
3. *Die Randomized-Group Technique.* Bei dieser Methode geht man noch etwas blinder vor. Nach der Zusammenstellung „gleicher" Vpn, d.h. solcher Vpn, die sich unter dem Gesichtspunkt relevanter Merkmale ähneln, werden diese zufällig („blind") auf die Untersuchungs- und Kontrollgruppe verteilt. So kommen etwa alle Vpn mit geraden Nummern in die Experimentalgruppe und alle Vpn mit ungeraden Nummern in die Kontrollgruppe. Auf diese Weise entstehen zwei Zufallsstichproben, d.h. Parallelgruppen, die sich nur zufällig bezüglich der relevanten Einflußvariablen unterscheiden.
Das Modell von Experimental- und Kontrollgruppe ist kein generell gültiges Modell. Wird beispielsweise nur *eine* Dimension (etwa die Auswirkung der Intelligenz auf die Schulleistung) untersucht, so ist eine Kontrollgruppe überflüssig. Siehe ergänzend die früheren Ausführungen auf S. 59ff.
Schließlich wäre hier noch die Unterscheidung von *abhängig* gewonnener und *unabhängig* gewonnener Stichprobe von Bedeutung. Unter *abhängigen* oder *korrelierenden* Stichproben versteht man Datenpaare, die von *identischen* Vpn-Gruppen – aus ein und derselben Stichprobe (etwa bei einem sog. Prätest vs. Posttest erhoben) – stammen. Beispielsweise werden im Rahmen eines Trainingsprogramms zur Begabungsförderung die IQ-Werte der Versuchsgruppe vor *und* nach dem Einsatz der Fördermaßnahmen, also zweimal innerhalb eines bestimmten Zeitintervalls, erfaßt (s. S. 194f.). Ferner werden im engeren Sinne *parallelisierte* Vpn-Gruppen, d.h. gematchte Parallelstichproben (siehe oben), als abhängige oder korrelierende Stichproben bezeichnet.
Hingegen gelten als *unabhängige* Stichproben Vpn-Gruppen, die durch *Randomisierung* (nach dem Zufallsprinzip) aus ein und derselben Population entnommen wurden, wenn also jede Stichprobe unabhängig von der anderen zustande kam. So können beispielsweise die Mütter von Germanistik- und Jurastudenten – unter der Voraussetzung, daß diese nicht gleichzeitig Kinder beider Studienrichtungen aufweisen – als unabhängig gewonnene *Samples* aufgefaßt werden (s. S. 192f.).

Die Unterscheidung in „abhängige“ und „unabhängige“ Stichproben ist vor allem im Hinblick auf den Einsatz diverser Signifikanztests (vgl. Kap. 5.3 u. 5.4) von Bedeutung.

2.2.5.2. Datenerhebung und Auszählung der Rohdaten

In der Regel wird man vor dem eigentlichen Experiment in sog. *Vorversuchen* die äußeren Bedingungen bzw. technischen Voraussetzungen zur Durchführung des Hauptversuchs testen müssen. Dazu gehört das Erfahrungen-Sammeln anhand der vorgesehenen Instruktion, Versuchszeit, Räumlichkeiten u.ä. genauso wie die Erprobung der technischen Einrichtungen zur Protokollierung bzw. Registrierung der Verhaltensdaten, Reaktionsmessung u.ä. Nicht selten werden dann noch einmal mehr oder weniger umfangreiche *Verbesserungen,* z.B. bezüglich der Versuchszeit, des geplanten Kategoriensystems zur Verhaltensdeskription oder auch an den Apparaturen, notwendig. Ganz allgemein wird man sagen können, daß ein Gelingen des Experiments um so eher gewährleistet ist, je gründlicher die Versuchsplanung vorgenommen und desto eindeutiger die Instruktions- und Ausführungsvorschriften definiert sind. Selbstverständlich ist der Vl dann während des eigentlichen Ablaufs der Untersuchung strikt an die Einhaltung der einmal fixierten Bedingungen (Richtlinien) gebunden. Jede weitere Variation würde ja die Objektivität des Experiments und damit die experimentellen Ergebnisse überhaupt gefährden.
Die *Registrierung* beobachteter Verhaltensdaten bildet zweifellos das Herzstück der Versuchs*durchführung,* zu der im weiteren Sinne auch die *Auszählung der Rohdaten* (das sog. *Scoring*) – nicht aber deren statistische Verarbeitung – gerechnet werden muß. Die Erfassung der Dateninformationen birgt vor allem bei größeren Stichproben (z.B. N > 500) und/oder umfangreicheren Datenmengen (d.h. größerer Variablenzahl) eine Reihe von technischen Problemen, zu deren Bewältigung sich zahlreiche Instrumente der elektronischen Datenverarbeitung (EDV-Technik) anbieten. Diese können den Arbeits- und Zeitaufwand und somit die Kosten größerer Forschungsvorhaben erheblich reduzieren und sollten deshalb vor allem im erziehungswissenschaftlichen Bereich stärker als bisher genutzt werden; in der psychologischen, teilweise auch in der soziologischen Forschungspraxis gehört der Einsatz von EDV-Techniken bereits seit Jahren zu den Selbstverständlichkeiten. Studierende der Sozialwissenschaften, die eine empirische Arbeit anstreben, sollten sich deshalb möglichst frühzeitig mit den Grundlagen und technischen Hilfen der EDV vertraut machen.
Im Rahmen der Datenerfassung – Registrierung und *Scoring* – gewinnen die *Lochkarte* bzw. das *Magnetband* als Datenträger sowie die sog. *Markierungsbelege* (elektronisch auswertbare Antwort- oder Protokollbögen) vorrangige Bedeutung. Ihre Verwendung ist denkbar einfach. Anhand eines Illustrationsbeispiels wollen wir zunächst die einzelnen Schritte zur Erstellung der Lochkarte und ihre Verwendungsmöglichkeiten beschreiben. Anschließend werden wir auf die elektronische Auswertung von Markierungsbelegen zu sprechen kommen.
Das abgebildete Lochkartenbeispiel mit eingestanzten Dateninformationen (▮) – aus einer größeren Begabungsuntersuchung d. Verf. von 1968 – läßt bereits die wichtigsten Funktionsmerkmale dieses Datenträgers erkennen:

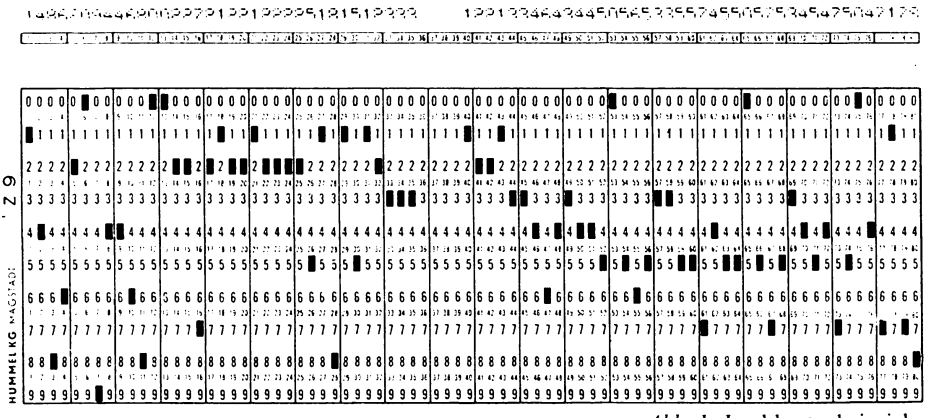


Abb. 1: Lochkartenbeispiel

1. Die genormte Lochkarte (nach dem deutschamerikanischen Erfinder der Lochkartenmaschine auch Hollerithkarte genannt) besteht aus 80 durchnumerierten Spalten.
2. Jede Spalte enthält 10 gekennzeichnete Lochungspositionen (von oben nach unten: 0, 1, 2, 3, 4, 5, 6, 7, 8 und 9) sowie die nicht gekennzeichneten Positionszeilen 12 und 11 (über der 0-Zeile oben). Insgesamt stehen somit $10 \cdot 8 = 800$ bzw. $12 \cdot 8 = 960$ Positionen oder Registriereinheiten pro Karte zur Verfügung.
3. In der Regel wird für jede Vp eine eigene Datenkarte angelegt. Sofern die Kapazität der Positionen nicht ausreicht, können für eine Vp auch mehrere – prinzipiell beliebig viele – Lochkarten, die dann durch Buchstabenmarkierungen (z. B. A, B, C usw.) oder eine andere Kennzeichnung identifiziert werden, reserviert werden.
4. Für die Registrierung der Informationsdaten kann man a) *Ziffern*symbole (Lochung in den betr. Positionen von 0 bis 9); b) *Buchstaben*symbole (jeweils zwei Lochungen in derselben Spalte, z. B. für Buchst. A Lochungen in Pos. 12 und 1, für Buchst. B Lochungen in Pos. 12 und 2, für Buchst. C Lochungen in Pos. 12 und 3 oder für Buchst. J Lochungen in Pos. 11 und 1, für Buchst. K Lochungen in Pos. 11 und 2 usw., für Buchst. S Lochungen in Pos. 0 und 2..., für Buchst. Z Lochungen in Pos. 0 und 9); c) *andere Zeichen* (z. B. für Symbol "$_0$ Lochungen in Pos. 0, 4 und 8) verwenden.
5. Die einfachen Lochungsmaschinen sind mit numerischer oder alphabetisch-numerischer Tastatur (ähnlich wie bei einer Schreibmaschine) ausgestattet. Manche Kartenlocher enthalten zusätzlich einen Lochschriftübersetzer, der dann den Aufdruck des Klartextes (vgl. Abbildung) als Orientierungshilfe gestattet.
6. Eventuelle Lochungsfehler werden im sog. Prüflocher ermittelt. Dazu werden die Originalkarte und die Kontrollkarte (unabhängig von der Originalkarte gelochte Datenkarte) miteinander verglichen.

Bevor nun die Dateninformationen abgelocht werden können, muß zunächst ein Plan für die Kodierung, d.h. Übertragung der Ausgangsdaten in das jeweils verwendete Symbolsystem (gewöhnlich Ziffernsystem), erstellt werden. Unser nachstehendes Beispiel eines solchen Planes bezieht sich wiederum auf die oben zitierte Begabungsuntersuchung des Jahres 1968; derselbe Plan hatte sich bereits einige Jahre zuvor (in variierter Form) bei Untersuchungen zur Begabungsreservenermittlung (vgl. Heller 1970) bewährt. Der hier wiedergegebene *Kodierungsplan* wurde gegenüber dem Originalplan, der sich auf die Kapazität zweier Lochkarten (160 Spalten) pro Vp bezog, gekürzt. Auf die Explikation des Vorgehens dürfte die geraffte Darstellung jedoch ohne Einfluß sein.

Kodierungsplan

Variable (Merkmal usw.)	Kennziffer (= Position)	Nr. der Lochkartenspalte
Schüler:	Pb-Nr. (4stellige Zahl)	1 + 2 + 3 + 4
Geschlecht:	männlich = 1	5
	weiblich = 2	
Alter:	jünger als 8; 0 = 00	6 + 7
	8; 0– 8; 2 = 01	
	8; 3– 8; 5 = 02	
	8; 6– 8; 8 = 03	
	8; 9– 8; 11 = 04	
	9; 0– 9; 2 = 05	
	9; 3– 9; 5 = 06	
	9; 6– 9; 8 = 07	
	9; 9– 9; 11 = 08	
	10; 0–10; 2 = 09	
	10; 3–10; 5 = 10	
	10; 6–10; 8 = 11	
	10; 9–10; 11 = 12	
	11; 0–11; 2 = 13	
	11; 3–11; 5 = 14	
	⋮ ⋮ ⋮	
	21; 0 und älter = 53	
Schulart:	Grundschule (1-klassig) = 1	8
	Grundschule (2-klassig) = 2	
	Grundschule (3-klassig) = 3	
	Grundschule (4-klassig) = 4	
	Realschule = 5	
	altsprachl. Gymnasium = 6	
	neusprachl. Gymnasium = 7	
	math./nat. Gymnasium = 8	
Schuljahr/Klasse:	Klasse 4 = 4	9
	Klasse 5 = 5	
	Klasse 6 = 6	
	Klasse 7 = 7	
	Klasse 8 = 8	
	Klasse 9 = 9	
	Klasse 10 = 0	
	Klasse 11 = 1	
	Klasse 12 = 2	
	Klasse 13 = 3	
Schulort:	Postleitzahl (4stellige Zahl)	10 + 11 + 12 + 13
Konfession:	evangelisch = 1	14
	katholisch = 2	
	sonstige Konfession = 3	
	konfessionslos = 4	

Variable (Merkmal usw.)	Kennziffer (= Position)	Nr. der Lochkartenspalte
Einschulungsjahr Zurückstellung Repetition:	1963 eingeschult = 1	15
	1964 eingeschult = 2	
	1965 eingeschult = 3	
	1966 eingeschult = 4	
	1 mal zurückgestellt = 5	
	2 mal zurückgestellt = 6	
	1 mal repetiert = 7	
	2 mal repetiert = 8	
Liebstes Schulfach: (Bei Mehrfachbenennungen nur das erstgenannte Schulfach verlisten!)	Aufsatz (Deutsch) = 1	16
	Rechtschreiben = 2	
	Lesen = 3	
	Rechnen (Mathematik) = 4	
	Heimatkunde = 5	
	Religionslehre = 6	
	Sport/Spiel = 7	
	Zeichnen/Werken/Handarb. = 8	
	Musik = 9	
	Fremdsprache(n) = 0	
Berufswunsch des Pb:	(S. Berufskategorien in Sp. 22 + 23)	17 + 18
Wievieltes Kind?	1. (ältestes) Kind = 1	19
	2. Kind = 2	
	3. Kind = 3	
	4. Kind = 4	
	5. Kind = 5	
	6. Kind = 6	
	7. Kind = 7	
	8. Kind = 8	
	9. Kind = 9	
	ab 10. Kind = 0	
Zahl der Geschwister:	0 Geschwister = 0	20
	1 Geschwister = 1	
	2 Geschwister = 2	
	3 Geschwister = 3	
	4 Geschwister = 4	
	5 Geschwister = 5	
	6 Geschwister = 6	
	7 Geschwister = 7	
	8 Geschwister = 8	
	9 und mehr Geschw. = 9	
Zahl der Geschwister in Realschule und/oder Gymnasium:	0 Geschwister = 0	21
	1 Geschwister = 1	
	2 Geschwister = 2	
	3 Geschwister = 3	
	(usw.; wie oben)	
Erlernter Beruf des Vaters:	Akademiker (ohne Spezif.) = 11	22 + 23
	freiberuflich tätige Akademiker (z. B. Arzt, Rechtsanwalt usw.) = 12	
	Akademiker im Angest.- oder Beamtenverhältnis (Dipl.-Kaufm., Studienrat) = 13	

Variable (Merkmal usw.)	Kennziffer (= Position)		Nr. der Lochkartenspalte
Erlernter Beruf des Vaters: (Forts.)	Großkaufleute u. Großunternehmer	= 14	22 + 23 (Forts.)
	Gutsbesitzer	= 15	
	Fabrikbesitzer	= 16	
	Fabrikdirektoren	= 17	
	Volksschullehrer	= 21	
	Fachschulingenieure (ohne Diplom)	= 22	
	Stabsoffiziere (ab Major)	= 23	
	Journalisten	= 24	
	gehob. Beamte (Inspektor, Amtmann); gehob. Angestelltenberufe (Prokur.)	= 25	
	freiberuflich tätige Nichtakademiker (Kunstmaler, Privatmusiklehrer usw.)	= 26	
	mittl. selbst. Gewerbetreibende (z. B. Lebensmittelhändler, Gastwirte)	= 31	
	Bauern	= 32	
	mittlere Beamte (Sekret.)	= 33	
	selbst. Handwerker und Handwerksmeister	= 34	
	mittl. Angest. (Handlungsbevollmächtigter, Unteroffizier, Vertreter usw.)	= 35	
	gehobene techn, Berufe (z. B. techn. Zeichner, Werkmeister, Polier, Konstrukteur, Techniker)	= 36	
	Offiziere (bis Hauptmann)	= 37	
	Industriefacharbeiter, Handwerksgesellen	= 41	
	einf. Angestellte (Ratschreiber, kfm. Angestellter usw.)	= 42	
	einf. Beamte (z. B. Postschaffner)	= 43	
	angelernte industrielle Tätigkeiten	= 44	
	angelernte, nichttechn. Tätigkeiten (Kellner, Busschaffner, Verkäufer, Krankenpfleger, Kraftfahrer usw.)	= 51	
	ungelernte Arbeiter und Landarbeiter	= 61	
	Hausierer	= 62	
	Rentner	= 71	
	Hausfrau	= 81	
	Sonstige	= 91	
Ausgeübter Beruf des Vaters:	(s. Berufskategorien in Sp. 22 + 23)		24 + 25
Erlernter Beruf der Mutter:	(s. Berufskategorien in Sp. 22 + 23)		26 + 27
Ausgeübter Beruf der Mutter:	(s. Berufskategorien in Sp. 22 + 23)		28 + 29
Schulbildung des Vaters:	Volksschule	= 1	30
	Realschule	= 2	
	Mittlere Reife (R)	= 3	
	Gymnasium	= 4	
	Mittlere Reife (G)	= 5	
	Abitur (G)	= 6	
	Päd. Hochschule	= 7	
	Universität	= 8	
		= 9	
		= 0	

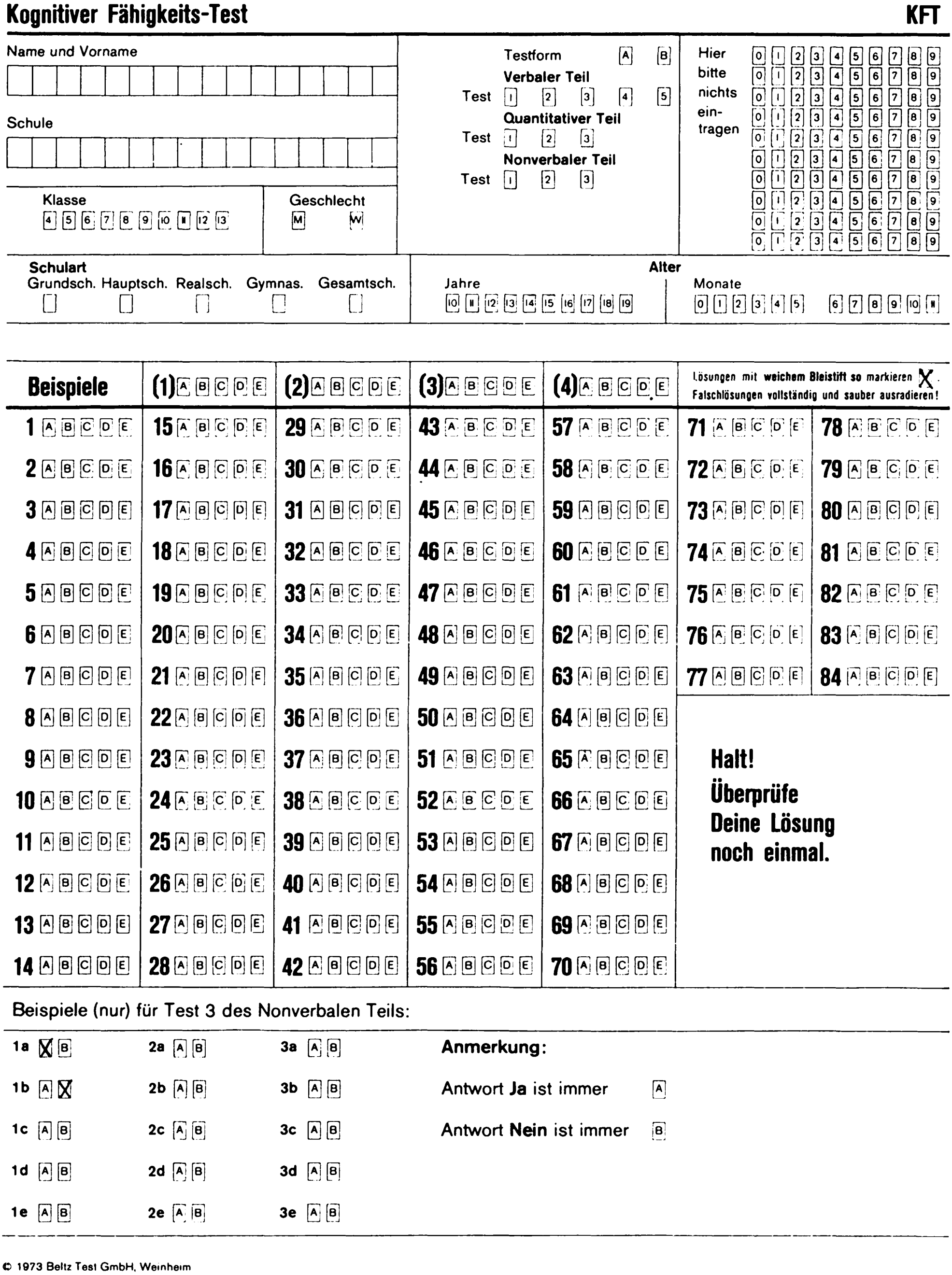
Variable (Merkmal usw.)	Kennziffer (= Position)	Nr. der Lochkartenspalte
Schulbildung der Mutter:	(s. Schulbildungskategorien des Vaters in Sp. 30)	31
Bildungswunsch der Eltern:	Übertritt nach der 4. Klasse:	32
	in die Hauptschule = 0	
	in die Realschule = 1	
	in das Gymnasium (ohne Spez.) = 2	
	in das altspr. Gymnasium = 3	
	in das neuspr. Gymnasium = 4	
	in das math./nat. Gymnasium = 5	
	Übertritt nach der 5. Klasse:	
	in die Hauptschule = 6	
	in die Realschule = 7	
	in das Gymnasium (ohne Spez.) = 8	
	in sonstige Bildungseinrichtg. = 9	
Zeugniszensuren:	Aufsatz (Deutsch) 1 = 1	33
	2 = 2	
	3 = 3	
	4 = 4	
	5 = 5	
	6 = 6	
	Rechtschreiben (Zensuren wie oben)	34
	Rechnen/Mathem. (Zensuren wie oben)	35
	Englisch (Zensuren wie oben)	36
	Französisch (Zensuren wie oben)	37
	Latein (Zensuren wie oben)	38
	Griechisch (Zensuren wie oben)	39
Lehrerurteil (LU) über Arbeitshaltung d. Pb:	sehr anstrengungsbereit = 1	40
	anstrengungsbereit = 2	
	durchschnittl. anstrengungsber. = 3	
	gering anstrengungsbereit = 4	
	sehr gering anstrengungsbereit = 5	
LU über Konzentration:	gut konzentriert = 1	41
	durchschnittl. konzentriert = 2	
	wenig konzentriert = 3	
LU über Selbständigkeit:	sehr selbständig = 1	42
	selbständig = 2	
	wenig selbständig = 3	
LU über Sozialverhalten:	sehr einordnungsbereit = 1	43
	einordnungsbereit = 2	
	wenig einordnungsbereit = 3	
Testergebnisse im PSB:	Subtests 1 + 2 (T-Werte)	44 + 45
	Subtest 3 (T-Werte)	46 + 47
	Subtest 4 (T-Werte)	48 + 49
	Subtests 3 + 4 (T-Werte)	50 + 51
	Subtest 5 (T-Werte)	52 + 53
	Subtest 6 (T-Werte)	54 + 55
	Subtests 5 + 6 (T-Werte)	56 + 57

Variable (Merkmal usw.)	Kennziffer (= Position)		Nr. der Lochkartenspalte
Testergebnisse im PSB: (Forts.)	Subtest 7 (T-Werte)		58 + 59
	Subtest 8 (T-Werte)		60 + 61
	Subtests 7 + 8 (T-Werte)		62 + 63
	Subtest 9 (T-Werte)		64 + 65
	Subtest 10 (T-Werte)		66 + 67
	Subtests 9 + 10 (T-Werte)		68 + 69
	Gesamtleistung (T-Werte)		70 + 71
Testergebnisse im AzN 4+:	Gesamtleistung (T-Werte)		72 + 73
Testergebnisse im CFT 2:	Gesamtleistung (T-Werte)		74 + 75
Testergebnisse im WST:	Gesamtleistung (T-Werte)		76 + 77
Schullaufbahnempfehlung der Grundschule (LU):	Gymnasium geeignet	= 1	78
	Gymn. eingeschränkt geeignet	= 2	
	Realschule geeignet	= 3	
	Real. eingeschränkt geeignet	= 4	
	Hauptschule A-Kurs	= 5	
	Hauptschule B-Kurs	= 6	
	Sonderschule (Lernbehinderte)	= 7	
	kein Vorschlag	= 8	
Schullaufbahnempfehlung des Bildungsberaters:	GG (Gymnasium gut geeignet)	= 1	79
	GB (Gymn. bedingt geeignet)	= 2	
	AG (Aufbaugymn. geeignet)	= 3	
	RG (Realschule gut geeignet)	= 4	
	RB (Realsch. bedingt geeignet)	= 5	
	RE (Realschul-Entwicklungsfall)	= 6	
	HA (geeignet f. Hauptschule A)	= 7	
	HB (geeignet f. Hauptschule B)	= 8	
	SO (Sonderschulverdacht: Lb)	= 9	
	NK (nicht kategorisierbar)	= 0	
Vollzogener Übergang im August 1968:	in Hauptschule A-Kurs	= 1	80
	in Hauptschule B-Kurs	= 2	
	in Hauptschule (ohne Spez.)	= 3	
	in Sonderschule (Lernbeh.)	= 4	
	in Realschule	= 5	
	in altspr. Gymnasium	= 6	
	in neuspr. Gymnasium	= 7	
	in math./nat. Gymnasium	= 8	
	in Gymnasium (ohne Spez.)	= 9	
	unbekannt	= 0	

Die anhand dieses oder eines ähnlichen Schlüssels kodierten Daten werden mit Hilfe eines Kartenlochers abgelocht und stehen dann für die weitere Verarbeitung bereit. So lassen sich die kodierten Daten bzw. Lochkarten mittels Sortier- und einfachen Zählmaschinen – selbstverständlich auch mit dem Computer – nach den verschiedensten Gesichtspunkten gruppieren, auszählen usw. Damit hätten wir freilich schon den ersten Schritt zur eigentlichen Daten*auswertung* (Verrechnung) getan, wovon in den folgenden Kapiteln ausführlich die Rede sein wird. Hier sei

nur noch angemerkt, daß man die Lochkarten jederzeit mit Hilfe sog. Kartendoppler beliebig häufig kopieren kann, so daß dann mehrere (identische) Daten-Kartensätze für Verrechnungszwecke u.ä. bereitstehen. Auch kann man die Lochkartendaten auf die platzsparenden Magnetbänder übertragen.
Der Kodierungsvorgang ist bei umfangreicherem Datenmaterial relativ arbeits- und zeitaufwendig. Dies gilt in weit größerem Maße noch im Hinblick auf die eigentliche Rohdatengewinnung, z.B. die Auswertung von Fragebögen, Interviews, Tests u.ä. In anglo-amerikanischen Ländern werden deshalb schon seit geraumer Zeit die Beobachtungsdaten, Testreaktionen usw. unmittelbar auf elektronisch auswertbare *Markierungsbelege* eingetragen. Registrierungen dieser Art können unter Einsatz sog. Belegleser oder *Test Scoring Machines* automatisch ausgewertet und die entsprechenden Daten (Markierungen in Form von Lösungen, Fehlern, Antwortreaktionen u.ä.) entweder auf Lochkarten oder Magnetband übertragen (*off-line*-Verfahren) oder unmittelbar der Computerverrechnung zugeführt werden (*on-line*-Verfahren). Der technisch interessierte Leser mag sich selbst anhand der von den einschlägigen Firmen bzw. Rechenzentren (z.B. IBM, NCS, MRC, DRZ, ZPL[1] u.a.) herausgegebenen und auf Anforderung zur Verfügung gestellten Informationsbroschüren über die Funktionsmerkmale automatischer (Test-)Auswertungsmaschinen sowie die neueste Entwicklung auf diesem Gebiet auf dem laufenden halten; siehe auch de Landsheere 1969, S. 243ff.
Nachstehend bringen wir eine stark verkleinerte Abbildung eines solchen Markierungsbeleges, wie er in dem v. Verf. bearbeiteten Kognitiven Fähigkeits-Test (KFT 4–13) als Schüler-Antwortbogen Verwendung findet. In dieser oder einer ähnlichen Form sind alle elektronisch auswertbaren Belegformulare gestaltet. Mit ihrem Einsatz spart man sich nicht nur mühevolle Auszählarbeit, auch die Zahl der Auswertungsfehler läßt sich auf diese Weise auf ein Minimum reduzieren. Viele der heutigen Belegleser sind bereits technisch soweit ausgereift, daß ihre Verwendung auch im schulischen Bereich ohne Bedenken empfohlen werden kann.
Mit der bloßen Auszählung der Rohdaten bzw. der Übertragung der wie auch immer protokollierten oder registrierten Informationen auf die Datenträger (Lochkarte oder Magnetband) ist der Vorgang der Datenerhebung praktisch abgeschlossen. Mehr oder weniger unmittelbar schließt sich nun daran der nächste Schritt an: die Datenverarbeitung oder Datenverrechnung. Damit werden wir uns in den folgenden Kapiteln sowie ausführlicher im statistischen Teil (s. S. 93ff.) befassen. Es kann nicht Aufgabe dieses Lehrbuches sein, zusätzlich eine Einführung in die EDV-Techniken zu bieten. Wer sich der Hilfe elektronischer (Groß-)Rechenanlagen bedienen möchte oder muß, sollte nach Möglichkeit an einem (in der Regel 14 Tage dauernden) Programmierkurs für Anfänger oder ggf. für Fortgeschrittene teilnehmen. Solche Kurse werden regelmäßig von jedem größeren Rechenzentrum, z.B. dem Deutschen Rechenzentrum (DRZ) in Darmstadt oder den hochschulinternen bzw. -angeschlossenen Rechenzentren, durchgeführt und stehen gegen Entrichtung einer geringen Unkostengebühr grundsätzlich allen Studierenden offen. Darüber hinaus wird man sich von Fall zu Fall von den am betr. Rechenzentrum angestellten Mathematikern bzw. Fachstatistikern beraten lassen können.

1 IBM = Intern. Büro-Maschinen GmbH (IBM Deutschland: Hauptsitz in Sindelfingen); NCS = National Computer Systems (Hauptsitz: New York); MRC = Measurement Research Center in Iowa City; DRZ = Deutsches Rechenzentrum in Darmstadt; ZPL = Zentrum für Personal- und Lehrsysteme in Stuttgart.


Kognitiver Fähigkeits-Test **KFT**

Name und Vorname

Schule

Klasse 4 5 6 7 8 9 10 11 12 13

Geschlecht M W

Testform A B

Verbaler Teil
Test 1 2 3 4 5

Quantitativer Teil
Test 1 2 3

Nonverbaler Teil
Test 1 2 3

Hier bitte nichts eintragen 0 1 2 3 4 5 6 7 8 9

Schulart
Grundsch. Hauptsch. Realsch. Gymnas. Gesamtsch.

Alter
Jahre 10 11 12 13 14 15 16 17 18 19
Monate 0 1 2 3 4 5 6 7 8 9 10 11

Beispiele	(1) A B C D E	(2) A B C D E	(3) A B C D E	(4) A B C D E	Lösungen mit weichem Bleistift so markieren X. Falschlösungen vollständig und sauber ausradieren!	
1 A B C D E	15 A B C D E	29 A B C D E	43 A B C D E	57 A B C D E	71 A B C D E	78 A B C D E
2 A B C D E	16 A B C D E	30 A B C D E	44 A B C D E	58 A B C D E	72 A B C D E	79 A B C D E
3 A B C D E	17 A B C D E	31 A B C D E	45 A B C D E	59 A B C D E	73 A B C D E	80 A B C D E
4 A B C D E	18 A B C D E	32 A B C D E	46 A B C D E	60 A B C D E	74 A B C D E	81 A B C D E
5 A B C D E	19 A B C D E	33 A B C D E	47 A B C D E	61 A B C D E	75 A B C D E	82 A B C D E
6 A B C D E	20 A B C D E	34 A B C D E	48 A B C D E	62 A B C D E	76 A B C D E	83 A B C D E
7 A B C D E	21 A B C D E	35 A B C D E	49 A B C D E	63 A B C D E	77 A B C D E	84 A B C D E
8 A B C D E	22 A B C D E	36 A B C D E	50 A B C D E	64 A B C D E		
9 A B C D E	23 A B C D E	37 A B C D E	51 A B C D E	65 A B C D E	**Halt! Überprüfe Deine Lösung noch einmal.**	
10 A B C D E	24 A B C D E	38 A B C D E	52 A B C D E	66 A B C D E		
11 A B C D E	25 A B C D E	39 A B C D E	53 A B C D E	67 A B C D E		
12 A B C D E	26 A B C D E	40 A B C D E	54 A B C D E	68 A B C D E		
13 A B C D E	27 A B C D E	41 A B C D E	55 A B C D E	69 A B C D E		
14 A B C D E	28 A B C D E	42 A B C D E	56 A B C D E	70 A B C D E		

Beispiele (nur) für Test 3 des Nonverbalen Teils:

1a X B	2a A B	3a A B	**Anmerkung:**
1b A X	2b A B	3b A B	Antwort **Ja** ist immer A
1c A B	2c A B	3c A B	Antwort **Nein** ist immer B
1d A B	2d A B	3d A B	
1e A B	2e A B	3e A B	

Abb. 2: Muster eines automatisch auswertbaren Test-Antwortbogens (stark verkleinert)

2.2.6. Datenverarbeitung und Interpretation der Ergebnisse

Bevor wir auf die Probleme und Methoden der Quantifizierung im einzelnen zu sprechen kommen, möge eine knappe Einführung in die Meßtheorie die allgemeinen Grundlagen für statistische Anwendungsmodelle vorbereiten und entsprechende Probleme einsichtig machen. Im Zentrum unserer theoretischen Erörterungen werden Begriff und Formen des Messens stehen. Deren Kenntnis ist von unmittelbarer Bedeutung für die statistischen Verfahren zur Datenverarbeitung.

2.2.6.1. Zum Begriff des Messens

Der Begriff „*Messen*" stammt aus der Physik. Allgemein versteht man darunter den *Vergleich mit einem vorliegenden Maßstab*. Dabei werden Zahlen zugeordnet, so daß mit diesen nach bestimmten Regeln bestimmte Operationen durchgeführt werden können, die zu *neuen Informationen* über die Beobachtungsdaten führen (nach Süllwold). Als klassische Kriterien des physikalischen Meßbegriffs gelten folgende: 1. festgelegte Einheiten mit 2. gleichen (konstanten) Intervallen und 3. mit absolutem Nullpunkt.

Nun gibt es aber physikalische Skalen ohne einen absoluten Nullpunkt: die Temperaturskalen. Hierbei ist der Nullpunkt willkürlich – auf den Gefrierpunkt des Wassers bezogen – gesetzt. Auch in den Sozialwissenschaften, z.B. der Testpsychologie, arbeiten wir mit Meßskalen ohne absoluten Nullpunkt (vgl. Heller 1973). Weiterhin könnten wir fragen, ob man auch dann noch von „Messen" sprechen kann, wenn zusätzlich das zweite Kriterium (Intervallkonstanz) entfällt. Kann eine solche Skala noch verwertbare Informationen geben? Um diese oder ähnliche Fragen beantworten zu können, müssen wir zunächst auf den Skalenbegriff eingehen.

2.2.6.2. Meßskalen

Unter einer „*Skala*" versteht man einen *Bereich, innerhalb dessen Meßergebnisse* der verschiedensten Art *variieren*. Sie dient allgemein als Vergleichsbasis oder Maßstab für die Meßresultate. In den Sozialwissenschaften müssen wir nun verschiedene Ebenen des Messens, d.h. vier Skalentypen unterscheiden:

1. *Nominal-* oder *Klassifikationsskala* (= unterste Ebene des Messens). Dieser Skalentyp genügt nur einem einzigen Kriterium, dem der *Äquivalenz*. Wenn wir beispielsweise alle männlichen Schüler mit der Ziffer 1 und alle weiblichen Schüler mit der Ziffer 2 verschlüsseln (kodifizieren), dann liegt diesem Vorgang bereits eine Nominalskala zugrunde. Analog handelt es sich bei der Klassifizierung in einzelne Berufstätigkeiten, Schularten, Bildungsempfehlungen usw. oder auch bei den Rückennummern der Fußballspieler, der Autokennzeichen, der Postleitzahlen, der Haarfarben usw. um Meßvorgänge auf dem Niveau einer Nominal- oder Klassifikationsskala. Äquivalenz kann also in bezug auf vielerlei Merkmale bestehen. Mit Hilfe der Nominal- oder Klassifikationsskala werden praktisch nur Positionen unterschieden, die bezüglich eines bestimmten Merkmals gleich oder

äquivalent sind, sonst aber keine Beziehungen untereinander aufweisen. So kann man nicht sagen, die Spieler mit der Nummer 3 seien besser als die mit der Nummer 6 oder die Autos mit den Kennbuchstaben HD seien schneller (besser, größer usw.) als die Autos mit den Kennbuchstaben BN. Wohl aber läßt sich ausmachen, daß die Spieler mit der Rückennummer 3 dieselbe Position im Fußballspiel innehaben, alle Autos mit den Kennbuchstaben HD aus dem Heidelberger Bezirk kommen usw. Auf der Grundlage von Nominalskalen lassen sich demnach (nur) folgende mathematisch-statistische Operationen ausführen: Zählvorgänge sowie Ermittlung von Häufigkeitsverteilungen einschließlich deren inferenzstatistische Kontrolle via Chi^2-Test bzw. Punkt-Vierfelderkorrelation oder Kontingenzbestimmung (s. S. 146ff. u. 209ff.).

2. *Ordinal-* oder *Rangskala* (= zweite Ebene des Messens). Während die Nominalskala eine bloße Kategorienskala darstellt, d.h. eine Zuordnung nach Begriffen (nomina) oder Klassen (Kategorien) gestattet, erlaubt die Ordinalskala darüber hinaus eine *Rang*ordnung. Mit der Rangreihenbildung, d.h. der Bestimmung eines „Mehr" oder „Weniger", sind nun Aussagen über die Beziehung zwischen den Klassen oder Kategorien möglich, z.B. im Sinne des „größer als", „besser als", „lebhafter als" usw. Freilich ist damit (noch) nichts über die Größe des Unterschiedes gesagt. Die Rang- oder Ordinalskala genügt somit zwei Kriterien: a) dem Kriterium der *Äquivalenz* und b) dem Kriterium der *Beziehung zwischen den Klassen*, z.B. der Beziehung „größer als" versus „kleiner als".

Viele Schul-(Leistungs-)Tests verwenden (Prozent-)Rangskalen als Normenskalen (vgl. Heller 1973). Auch die meisten Lehrerurteile sind Einschätzungen hinsichtlich „besser als" oder „schlechter als". Die Schüler werden rangmäßig eingestuft, ohne daß damit angegeben wäre, wie groß der Abstand zwischen den einzelnen Rangpositionen tatsächlich ist. So sind die Schüler mit der Note 2 besser als die Schüler mit der Note 3 oder die Schüler mit der Note 6 schlechter als die Schüler mit der Note 5 – innerhalb desselben Klassenbezugssystems, d.h. in bezug auf dieselbe Schulklasse. Es ist jedoch völlig offen, ob die beiden zum Vergleich anstehenden Schülerpaare oder -gruppen konstante Leistungsabstände aufweisen. Der Leistungsunterschied zwischen Note 2 und 3 ist nicht ohne weiteres der gleiche wie der zwischen Note 5 und 6. Entsprechende mathematisch-statistische Operationen wären hier die Berechnung von Median, Quartilen, Rangkorrelationen usw. (vgl. Kap. 3.3.2, 3.4.3 u. 3.5.1 sowie S. 137ff.).

3. *Intervallskala* (= dritte Ebene des Messens). Waren die ersten beiden Skalentypen mehr ordnend und qualifizierend, so liegt nun mit der Intervallskala die erste im engeren Sinne quantifizierende Skala vor. Als weitere Voraussetzung kommt hier die *Intervallkonstanz* hinzu, so daß diese Skala folgenden drei Kriterien genügt: a) dem Kriterium der Äquivalenz, b) dem der Beziehung im Sinne des „größer als", c) dem der Gleichheit der Intervalle. Der Nullpunkt ist hier nur konventionell festgelegt. Die Intervallskala kann somit als eine Skala im Sinne der genannten Kriterien, jedoch unabhängig von Maßeinheit und absolutem Nullpunkt, definiert werden. Neben der bereits erwähnten Thermometerskala sind hier zahlreiche Normenskalen vollstandardisierter Tests einschlägig, z.B. IQ-, Z-, T-, C- oder z-Skala.

Auf Intervallskalenniveau können praktisch alle mathematisch-statistischen Operationen ausgeführt werden, ausgenommen Proportionen (Bildung von Quotienten). Die Berechnung von arithmetischem Mittel, Standardabweichung u.ä. ist genauso legitim wie die Produkt-Moment-Korrelation und ihre zahlreichen Varianten sowie die in Kap. 5.3 aufgeführten Signifikanztests.

4. *Verhältnis-* oder *Rationalskala* (oberste Ebene des Messens). Diese Skala genügt nunmehr allen vier Kriterien: a) der Äquivalenz, b) der Beziehung, c) der Intervallkonstanz und d) dem *absoluten Nullpunkt*. Willkürlich ist nur noch die Maßeinheit, z. B. cm, g oder sec (c-g-s-System). Somit sind alle mathematisch-statistischen Operationen auf Rationalskalenniveau möglich, auch Quotientenbildung. Während es unsinnig wäre, zu sagen, daß 100 IQ doppelte Intelligenz im Vergleich zu 50 IQ anzeige oder daß 15 Grad Celsius halb so warm sei wie 30 Grad Celsius, können wir sehr wohl aussagen, daß 60 cm doppelt so lang ist wie 30 cm oder daß 50 kg halbes Gewicht im Vergleich zu 100 kg bedeuten. Praktisch spielen Verhältnisskalen im Bereich der Sozialwissenschaften fast nur in der Psychophysik eine Rolle, theoretisch immer dann, wenn das *Verhältnis* zweier Skalenpositionen *konstant* ist, d. h. unabhängig von der Maßeinheit.

Für die empirische Forschung in der Pädagogik bzw. den Sozialwissenschaften ist die Unterscheidung der verschiedenen Meßniveaus oder Skalentypen sehr wichtig, vor allem im Hinblick auf den Einsatz mathematisch-statistischer Operationen und Tests. Die größere Bedeutung kommt hierbei den ersten drei Meßarten (Nominal-, Ordinal- und Intervallskala) zu.

Schließlich wäre noch die Frage nach der Transformierbarkeit der Skalentypen zu stellen: Wie kann man Skalen transformieren, ohne daß ihre Funktionen verlorengehen? Dies ist möglich bei der:

Rationalskala durch Multiplikation mit einer Konstanten,
Intervallskala durch Addition einer Konstanten,
Ordinalskala durch Quadrieren und Potenzieren mit 3,
Nominalskala durch völliges Permutieren (vgl. Bartel u. a. 1971 oder Traxel 1973).

Grundsätzlich läßt sich ein höheres Skalenniveau auf ein niedrigeres zurückführen, während ein niedrigeres Meßniveau nur gelegentlich und bedingt erhöht werden darf, z. B. Rangnormalisierung (vgl. Kap. 3.5.4). Es gibt auch statistische Verfahren, die simultan unterschiedliche Skalenniveaus als Voraussetzung gestatten, so bei der Zweizeilenkorrelation (s. S. 133ff.).

Die im zweiten Teil dieses Buches dargestellten *statistischen* Methoden und Entscheidungsmodelle sind nach dem Prinzip der Skalenabhängigkeit geordnet: Das *Skalenniveau* erhobener Merkmalsdaten bzw. gesammelter Informationen ist u. E. *das* meßtheoretische *Kriterium* für die Wahl der jeweils einschlägigen Rechenoperation bzw. die Verwendung des adäquaten statistischen Modells, d. h. deskriptiver oder korrelativer Methoden versus stichprobenstatistischer Entscheidungstechniken. Diesem Ordnungsgesichtspunkt folgen die meisten statistischen Lehrbücher, so Siegel (1956), Fröhlich & Becker (1971) u. a. Freilich sind auch andere Angemessenheits-Kriterien denkbar und neuerdings in Diskussion, ohne daß wir im Rahmen unserer Einführung auf entsprechende Konzepte näher eingehen können (vgl. z. B. Selg & Bauer 1971, S. 100ff.).

2.2.6.3. *Quantifizierung qualitativer Variablen*

Wenn bisher vom Quantifizieren (Messen) die Rede war, so bezog sich dieser Begriff meistens auf *quantitative* Variablen, z.B. der UV und AV im Experiment. Sehr oft haben wir es aber in den Sozialwissenschaften nicht mit quantitativen Variablen i.e.S. zu tun, sondern mit sog. *qualitativen* Variablen (z.B. Haarfarbe, Augenfarbe, Geschlecht, Milieu, Stimmungen, Affekten usw.). Dabei hört man oft die Meinung, qualitative Variablen seien nicht quantifizierbar. Ja, manche Vertreter der Pädagogik oder Soziologie, seltener der modernen Psychologie, gehen in ihrer Polemik gegen das Messen in den Sozialwissenschaften sogar so weit, daß sie allen Ernstes behaupten, mit den quantitativen Untersuchungsmethoden würde man „das Wesen des Menschen verfehlen". Uns konnte allerdings bislang noch niemand sagen, worin dieses vielzitierte „Wesen" besteht. Wie kann man aber dann das Wesen des Menschen verfehlen, wenn es noch gar nicht bekannt ist? Kein Geringerer als Leonardo da Vinci (zit. nach Hofstätter & Wendt 1967) tat den Ausspruch: „Wer die Hilfe der Mathematik verschmäht, nährt sich von Verwirrung und kann niemals die sophistischen Disziplinen zum Verstummen bringen, durch die nur fortgesetzt Geschrei erregt wird."

Im Hinblick auf die eingangs gestellte Frage der Quantifizierbarkeit qualitativer Variablen können wir nunmehr unseren eigenen methodologischen Standpunkt pointieren: Wir sind der Ansicht, daß *prinzipiell* alles im Bereich der Sozialwissenschaften quantifizierbar ist, insofern alles dimensionierbar und somit skalierbar ist und weil es eine echte Dichotomie zwischen Psychischem (Humanem usw.) und Metrischem (d.h. der Quantifikation) nicht gibt. Grundsätzliche Quantifizierbarkeit bedeutet jedoch nicht, daß in jedem konkreten Falle Quantifizierung möglich wäre.

Quantifikation vollzieht sich immer in zwei Richtungen, wie Graumann (1965) darlegen konnte. Der erste Schritt ist hierbei sensu Stevens die *Unterscheidung (discrimination)*, d.h. die Feststellung: Etwas ist, oder es ist nicht – Ja oder Nein. Die dichotome Betrachtungsweise der Phänomene steht somit am Anfang wissenschaftlichen Tuns, sie verlangt eine *Kategorisierung* (z.B.: grün – andere Farben; Freude – andere Zustände; Aggression – andere Kontakte; tot – lebendig; männlich – weiblich). Der zweite Schritt betrifft die *Dimensionierung:* Ein Phänomen wird nach der Identifizierung oder Abhebung unter den Aspekt des Mehr oder Weniger gerückt, d.h. dimensional aufgefaßt *(= skaliert).*

Weg 1 führt zu deskriptiven Kategorien, Weg 2 zur eigentlichen Quantifikation. Dabei ist die Quantifikation nicht auf den zweiten Weg beschränkt, wie oft angenommen wird. Auch via Kategorisierung besteht die Möglichkeit zur Quantifizierung – auf Nominalskalenniveau sowieso, aber auch darüber –, indem nämlich ein einzelnes nicht als einzelnes, sondern als Item seiner Klasse aufgefaßt wird. So kann der einzelne Proband zunächst als bester, größter, zuverlässigster usw. seiner Gruppe betrachtet werden, wobei Qualifikationsmöglichkeiten dieser Art anschließend stets hinsichtlich „mehr" oder „weniger" analysiert werden können. Auch die Klasse „Mann" vs. „Frau" erlaubt eine Quantifizierung: Nach der Diskrimination in „männliche" vs. „weibliche" Populationsmitglieder kann man jedes Individuum weiterhin unter dem Aspekt des mehr oder weniger männlich vs. weiblich untersuchen, also quantifizierend betrachten. Auch in bezug auf andere qualitative Variablen (traurig – nicht traurig; aufrichtig – nicht aufrichtig; sozial –

nicht sozial; usw.) ist eine Steigerung oder Minderung, d.h. eine Dimensionierung, möglich. Die wissenschaftlichen Maßstäbe (Eindeutigkeit bzw. Objektivität, Reliabilität bzw. Konsistenz und Validität) gelten für beide Vorgehensweisen: die deskriptive Kategorisierung und die quantifizierende Dimensionierung. Dabei ist es gleichgültig, ob wir mit Ziffern oder anderen Symbolen (i.e.S. metrisch oder kategorial) arbeiten, also Quantifikationen oder Qualifikationen benutzen. Insofern ist auch die Unterscheidung in quantitative und qualitative Variablen nur von sekundärer Bedeutung (vgl. Kap. 2.2.6.2).

2.2.6.4. Probleme der Generalisierung

Ist die Verrechnung abgeschlossen, so gilt es, die gesamten Untersuchungsergebnisse zu interpretieren. Die *Interpretation der Ergebnisse* impliziert praktisch immer einen Akt der generalisierenden Abstraktion, womit wir uns noch ausführlich im abschließenden Kapitel der Methodenlehre (Theoriebildung) zu befassen haben. „In der Interpretation werden die ausgewerteten Beobachtungsdaten auf die Ausgangslage zurückbezogen. Die Interpretation gibt also eine neue, empirisch begründete Antwort auf die im Problem gestellte Frage. Sie kann die als Hypothese formulierte erwartete Antwort bekräftigen, modifizieren oder verdrängen“ (Selg & Bauer 1971, S. 141). Freilich, um diese „neue“ Antwort bzw. Hypothesenentscheidung zu ermöglichen, bedarf es einer Reihe von Schlußfolgerungen, zunächst stichprobenstatistischer Art und dann i.e.S. auf Interpretationsebene.
Beim ersten Schritt handelt es sich um den Schluß von der Stichprobe auf die Population, wozu wir uns der sog. *Signifikanztests* (vgl. Kap. 5) bedienen. Je nach Verteilungscharakter und Skalenniveau der Erhebungsdaten (Informationen) kommen dabei *parametrische* oder *nonparametrische* Tests zum Einsatz. Bei Dependenzanalysen (Aufhellung der Beziehungen zwischen UV und AV) bzw. Interdependenzanalysen (s. S. 56f.) ist die *Korrelationsstatistik* (vgl. Kap. 4) einschlägig. Sowohl der Korrelations- als auch der Stichprobenstatistik liegen Informationsdaten der sog. *deskriptiven Statistik* (vgl. Kap. 3) zugrunde, z.B. Angaben über Mittelwert und Streuung. Bei der Kontrolle statistischer Schlußfolgerungen werden nun die Stichprobenresultate (Statistiken) auf Zufälligkeit versus Überzufälligkeit hin überprüft; hierbei nimmt man ein größeres oder kleineres Fehlerrisiko von Typ α versus β in Kauf (s. S. 187f.).
Als Hauptfehlerquelle auf Interpretationsebene muß die Gefahr der *Überinterpretation* angesehen werden. Im Gegensatz zur stichproben*statistischen* Absicherung (s. oben) gibt es hierfür keine generell gültigen, präzisen Richtlinien. So können Untersuchungsergebnisse, die bei Zehnjährigen gewonnen wurden, nicht ohne weiteres auf die gesamte Grundschulpopulation übertragen werden; Untersuchungsresultate von Pädagogik- oder Psychologiestudenten sind nicht unbedingt repräsentativ für das Studentenkollektiv, d.h. die Studierenden aller Fachrichtungen; auch die Problematik der sog. Freiwilligen, also Selektionstendenzen bereits bei der Stichprobenbildung (vgl. Kap. 2.2.5.1), u.dgl.m. beeinflussen die Möglichkeiten zur Generalisierung auf Interpretationsebene mehr oder weniger empfindlich. Im allgemeinen wird man sich also gegen Verallgemeinerungen auf zu schmaler Generalisationsbasis in dieser oder jener Form schützen müssen, ohne

daß wir dem Leser verbindliche Sicherungskriterien für jeden konkreten Fall benennen könnten. Der Schritt zur *generalisierenden Abstraktion* muß freilich an irgendeiner Stelle im Wissenschaftsvollzug immer geleistet – und gewagt – werden, sollen wissenschaftliche Ergebnisse nicht im vielzitierten Elfenbeinturm, d.h. gleichermaßen im theoretisch unverbindlichen und praktisch irrelevanten Raum, angesiedelt bleiben. Siehe ergänzend dazu die früheren Ausführungen auf S. 62ff. Immerhin bietet die wissenschaftstheoretische Diskussion der letzten Jahre einen beachtlichen Katalog – typisierender – Entscheidungshilfen auf Theoriebildungsebene an, die wir uns im Rahmen der Interpretation empirischer Untersuchungsergebnisse in dieser oder jener Form zunutze machen sollten. Die Einordnung wissenschaftlicher Erkenntnisse in ein Bezugssystem, d.h. hier empirisch fundierter, konkreter Aussagen in eine mehr oder minder allgemeingültige Theorie, ist ja letzten Endes immer das Ziel wissenschaftlicher Forschung – auch in den Sozialwissenschaften. Daß es dazu – neben gründlicher Methodenkenntnisse – auch eines intensiven *Literatur*studiums (bereits *vor* der Aufstellung operationaler Designs) bedarf, ist wohl spätestens an dieser Stelle für jeden Leser einsichtig; entsprechende Forderungen wurden verschiedentlich bei früheren Erörterungen schon impliziert (z. B. S. 19ff. oder S. 28ff.) und brauchen im einzelnen nicht mehr rekapituliert zu werden. Es wäre vermessen und töricht zugleich, bei jeder Untersuchung quasi am Nullpunkt beginnen zu wollen und sich nicht über die bereits geleistete Arbeit auf dem speziellen Untersuchungsgebiet hinreichend zu informieren. Wie soll man empirische Befunde im theoretischen Kontext verankern, wenn man diesen gar nicht kennt? Die Artikulierung dessen, was man untersuchen möchte, stellt sich somit am Anfang *und* am Ende jeder Untersuchung, wenngleich sehr oft unter verschiedenen Aspekten, als Aufgabe. Die nachfolgenden Erörterungen zur Theoriebildung sollen deshalb die Einführung in die Methodenlehre i.e.S. beschließen.

2.2.6.5. Theoriebildung im Wissenschaftsvollzug

Etymologisch betrachtet heißt *Theorie* soviel wie „Anschauung", „Sicht", „Betrachtung" u.ä. Der wissenschaftliche Begriff von „Theorie" ist freilich anspruchsvoller, hier ist nicht nur größere Eindeutigkeit, sondern zugleich ein Höchstmaß an intersubjektiver Verbindlichkeit gefordert (s. S. 19ff.). Wenn hier von Theorie die Rede ist, so ist damit bereits eine Metatheorie, eine Theorie der Theorie, angesprochen.

Oft werden „Theorie" und „Praxis" als Gegensätze betrachtet. Dabei müßte man sich die Fragen vorlegen: „Stimmt die Theorie?" und „Welchen Wert hat die Theorie für den Wissenschaftler vs. Praktiker?"

Ob man nun Theorie und Praxis gegensätzlich sieht oder nicht, Theorien liegt immer ein *Abstraktionsprozeß* zugrunde, worauf bereits im vorherigen Kapitel hingewiesen wurde. Diesen Prozeß, d.h. den Weg von den beobachteten Phänomenen hin zur Theorie (z.B. im Sinne eines Systems von *Prinzipien)*, müßte man also kennen, um eine wissenschaftlich fundierte Aussage über die Theorie machen zu können. Sehr anschaulich hat Goethe, der bekanntlich alle Theorie „grau" nannte, auf den kritischen Punkt bei der Theoriebildung aufmerksam gemacht (in: Maximen und Reflexionen):

„Theorien sind Übereilungen eines ungeduldigen Verstandes, der die Phänomene los sein möchte und statt dessen Bilder oder Worte einschiebt."

Entscheidung für die Theoriebildung ist somit die Wendung von den Phänomenen weg hin zur Theorie. Dabei steht im Zentrum des Interesses die systematische Ordnung, d.h. Darstellung allgemeiner Prinzipien, die zur Klärung bestimmter Phänomene bzw. Phänomenkomplexe dienen sollen. Allen Theorien sind deshalb folgende *Merkmalskriterien* gemeinsam:

1. Theorien haben *Erklärungsfunktionen*, sie sollen etwas verständlich machen oder erklären. Diese Erklärung kann erfolgen a) durch *Herstellung des Kontextes* (z.B. Aufweis motivationaler Zusammenhänge), b) durch ein *Konstruktum* (z.B. Aufdeckung einer Gesetzmäßigkeit[1] oder eines Prinzips), c) durch *Reduktion*, d.h. Zurückführung komplexer Tatbestände auf einfachere (z.B. Angstverhalten auf physiologische Prozesse), was sehr oft zur Vermischung der Bezugssysteme führt.
2. Die Erklärung erfolgt durch den Aufweis *allgemeiner Prinzipien*. Damit ist das Aufspüren von Regelmäßigkeiten oder Gesetzmäßigkeiten in der Vielfalt der Phänomene gemeint, von Prinzipien i.e.S., d.h. des letztlich Verursachenden oder Wesentlichen in der philosophischen Terminologie.
3. Diese *Prinzipien* sollen sich auf *größere Phänomenbereiche* erstrecken – im Gegensatz zu *Hypothesen*, die mehr auf *einzelne* Phänomene und Phänomenverknüpfungen gerichtet sind. Die von Natur aus ungleichartigen Phänomene werden also erst durch unsere theoretische Behandlung mehr oder weniger gleichförmig. Dem geht in der Regel die Beobachtung der Phänomene – und deren Beschreibung (vgl. Kap. 2.1.2) – voraus, wiewohl nicht selten auch der umgekehrte, u.E. gefährlichere Weg eingeschlagen wird (s. unten).

Aus den bisherigen Ausführungen ergeben sich bereits die *Elemente* der wissenschaftlichen Theoriebildung, nämlich 1. die *Phänomene* (Daten oder Fakten) als *empirische* Grundlage einer Theorie, 2. die *Datensprache*, d.h. die Deskription und somit Benennung der Daten, 3. die *Theoriesprache*, d.h. die Artikulierung der Beziehungen zwischen Phänomenen mit Hilfe von Prinzipien oder Gesetzen. Die Kodifizierung kann dabei durch beliebige Symbolsysteme, z.B. sprachlich oder numerisch, erfolgen.

Zu jeder Theoriebildung gehören einerseits *Phänomene* oder *Daten* und andererseits *Prinzipien* oder *Gesetze* bzw. Gesetzmäßigkeiten. Dabei können – wie schon angedeutet – die Prinzipien *über* den Daten stehen (Weg von 1 bis 3 oder von „unten nach oben") versus jene den Daten *zugrunde* liegen (Weg von „oben nach unten"). Nach Möglichkeit sollte man den ersten Weg einschlagen (s. noch S. 23f.).

Theorien bilden häufig den Abschluß einer wissenschaftlichen Arbeit, etwa einer empirischen Untersuchung (= Erklärungsfunktion); sie dienen aber auch als Werkzeuge der Vorhersage (z.B. in der Psychologie). Im Rahmen unserer Erörterung ist besonders der erste Aspekt thematisiert.

Unter metatheoretischen Gesichtspunkten lassen sich nach Marx (1963) (s. noch Topitsch [1960, 1965], Graumann [1965, 1969], Popper [1971] u.a.) mindestens

1 Von „Gesetzen" kann man strenggenommen nur in den Naturwissenschaften reden; in den Sozialwissenschaften sollte man deshalb besser von „Gesetzmäßigkeiten" oder Prinzipien sprechen. Damit wird der Konstruktcharakter sozialwissenschaftlicher Theorien deutlicher zum Ausdruck gebracht, weshalb z.B. Graumann den Begriff „Gesetz" für die (allgemeine) Psychologie ablehnt.

vier verschiedene Konzeptionen einer wissenschaftlichen Theoriebildung (in der Psychologie bzw. den Sozialwissenschaften) unterscheiden.
Die erste Form des Theoretisierens betrifft die *Begriffsbildung*, d.h. einen Vorgang der generalisierenden Abstraktion. *Abstraktion* ist hier als Selektion, als Hervorhebung bestimmter, spezifischer Merkmale einer Situation, zu verstehen. *Generalisation* meint die Verallgemeinerung im Sinne einer Zusammenfassung spezifischer Merkmale zu einer Klasse von sonst variablen Situationen. Via *generalisierende Abstraktion* entstehen so – je nach Universalität vs. Spezifität der jeweiligen Merkmale – Klassen verschiedenen Umfanges. Im weiteren Sinne sind mit dem Vorgang der generalisierenden Abstraktion alle Operationen oberhalb der Beobachtung der Phänomene, d.h. alle formalen (begrifflichen und schlußfolgernden) Prozesse, verstanden. Theoretische Arbeit in diesem Sinne wären beispielsweise Formelableitungen. Dem stünde die *empirische* Arbeit, z.B. Verhaltensbeobachtung, Experiment u.ä., gegenüber.
Die zweite Form des Theoretisierens ist die des *generalisierten Erklärungsprinzips*. Hiermit ist der funktionale Aspekt der Theorie angesprochen. Erklärungsprinzipien, die auf die funktionale Beziehung zwischen den Variablen verweisen, haben im Sinne Graumanns u.a. Gesetzescharakter. Dabei können die Variablen rein empirisch konstituiert (operational gewonnen) oder abstrakter definiert sein (z.B. intervenierende Variablen); zum Variablenbegriff s. die früheren Ausführungen auf S. 51 ff.
Die dritte Form der Theoriebildung impliziert die strengste Konzeption einer Theorie: Theorie als *System logischer Gesetzesverknüpfungen*, wobei die Gesetze *deduktiv* verknüpft werden. Theorien dieser Art finden sich vor allem in der Physik, aber auch in den Sozialwissenschaften, z.B. der Psychoanalyse. Die Systematik solcher Theorien steht freilich nicht selten in krassem Gegensatz zur Empirie. In den empirischen Wissenschaften begegnet man deshalb dieser Theorieauffassung häufig mit einer gewissen Skepsis.
Die vierte Form der Theorie bedient sich der *beschreibenden Feststellung* von Gesetzmäßigkeiten oder Ordnungen, und zwar auf rein operationistischer Grundlage (z.B. via Korrelation). Ihr Ergebnis sind Bündel mehr oder weniger summarischer Feststellungen *(cluster of laws)*. „Gesetze" i.e.S. lassen sich bei diesem Vorgehen, d.h. beim radikalen Beharren auf Operationen unter Vermeidung jeglicher generalisierender Abstraktionen, kaum entwickeln. Dies gilt für alles *induktive* Vorgehen, soweit man dabei auf Hypothesen und Vorannahmen verzichten will. Allerdings darf nicht verkannt werden, daß auch der induktive Theoretiker ungewollt häufig logisch-deduktiv vorgeht, wenngleich seine Prämissen in der Regel implizit bleiben, während der funktionale und deduktive Theoretiker seine Annahmen und Hypothesen zu explizieren gewohnt ist. Sowohl in der Methodenwahl wie auch schon in der Bestimmung seines Untersuchungsgegenstandes stecken solche (impliziten) Voraussetzungen, deren sich der induktive Theoretiker unausgesprochen bedient. Freilich: Das Postulat, die Generalisierung auf möglichst niedriger Ebene – nahe an den Phänomenen – anzusetzen, sollte stets Anliegen wissenschaftlichen Tuns sein. Darauf haben wir des öfteren nachdrücklich hingewiesen, ohne unangebrachtem Methoden-Monismus das Wort zu reden.
Unser Überblick über die verschiedenen Formen – Typen oder auch nur Akzentuierungen – der Theoriebildung versuchte zu belegen, daß Theorien Modi der Generalisierung darstellen, deren Zweck die Erhellung oder Erklärung bestimmter

Phänomene sowie komplexerer Tatbestände ist. Abschließend bringen wir eine in Anlehnung an Graumann aufgestellte Modellskizze, die den Weg der Theoriebildung im Wissenschaftsvollzug noch einmal verdeutlichen soll.

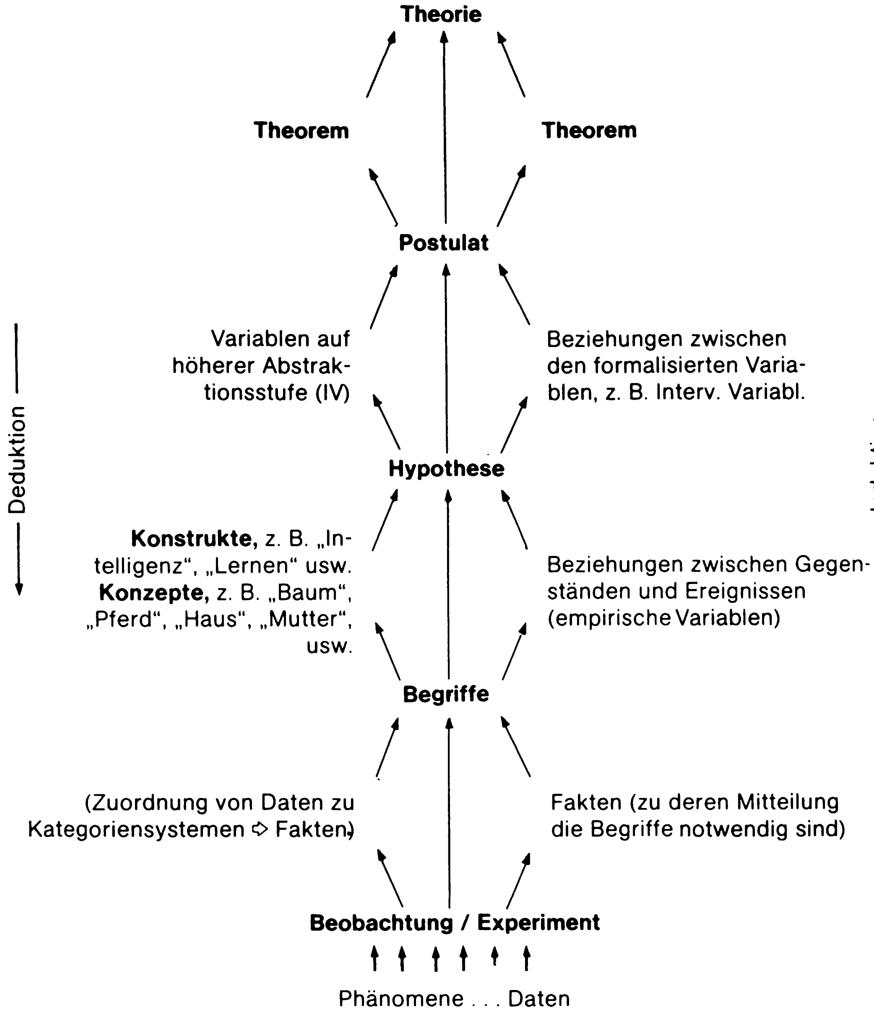


Abb. 3: Schematische Darstellung der Theorienbildung

Einführung in die Forschungsstatistik

3. Deskriptive Statistik

3.1. Wichtige Begriffe

Nach der mehr allgemein gehaltenen Einführung in die Methodik empirisch-pädagogischer bzw. sozialwissenschaftlicher Forschung wenden wir uns nun der sog. *Statistik (statistics)* als jenem Teil der Methodenlehre zu, der Bereiche der angewandten Mathematik repräsentiert. Die gesamte Statistik läßt sich in zwei oder drei große Teilgebiete aufgliedern: die *deskriptive* Statistik (Kap. 3), die *Inferenz*statistik (vgl. Kap. 5) und die *Korrelations*rechnung (vgl. Kap. 4), die eigentlich noch der deskriptiven Statistik zuzuordnen wäre. Während die *deskriptive* Statistik sich der ordnenden *Beschreibung* – graphischer und/oder numerischer Art – quantitativer Merkmalsdaten bedient, ermöglicht die *analytische* oder *Inferenz*statistik (stichproben)statistische *Schlußfolgerungen* (s. S. 163 u. 180ff.) auf induktivem Wege. Beide Ansätze sind in der empirischen Forschung notwendig und informativ.
Als *deskriptive Statistik* bezeichnet man alle Techniken oder Methoden, die der *unmittelbaren Beschreibung einer Menge vorhandener Daten – in quantitativer Hinsicht* – dienen. Süllwold definiert deshalb die deskriptive Statistik als „Beschreibung der quantitativen Eigenschaften einer direkt beobachteten, vollständig erfaßten Gruppe“. Damit ist die Anwendung der deskriptiven Statistik im Hinblick auf Populationen angesprochen.
Unter *Population* (Kollektiv oder Grundgesamtheit) wird die *Gesamtheit aller Merkmalsträger* verstanden, d.h. aller Individuen, die gleiche Merkmale oder Merkmalskombinationen tragen. Populationen sind keine natürlich gewachsenen, sondern immer *definierte* Gruppen von Individuen oder Beobachtungsdaten, z.B. Grundschulpopulation (= *alle* Grundschüler von der 1. bis 4. Klasse), Abiturientenpopulation, Kollektiv der Realschullehrer, PH-Studenten usw.
Statistische Kenndaten oder Maße, die sich auf eine Population beziehen, werden *Parameter* genannt. Parameter sind somit „exakte, endgültige oder wahre Werte, durch die eine vollständig erfaßte Gruppe charakterisiert wird“ (im Sinne Süllwolds). Parameter können Maße der zentralen Tendenz (Mittelwerte), Variabilitätswerte, Korrelationskoeffizienten u.ä. sein; aber auch Sätze wie „Physikstudenten sind im rechnerischen Denken besser als Philologiestudenten“ oder „Mädchen haben bessere Schulzeugnisse als Jungen“ gelten als Parameter, sofern die genannten Aussagen die Population, d.h. *alle* Mitglieder der betr. Gruppe berücksichtigen.
Aus zeitlichen, ökonomischen oder anderen Gründen ist es jedoch selten möglich, eine vollständige Population zu erfassen. Hier ist man dann auf mehr oder weniger große Teile der Population, auf sog. Stichproben, angewiesen. Eine *Stichprobe* kann als *relativ kleine Gruppe von Angehörigen einer bestimmten Population* definiert werden (vgl. Kap. 2.2.5.1). Statistische Maße, die sich auf Stichproben beziehen, nennt man *Statistiken* (z.B. Mittelwerte, Prozentzahlen, Koeffizienten usw.).

Bei den Erhebungsdaten werden – je nach Skalentyp (s. S. 81 ff.) – quantitative und qualitative Merkmalsdaten unterschieden. *Quantitative Merkmale* sind alle Merkmale, die eine *kontinuierliche* Reihe mit *definierten Einheiten* bilden können, d.h. als *Maßzahlen auf einem Kontinuum* erscheinen. Demnach repräsentieren Intelligenz, Schulleistung, Motivation u.ä. oder Größe, Gewicht, Lernzeit usw. Häufigkeiten einer Intervall- oder Rationalskala und somit quantitative Merkmale. *Qualitative Merkmale* sind hingegen *Kategorien-Häufigkeiten*, d.h. Häufigkeiten ohne ein Kontinuum, z.B. Autotypen, Augen- oder Haarfarbe, Geschlecht, Berufsklasse, Schuleignungskategorie u.ä.
Sofern sich die Meßwerte auf eine kontinuierliche Skala beziehen, also das betr. Merkmal mindestens auf Intervallskalenniveau erhoben wurde, spricht man von *Maßzahlen (scores)*. Maßzahlen werden in der Statistik mit dem Großbuchstaben X symbolisiert. Solche Maßzahlen können beliebige Werte auf der Skala annehmen, sie sind also veränderliche Größen oder *Variablen* i.e.S.

„Die *abhängige* Variable einer Untersuchung ist genaugenommen nicht das untersuchte *Merkmal*, sondern die zahlenmäßige Repräsentation der Merkmalsausprägung, wie sie mittels der Skala gemessen wurde. Sagt jemand z.B., er wolle die ‚Variable' namens ‚Handgeschicklichkeit' messen, so meint er eigentlich die Feststellung des *variablen* Maßes für Handgeschicklichkeit an einer Gruppe von Individuen. Dieses Maß ist deshalb Variable, weil es bei verschiedenen Beobachtungen an verschiedenen Fällen zu verschiedenen Gelegenheiten jeweils unterschiedliche zahlenmäßige Größen annimmt, die den individuellen Grad der Merkmalsausprägung von Handgeschicklichkeit wiedergeben. Ist X_i die Bezeichnung für irgendein erhaltenes Maß, für eine *Maßzahl*, in bezug auf ein zahlenmäßiges Merkmal X, so indizieren wir die erste Beobachtung, die erste Maßzahl, *ungeachtet* ihrer numerischen Größe, mit dem Laufindex $i = 1$ und schreiben X_1. Die letzte Messung wird bei einer Stichprobe vom Umfang N den Laufindex $i = N$ erhalten, so daß wir X_N schreiben. Die vollständige Reihe der Ergebnisse lautet dann: $X_1, X_2 \ldots X_i \ldots X_N$. Manchmal müssen mehrere Indizierungen verwendet werden, um die Meßdaten in der rechten Ordnung zu betrachten. Nehmen wir z.B. an, wir haben unsere Variable X bei k verschiedenen Stichproben zu j verschiedenen Gelegenheiten an i Individuen pro Stichprobe ermittelt. Um nun die Herkunft der Daten an der Indizierung ablesen zu können, wird jede Maßzahl in der Form X_{ijk} zu notieren sein. Die Maßzahl des fünften Individuums aus der Beobachtung bei der dritten Gelegenheit in der vierten Stichprobe wird dann als X_{354} bezeichnet" (Fröhlich & Becker 1971, S. 25).

Mit diesen Begriffen und Symbolen wollen wir es vorläufig bewenden lassen. Weitere Termini, Symbole und Indizes werden im Verlauf der Darstellung einzelner statistischer Methoden zu klären sein. Ihre kursorische Vorwegnahme würde den Leser vermutlich mehr verwirren als informieren. Überhaupt ist es unser Bestreben, die Methoden der Statistik immer eingebettet in konkrete Versuchspläne darzustellen und anhand eines Rechenbeispiels ihre Anwendung zu demonstrieren. Erst danach werden jeweils die Anwendungs*regeln* einschließlich relevanter Voraussetzungen und Einschränkungen formuliert. Wir erachten dieses „induktive" Vorgehen didaktisch für vorteilhafter als den umgekehrten Weg, zumal anschließend an jede Methodenexplikation durch eine Reihe von Übungsbeispielen dem Leser Gelegenheit geboten wird, sein Wissen und Können selbst zu überprüfen.

3.2. Ordnen und Beschreiben empirischer Häufigkeitsverteilungen

3.2.1. Erstellung von Urlisten und Häufigkeitstabellen

Aus den Untersuchungs- oder Testprotokollen werden zunächst alle Daten gesammelt und in die sog. Urliste übertragen. Dabei kann man die Daten fortlaufend (ungeordnet) festhalten oder bereits der Größe nach ordnen. Danach werden die Daten erst der eigentlichen Verrechnung zugeführt. Folgendes Beispiel, das sich auf eine Begabungsuntersuchung in der 5. Hauptschulklasse bezieht, möge das Vorgehen im einzelnen erläutern.

Beispiel für eine *ungeordnete* Urliste:
23 41 15 30 30 21 33 36 12 26 29 46 28 32 33 38 18 42 27 34 24 19 48 25 31
44 35 37 17 22 29 29 34 39 10 21 28 12 30 32 30 31 24 35 26 33 38 32 28 25

Beispiel für eine *geordnete* Urliste (anhand desselben Datenmaterials):
10 12 12 15 17 18 19 21 21 22 23 24 24 25 25 26 26 27 28 28 28 29 29 29 30
30 30 30 31 31 32 32 32 33 33 33 34 34 35 35 36 37 38 38 39 41 42 44 46 48

Die Erstellung der Urliste dient vor allem der weiteren Datenverarbeitung (Verrechnung). Darüber hinaus vermittelt diese Form der Datenanordnung schon einen ersten Überblick, z.B. über die niedrigste Leistung (X_{min}) und die höchste Leistung (X_{max}) in der Untersuchungsstichprobe. Aus der Differenz von X_{max} und X_{min} ergibt sich ferner die *absolute Streuweite* oder *Dispersionsspanne*, die in unserem Beispielfall von $X = 10$ bis $X = 48$ auf der Rohpunktskala des zugrundeliegenden Intelligenztests[1] reicht. Differenziertere Informationen, etwa über die Verteilung zwischen den Extremen (X_{max} und X_{min}), sind dem bisherigen Arrangement allerdings kaum zu entnehmen. Deshalb bedient man sich häufig einer modifizierten Form der Urliste, der sog. *Strichliste*. Bei umfangreicheren Datenmengen empfiehlt sich außerdem die Bildung von *Klassenintervallen*, indem man die einzelnen Maßzahlen (X) zu Maßzahl*klassen* ($\bar{X}$) zusammenfaßt. Auf diese Weise wird nicht nur die Übersichtlichkeit erhöht, man verringert auch den Arbeitsaufwand für die weiteren Verarbeitungsgänge.

Anleitungsregeln

a) Über die Extremwerte (X_{max} und X_{min}) wird die absolute Dispersionsspanne ermittelt. Diese ist für die Bildung der Maßzahlklassen ausschlaggebend.
b) Unter Berücksichtigung der Dispersionsspanne wird die Intervallbreite (i) so festgelegt, daß nicht weniger als 10 und nicht mehr als 20 Maßzahlklassen ($\bar{X}$) entstehen.
c) Schließlich werden die einzelnen Vpn oder Pbn unter Berücksichtigung des jeweils erzielten Test*scores* in die betr. Maßzahlklasse eingestrichelt und abschließend die Zeilensummen in die f-Spalte (f = Frequenzen) übertragen. Sofern es wünschenswert erscheint, kann man darüber hinaus die Namen der Vpn in einer weiteren Spalte in der zutreffenden Zeile (Maßzahlklasse) festhalten.

1 Die Erhebungsdaten beziehen sich auf die Rohpunktleistung in den Subtests 1 + 2 im Prüfsystem für Schul- und Bildungsberatung (PSB) von W. Horn (vgl. Heller 1973, S. 139f.).

$\bar{X}$	Strichliste	f	Name(n) d. Vp(n)
8–10	/	1	Fritz K.
11–13	//	2	Otto M., Karin F.
14–16	/	1	usw.
17–19	///	3	
20–22	///	3	
23–25	~~////~~	5	
26–28	~~////~~ /	6	
29–31	~~////~~ ////	9	
32–34	~~////~~ ///	8	
35–37	////	4	
38–40	///	3	
41–43	//	2	
44–46	//	2	
47–49	/	1	

i = 3 (Intervallbreite) N = 50 (Stichprobengröße)

Die mittels Strichliste erhaltene *Frequenztabelle* oder *Häufigkeitstabelle* gibt nun detailliert Aufschluß über die empirische Häufigkeitsverteilung hinsichtlich der PSB-Leistung in der Untersuchungsstichprobe. Auch der Kulm wird deutlich. Noch anschaulicher treten die Verteilungscharakteristika in der graphischen Darstellung zutage.

3.2.2. Graphische Darstellungsmethoden von Häufigkeitsverteilungen

Für die graphische Darstellung von Häufigkeitsverteilungen bedient man sich gewöhnlich der Intervall-Mittelpunkte (Mp). Laut Konvention werden hierbei alle Werte eines Klassenintervalls als identische Werte aufgefaßt. In unserem Beispiel ergeben sich folgende Mittelpunkte: 9, 12, 15, 18... 48 Mp. Nicht immer lassen sich die Klassenmitten so leicht bestimmen; in solchen Fällen verwendet man nachstehende *Formel zur Berechnung der Mittelpunkte* (Mp):

$$\mathrm{Mp} = \text{untere Grenze (des betr. Intervalls)} + \frac{\text{obere Grenze} - \text{untere Grenze}}{2}$$

Für unser Rechenbeispiel ergäbe sich somit:

$$\mathrm{Mp} = 7.5 + \frac{10.5 - 7.5}{2} = 9.0$$ [1]

$$\mathrm{Mp} = 10.5 + \frac{13.5 - 10.5}{2} = 12.0 \quad \text{usw.}$$

Es gibt verschiedene Methoden zur Abbildung von Häufigkeitsverteilungen. Am gebräuchlichsten sind *Polygon*züge und *Histogramm*darstellungen. Dabei sind folgende *Regeln* zu beachten:

1 Jeder Punkt auf einer kontinuierlichen Skala stellt einen Bereich dar. So erstreckt sich der Rohpunkt 8 auf den Bereich von 7.5 bis 8.5 oder genauer: von 7.50 bis 8.49, der Rohpunkt 9 auf den Bereich von 8.5 bis 9.5 usw.

a) Konventionell trägt man auf der Abszisse (X-Achse) die Werte der kontinuierlichen Skala (= Merkmalsskala) ab, auf der Ordinaten (Y-Achse) die gefundenen Häufigkeiten oder Frequenzen (f).
b) In der Regel wählt man die Skaleneinteilung so, daß die Zahlenwerte von links nach rechts bzw. von unten nach oben ansteigen. Innerhalb derselben Achse müssen die Einheiten (Punktabstände) konstant gehalten werden.
c) Hingegen sind unterschiedliche Maßstäbe für die Abszissen- und Ordinatenachse nicht nur erlaubt, sondern vielfach sogar erwünscht. Aus Gründen der Einheitlichkeit und somit des besseren Vergleichs unterschiedlicher Häufigkeitsverteilungen sollten Abszisse und Ordinate etwa im Verhältnis 2 : 1 oder 4 : 3 gewählt werden.
d) Aufgrund internationaler Konvention wird jede Kurve auf die X-Achse heruntergezogen. Zu diesem Zweck muß an den Enden des Polygonzugs links und rechts jeweils eine Skaleneinheit bzw. ein Klassenintervall angehängt werden.
e) Schließlich sei noch einmal darauf hingewiesen, daß bei gruppierten Daten (Klassenintervallen) die Abszisseneinheiten durch die Mittelpunkte der Intervalle (Mp) repräsentiert werden.
f) Die Wahl der Intervallbreite (i) hängt einmal von der Ordinaten/Abszissen-Proportion (1 : 2 oder 3 : 4) und zum andern von der benötigten Anzahl der Abszisseneinheit (in der Regel zwischen 10 und 20 Einheiten) ab.

Ob man die *Histogramm*darstellung (Abb. 4) oder die *Polygon*darstellung (Abb. 5) wählt, hängt mehr oder weniger vom persönlichen Geschmack jeweils ab. Die Darstellungsfunktion ist – jenseits ästhetischer Überlegungen – dieselbe: Die Gesamtzahl der Daten wird durch Histogramm und Frequenzpolygon (Häufigkeitspolygon) isomorph abgebildet, d.h., die dargestellte Gesamtfläche bleibt konstant (vgl. Abb. 6).

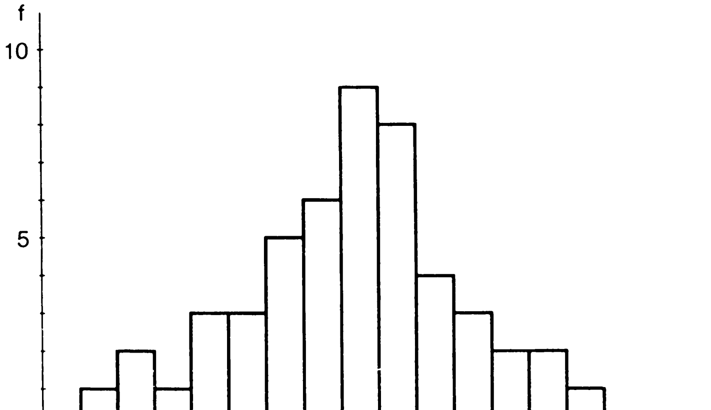


Abb. 4a: Histogramm (Blockdarstellung) unter Verwendung von Klassenintervallen: Rohpunktleistung (X) von 50 Hauptschülern der 5. Klasse im PSB 1 + 2 (s. Frequenztabelle auf S. 96)

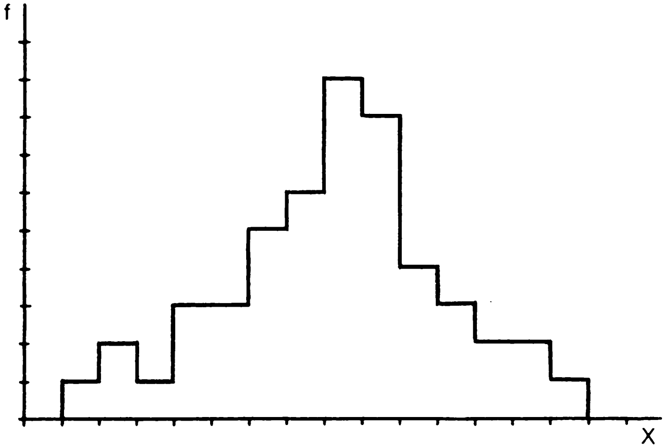


Abb. 4b: Treppenpolygon (vgl. Fröhlich & Becker 1971, Heller 1973 u.a.)

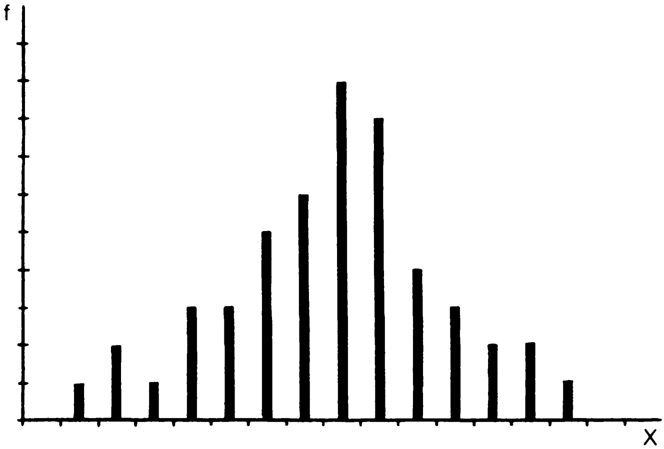


Abb. 4c: Säulendarstellung (vgl. Heller 1970)

Mit Hilfe solcher graphischen Darstellungen lassen sich Untersuchungsgruppen recht gut beschreiben. So kann man – neben dem Verlauf der Häufigkeitsverteilung – beispielsweise Aussagen darüber gewinnen, ob man eine homogene oder heterogene Gruppe vor sich hat, eine durchschnittliche oder leistungsstarke vs. leistungsschwache Schulklasse u.dgl.m. (s. unten Kap. 3.2.3).
Für die graphische Darstellung der Häufigkeitsverteilung quantitativer Variablen sind immer zwei Dimensionen erforderlich: die Merkmalsskala qua Kontinuum zur Dimensionierung der (quantitativen) Variablen und die Ordinatenachse zur Festlegung der Frequenzen. Hingegen genügt für die Darstellung *qualitativer* Variablen allein die Frequenzskala (Ordinate). Solche diskreten oder nichtkontinuierlichen Merkmalsdaten bzw. Kategorien auf Nominalskalenniveau sind

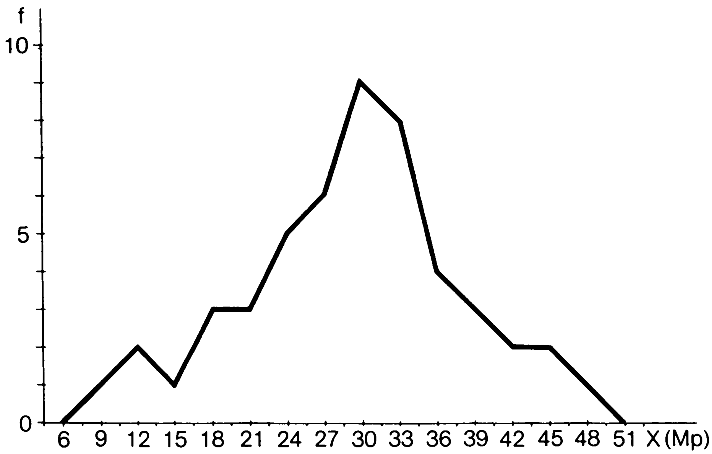


Abb. 5: Frequenzpolygon oder Häufigkeitspolygon (graphische Darstellung der Tabelleninformationen von S. 96 oben)

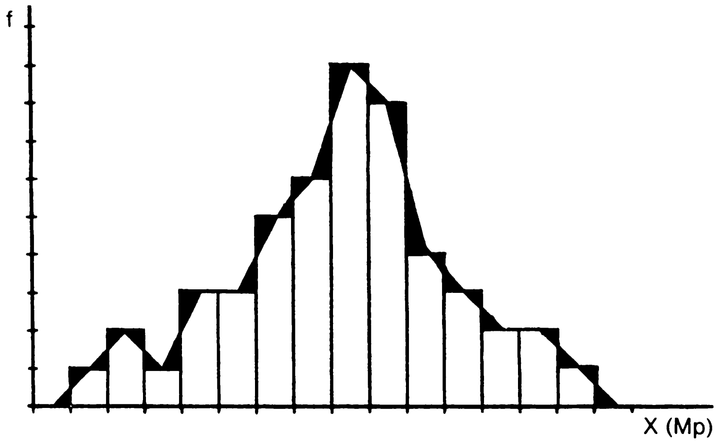


Abb. 6: Vergleich der Darstellungsfunktion von Histogramm und Frequenzpolygon (s. S. 97)

statistisch lediglich durch die *Häufigkeit* ihres Auftretens charakterisiert. Die wichtigsten Darstellungsmethoden sind hierbei das *Block-* oder *Säulendiagramm* (Abb. 7), das *Balken-* oder *Streifendiagramm* in der vertikalen Anordnung (Abb. 8a) versus horizontalen Anordnung (Abb. 8b) und die *Kreis-* oder *Sektorendarstellung* (Abb. 9). Gelegentlich werden noch andere Formen verwendet, z.B. das *Kartogramm* zur Darstellung der Besiedlungsdichte, unterschiedlicher Bildungsteilhabe, Begabungsreserven u.ä. (vgl. Heller 1970). Die Frequenzen können wiederum absolut oder relativiert (%) angegeben werden. Die Erstellung und Interpretation solcher Diagramme bereitet im allgemeinen keine große Mühe; ihr Anschauungswert wird durch nachstehend aufgeführte Beispiele hinreichend belegt.

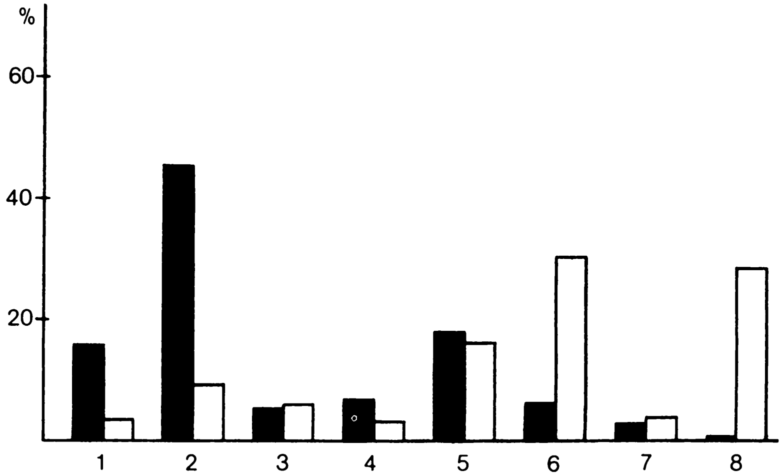


Abb. 7: Beispiel für *Block-* oder *Säulendiagramm:* Sozio-ökonomischer Status (gemessen am Vaterberuf) und Bildungsteilhabe an Gymnasium (weiße Säulen) versus Hauptschule (schwarze Säulen) – nach Heller 1970, S. 205
Legende: 1 = Rentner u. Hilfsarbeiter; 2 = Facharbeiter; 3 = selbst. Handwerker u. Handwerksmeister; 4 = Bauern; 5 = einf. Angestellte u. Beamte; 6 = mittl. Angestellte u. Beamte, Kaufleute usw.; 7 = Selbständige; 8 = Akademiker

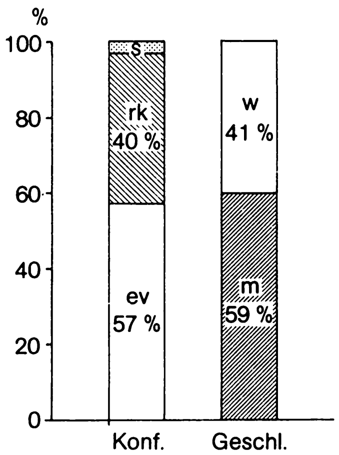


Abb. 8a: Balkendiagramm: Konfessionsanteile und Geschlechtszugehörigkeit von 881 Gymnasiasten (nach Heller 1970, S. 204)

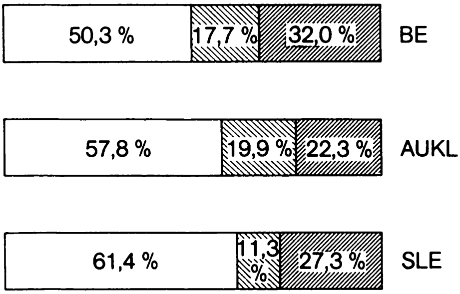
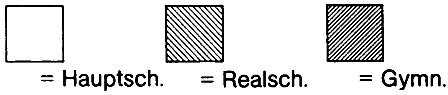
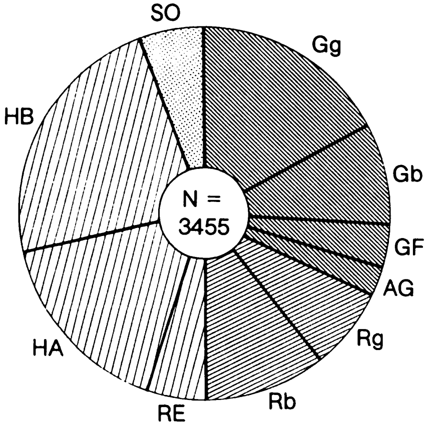


Abb. 8b: Streifendiagramm: Synoptische Darstellung des Schuleignungsbestandes (nach AUKL[1]), der Bildungsempfehlung (BE) und der tatsächlichen Schullaufbahnentscheidung (SLE) am Ende der Grundschule; Prozentuierungsbasis: N = 3500 (100%)

Legende:





Gg = 17,1 %
Gb = 8,5 %
GF = 3,8 %
AG = 2,6 %
Rg = 6,9 %
Rb = 10,8 %
RE = 5,0 %
HA = 16,4 %
HB = 22,9 %
SO = 6,1 %

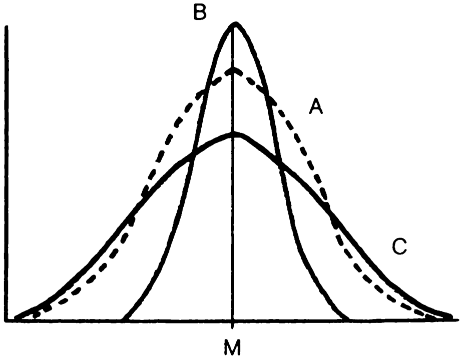
Abb. 9: Beispiel für die *Kreisdarstellung:* Schuleignungsquoten eines Großstadtkollektivs der 4. Grundschulklasse
Anm.: Die Eignungskategorien sind auf S. 78 in diesem Buch beschrieben.

3.2.3. Typische Verteilungsformen

Bei der graphischen Darstellung von Häufigkeitsverteilungen (quantitativer Variablen) begegnet man in der Praxis oft mehr oder weniger von der Normalverteilungsform abweichenden Kurvenverläufen. Eine erste Inspektion vorliegender Verteilungsgraphiken kann oft schon die Frage, ob *Normalverteilung* (s. Kap. 5.1.2 unten) der untersuchten Variablen gegeben ist, klären helfen. Eine exaktere Beantwortung dieser Frage ermöglicht der Chi^2-Test in der Anwendung als Anpassungstest (vgl. Kap. 5.4.1.3). Darüber hinaus lassen sich aber aus der jeweiligen Verteilungsform zahlreiche weitere Informationen entnehmen, wie beispielhaft an den nachstehend abgebildeten Verlaufstypen gezeigt werden soll.

---

1 *AUKL* = Automatische Klassifikation psychologischer Untersuchungsbefunde. Siehe dazu Heller 1970 u. 1973.



Legende: A = mesokurtische (Normal-) Kurve (mäßig steil), B = leptokurtische Kurve (steilgipflig), C = platykurtische Kurve (flachgipflig), M = arithm. Mittel (vgl. Kap. 3.3.3)

Abb. 10: Drei Frequenzpolygone unterschiedlicher Steilheitsgrade

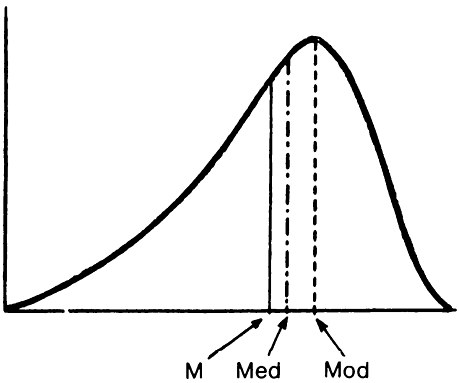


Abb. 11a:
Linksschiefe Kurve (negative Schiefe)

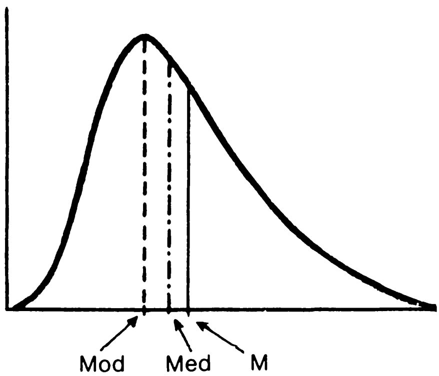


Abb. 11b:
Rechtsschiefe Kurve (positive Schiefe)

Abb. 11: Asymmetrische Verteilungskurven (mit negativer vs. positiver Schiefe)
Legende: M = arithm. Mittel, Med = Median, Mod = Modus (vgl. Kap. 3.3)

Die in Abb. 10 wiedergegebenen Frequenzpolygone haben alle denselben Mittelwert (M), sie unterscheiden sich aber hinsichtlich der Streuung. Während platykurtische Verteilungsformen eine große Streuung und somit *heterogene* Varianzen indizieren, deuten leptokurtische Verteilungskurven auf eine relativ geringe Streuung und somit *homogene* Varianzen hin. Im letzteren Beispiel (Kurve B) scharen sich die meisten Fälle eng um den Mittelwert, während sie bei echter Normalverteilung (Kurve A) im Sinne der Gaußschen *Glockenkurve* (s. S. 174) und bei flachgipfliger Verteilung relativ breit um den Mittelwert streuen.
Auch die *bimodale* Kurve (Abb. 12) ist symmetrisch zur Mitte der Verteilung (M) hin, weist jedoch im Gegensatz zu den unimodalen (eingipfligen) Kurven in Abb. 10 zwei Verteilungsgipfel auf. Die Frequenzmaxima liegen hierbei zwischen dem Mittelwert der Verteilung und den beiden äußeren Verteilungsgrenzen. Im Extremfalle rutschen die Kurvengipfel ganz nach außen, so daß dann die Frequenz-

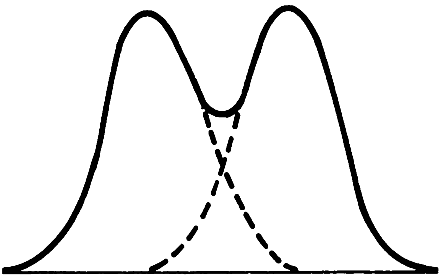


Abb. 12: Bimodale (zweigipflige) Verteilungskurve

maxima mit den Verteilungs*grenzen* zusammenfallen; diese Kurve bietet das Bild zweier sich zugekehrter J-Kurven (s. S. 165 unten). Bimodale Häufigkeitsverteilungen entstehen sehr oft durch Überlappung zweier Normalverteilungen. So können z. B. die Schulleistungsresultate im Lesen oder Aufsatz auf einer bestimmten Alters-/Klassenstufe bimodal verteilt sein, d. h. die Jungen- und die Mädchenkurve sich überlappen. Solche Überschneidungen sind auch zwischen Hauptschülern und Realschülern versus zwischen Realschülern und Gymnasiasten hinsichtlich diverser Lernleistungsmerkmale nachweisbar (vgl. Heller 1970, 1973). Mit bimodalen Verteilungen ist also immer dann zu rechnen, wenn zwei jeweils für sich genommen homogene (Unter-)Gruppen als eine einzige – heterogene – Gruppe betrachtet werden.

Wohl die häufigste Variante der Normalverteilungskurve bilden (unimodale) asymmetrische Verteilungsmodi der in Abb. 11 wiedergegebenen Formen. Dabei differieren die drei Mittelwerte M, Median und Modus – je nach Schiefegrad – mehr oder weniger stark. In der Extremform geht die schiefe Verteilung in die sog. J.-Kurve (s. S. 165) über.

3.2.4. Methoden der Kurvenglättung

Bei der graphischen Darstellung empirischer Häufigkeitsverteilungen stören zuweilen kleinere Unregelmäßigkeiten (z. B. Zacken) des Kurvenverlaufs. In solchen Fällen kann man eine Kurven*glättung* vornehmen. Glättungen sind allerdings keine reinen Beschreibungen mehr, insofern hierbei die Frequenzen verändert werden; der Boden bloßer Deskription wird somit bereits verlassen. Glättungen des Kurvenverlaufs sollen lediglich kleinere Unregelmäßigkeiten, die zu Lasten eines begrenzten Stichprobenumfanges gehen, ausgleichen, um so die Verlaufstendenz der betr. Kurve optisch besser zur Erscheinung zu bringen. Kurvenglättungen können auf graphischem bzw. optischem Wege (vgl. Abb. 13) oder auf rechnerischem Wege (Methode der gleitenden Mittelwerte) vorgenommen werden.

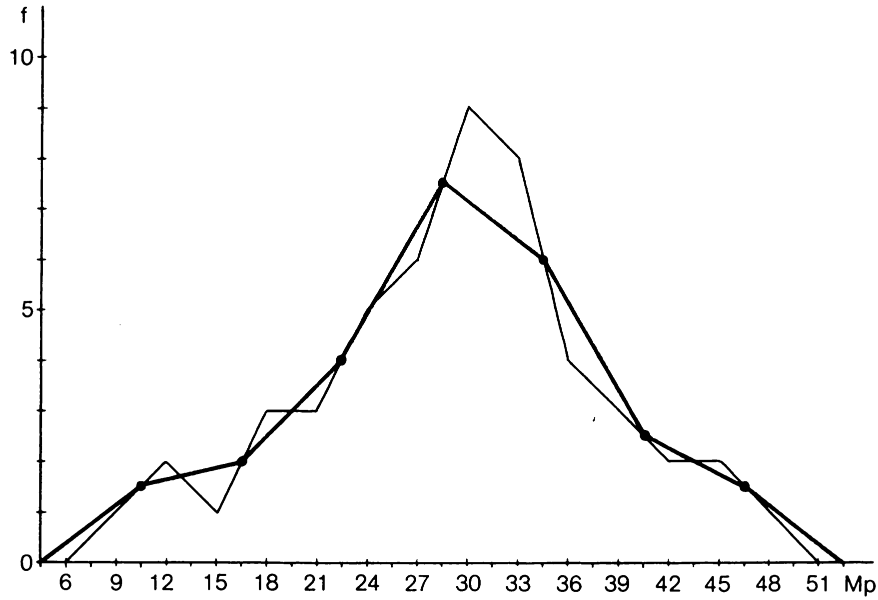


Abb. 13: Frequenzpolygon vor und nach der Kurvenglättung (*Sample*-N = 50)
Legende: Originalverteilung = dünne Kurvenlinie, Glättung = breite Kurvenlinie

Anleitungsregeln für die Durchführung der

A. Kurvenglättung auf graphischem Wege:

a) Zunächst werden die arithmetischen Mittel von je zwei aufeinanderfolgenden Punktwerten, also zwischen dem 9. und 12., dem 15. und 18., dem 21. und 24. usw. Mp in unserem Beispiel, *graphisch* bestimmt (Markierung durch Punkte).

b) Sodann werden die Markierungspunkte miteinander verbunden (= geglättete Kurve); s. Abb. 13.

c) Sofern es wünschenswert – und notwendig – ist, kann eine zweite Glättung nach demselben Verfahren angeschlossen werden. Dabei dient das Ergebnis der ersten Glättung als Ausgangskurve der zweiten Glättung.

B. Kurvenglättung auf rechnerischem Wege:

a) Hierzu bildet man die sog. adjustierte Frequenz: Diese ergibt sich, indem man die Frequenzen des kritischen Intervalls (d.h. hier der zu adjustierenden Mp-Klasse) sowie der benachbarten Intervalle (Mp-Klassen darunter und darüber) aufsummiert und die Summe durch 3 teilt. Im Hinblick auf unser Polygonbeispiel ergeben sich folgende neue Frequenzen:

$$\frac{0+1+2}{3} = 1.0 \quad \frac{1+2+1}{3} = 1.3 \quad \frac{2+1+3}{3} = 2.0 \quad \text{usw.}$$

b) Die adjustierten Frequenzen werden in das Koordinatensystem eingetragen und die betr. Werte (neuen Frequenzen) miteinander verbunden (= geglättete Kurve).

c) Analog zur Methode A können auch hier im Bedarfsfalle mehrere Glättungen vorgenommen werden.

Voraussetzungen und Einschränkungen

a) Kurvenglättungen dürfen nur dann vorgenommen werden, wenn der begründete Verdacht besteht, daß die beobachteten Unregelmäßigkeiten im Kurvenverlauf auf die Kleinheit der Stichprobe zurückzuführen sind. Ansonsten würde man eventuelle (echte) Verteilungscharakteristika gemessener Merkmale bzw. Untersuchungsgruppen vertuschen und damit in ihrem Aussagewert verkennen.
b) Aus den Frequenzen der geglätteten Kurve dürfen keine Statistiken (z.B. Mittelwerte, Streuungsmaße, $\%_0$-Werte u.ä.) abgeleitet werden. Hierzu müssen immer die Ausgangsdaten (Frequenzen der Originalkurve) herangezogen werden.

Übungsbeispiele zu Kap. 3.2

Nr. 3/1
Folgende Werte (ungeordnete Urliste) seien gegeben:
85 66 76 45 66 91 77 64 71 74 47 78 76 42 70 58 71 67 80 78 73 48 68 87 72
65 69 73 84 81 75 56 58 87 56 72 62 93 73 83 97 81 51 61 53 72 62 79 88 79
Nach der Ordnung der Werte (Rohpunkte) und Intervallbildung ($i = 5$: 40–44, 45–49 usw.) sollen diese via Strichliste tabelliert und soll die Frequenzverteilung in einem Polygon graphisch dargestellt werden. Ferner sind die adjustierten Frequenzen zu ermitteln und die Kurvenglättung durchzuführen (Methode B).
Nr. 3/2
Die Synopse von Abb. 8 (Streifendiagramme) ist auf die Kreis-/Sektorendarstellung zu übertragen. Die Zentriwinkel α berechnen sich für die einzelnen Sektoren nach folgender Beziehung:

$$\frac{100}{f} = \frac{360^\circ}{\alpha} \quad \text{bzw.} \quad \alpha = \frac{f \cdot 360^\circ}{100} \quad \text{oder} \quad \alpha = f \cdot 3.6^\circ$$

3.3. Maße der zentralen Tendenz (Mittelwerte)

Wesentlich prägnanter als bisher lassen sich die zentralen Tendenzen einer Häufigkeitsverteilung durch Herstellung von Mittelwerten beschreiben. Für die deskriptive Statistik sind vor allem drei Arten von Mittelwerten von Bedeutung: das arithmetische Mittel, der Median oder Zentralwert und der Modalwert. Wir beginnen mit den meßtheoretisch einfacheren, d.h. voraussetzungsärmeren Verfahren.

3.3.1. Modus oder Dichtemittel

Der Modus (auch Modalwert oder Dichtemittel genannt) ist der Wert in der Verteilung, der am häufigsten vorkommt, d.h., wo die Verteilungsdichte am größten ist. Man kann ihn einfach aus der Frequenztabelle ablesen oder graphisch durch Fällen des Lotes vom Kurvengipfel auf die Abszissenachse bestimmen. In unserem Rechenbeispiel (vgl. Abb. 5) liegt der Modus beim Mp = 30. Dies ist der sog. grobe Modus, der bereits auf dem niedrigsten Meßniveau (Nominalskala) verwendet werden darf. Er ist somit ein verteilungsunabhängiges Maß.
Von diesem *groben* oder – im Falle der Mp-Verwendung – *angenäherten* Modus

ist nach Fröhlich & Becker (1971, S. 72) der „wahre" Modus (manchmal auch als „echter" oder „mathematischer" Modus bezeichnet) zu unterscheiden. Der *wahre* Modus bestimmt sich nach folgender Formel: $\text{Mod} = 3 \cdot \text{Med} - 2 \cdot \text{M}$
Hierbei bedeuten:
Mod = Modus („echter" Modus),
Med = Median (vgl. Kap. 3.3.2),
M = arithm. Mittel (vgl. Kap. 3.3.3).

3.3.2. Median oder Zentralwert

Die Verwendung des Zentralwertes setzt nun bereits Ordinalskalenniveau der Meßdaten voraus, *nicht* jedoch Intervallskalenniveau und Normalverteilung. Der Median ist der Punkt in einer Verteilung, über dem 50% der Fälle liegen und unter dem 50% der Fälle liegen. Zu beachten ist, daß es sich hierbei um einen *Punkt* und nicht um einen Bereich handelt. Beispielsweise seien folgende Meßdaten vorhanden:

5 5 6 7 9 10 12

Hier ist der Median (Med oder Md) gleich dem Wert 7, also: Med = Md = 7.
Anderes Beispiel:

3 4 5 6 7 9

Hier ist laut Konvention der Median identisch mit dem Wert *zwischen* 5 und 6, d.h. gleich dem Mp des betr. Intervalls, also: Med = 5.5.
Bei größeren Datenmengen kann man den Median einfach nach folgender Formel bestimmen:

$$\text{Med} = \frac{N+1}{2}$$

Z.B.: $7+1/2=4$, d.h., der Median ist der Mp des 4. Wertes in der Rangfolge der Werte, im obigen (ersten) Beispiel also 7.
Den Median kann man auch bei schiefen Verteilungen mit mehr oder weniger starken Extremen verwenden (vgl. Abb. 11). Seine Anwendung setzt ferner nur Rangskalencharakter der Beobachtungsdaten voraus. Besonders pädagogisch wichtig ist in diesem Zusammenhang der Hinweis, daß Schulnoten (Zensuren) sehr häufig nur Meßwerte auf Rangskalenniveau darstellen. Somit dürfen – meßtheoretisch betrachtet – Zensuren entgegen weitverbreiteter Praxis *nicht* arithmetisch gemittelt werden; der adäquate Mittelwert wäre vielmehr hier der Median.
Sofern Frequenzverteilungen vorliegen, wird man den Median zweckmäßig über die *kumulierte* Häufigkeitsverteilung (cum f) ermitteln. Dieses Vorgehen soll wiederum anhand unseres Standardbeispiels (s. S. 95f.) erläutert werden.

Berechnung

a) Nach Erstellung der Frequenztabelle werden die kumulierten Frequenzen (cumf) ermittelt, d.h. die Frequenzen (f) fortlaufend addiert. Anschließend kann man noch die cumf-Werte – auf der Basis von N = 100% – prozentuieren und somit cumf% berechnen, was besonders für die spätere Quartilbildung von Bedeutung ist (s. S. 112).

X̄	f	cumf	cumf%
8–10	1	1	2
11–13	2	3	6
14–16	1	4	8
17–19	3	7	14
20–22	3	10	20
23–25	5	15	30
26–28	6	21	42
29–31	9	30	60
32–34	8	38	76
35–37	4	42	84
38–40	3	45	90
41–43	2	47	94
44–46	2	49	98
47–49	1	50	100

N = 50

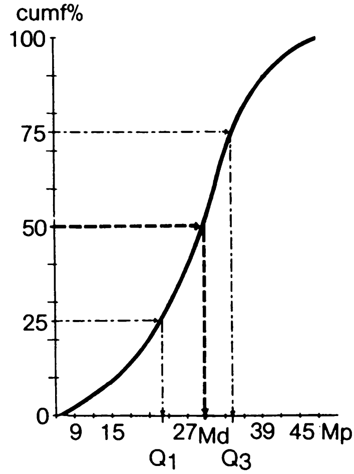


Abb. 14: Ogive oder Summenkurve (Kurve der cumf- bzw. cumf%-Werte)

b) Die Zahlen in der cumf-Spalte geben an, wie viele Fälle (absolut) – entsprechend in der cumf%-Spalte: wieviel % der Fälle – jeweils unter der oberen Grenze des kritischen Klassenintervalls liegen. So liegen laut der von uns ermittelten kumulierten Häufigkeitsverteilung z.B. 15 Fälle unter dem Wert 25,5 oder 45 Fälle unter dem Wert 40.5 (s. obige cumf-Tabelle).

c) Der Median muß nun gemäß der auf Seite 106 angeführten Definition zwischen 29 und 31 RP liegen. Seine genaue Bestimmung ist nur durch *Interpolation* möglich. Dazu dient uns folgende Formel:

$$\text{Med} = \text{Md} = \text{unt. Gr.} + \frac{\frac{N}{2} - \text{fo}}{\text{fi}} \cdot i$$

Dabei bedeuten:

unt. Gr. = untere Grenze des kritischen Intervalls, in dem der Md liegen muß,
fo = alle Fälle „oberhalb“ (graphisch gesehen) des kritischen Intervalls, in unserem Rechenbeispiel also 21 (abzulesen in der cumf-Spalte!),
fi = alle Fälle „innerhalb“ des kritischen Intervalls, in unserem Beispiel 9 (abzulesen in der f-Spalte!),
i = Intervallbreite,
N = Anzahl aller Fälle bzw. Maßzahlen (= Stichprobengröße).

Für unser Rechenbeispiel ergibt sich demnach folgender Medianwert:

$$\text{Med} = 28.5 + \frac{25 - 21}{9} \cdot 3 = 29.82$$

Sofern man die relativierten cumf-Werte, also die cumf%-Werte, graphisch in die Form der sog. Ogive übersetzt, kann man den Median auch *graphisch* bestimmen (s. Abb. 14). Laut Definition ist ja der Median mit dem 50. PR (Prozentrang) identisch (vgl. Kap. 3.5.1).

Voraussetzungen und Einschränkungen

a) Der Md ist der adäquate Mittelwert für Meßdaten auf Rangskalenniveau.
b) Da der Md von Extremwerten unbeeinflußt bleibt, kann er auch bei unregelmäßigen und schiefen (anormalen) Verteilungen verwendet werden.
c) Die Mittelung von Schulnoten, z.B. zur Berechnung von Zeugniszensuren, sollte wegen des Rangdatencharakters von Zensuren über den Median – und *nicht* arithmetisch – erfolgen.

3.3.3. Arithmetisches Mittel oder Durchschnitt

Das arithmetische Mittel oder Durchschnittsmaß (mit M bzw. μ [1] abgekürzt) ist definiert als Quotient aus der Summe ($\sum$) der Maßzahlen (X) dividiert durch die Anzahl der Fälle (N) oder formelhaft:

$$M = \mu = \frac{\sum X}{N}$$

Mathematisch exakt müßte man eigentlich schreiben: $\sum_{i=1}^{i=n} X_i$, d.h. Summe aller Maßzahlen von X_1 über X_i bis X_n, wobei i jeden beliebigen Score (X) bezeichnet. Bezogen auf die Meßdaten 5 5 6 7 9 10 12 wäre demnach:

$$M = \frac{\sum X}{N} = \frac{5+5+6+7+9+10+12}{7} = \frac{54}{7} = 7.71$$

Sofern es sich um eine *Frequenzverteilung* oder gruppiertes Datenmaterial handelt, benutzt man zweckmäßig folgende Formel:

$$M = \mu = \frac{\sum f \cdot X}{N} \text{ bzw. } M = \mu = \frac{\sum f \cdot \bar{X}}{N}$$

Berechnung

a) Die Häufigkeitsverteilung wird in der gewohnten Weise erstellt.
b) Dann wird der Ausdruck $f \cdot X$ berechnet, indem die einzelnen Frequenzen mit den jeweiligen Mittelpunkten der Intervalle multipliziert werden.
c) Schließlich wird die Summe $f \cdot X$ gebildet und in die obige Formel eingetragen. Siehe nachstehende *linke* Tabelle.

1 Sofern der Mittelwert eine Statistik darstellt – was meistens der Fall ist –, wird das Symbol M verwendet, andernfalls (bei einem Parameter) verwendet man gewöhnlich den griechischen Kleinbuchstaben μ (my).

Mp d. Interv.	f	f·X
9	1	9
12	2	24
15	1	15
18	3	54
21	3	63
24	5	120
27	6	162
30	9	270
33	8	264
36	4	144
39	3	117
42	2	84
45	2	90
48	1	48
$i = 3$	$N = 50$	$\sum f \cdot X = 1464$

Mp d. Interv.	f	x′	f·x′
9	1	−7	−7
12	2	−6	−12
15	1	−5	−5
18	3	−4	−12
21	3	−3	−9
24	5	−2	−10
27	6	−1	−6
30	9	0	0
33	8	+1	8
36	4	+2	8
39	3	+3	9
42	2	+4	8
45	2	+5	10
48	1	+6	6
$i = 3$	$N = 50$		$\sum fx' = -12$

Für unser Rechenbeispiel ergibt sich demnach:

$$M = \frac{\sum fX}{N} = \frac{1464}{50} = 29.28 \text{ RP}$$

Die gezeigte Methode ist somit wesentlich rationeller als die Berechnung anhand der Originalformel ($M = \sum X/N$); statt 50 Maßzahlen einzeln zu addieren mußten wir nur 14 Werte multiplizieren und deren Summe dann durch die Zahl 50 dividieren. Die Durchschnittsleistung unserer Untersuchungsgruppe beträgt somit 29.28 RP im PSB-Subtest 1 + 2. Freilich ist auch dieses Verfahren u. U. noch recht langwierig, besonders wenn keine Rechenmaschine zur Verfügung steht und das Datenmaterial umfangreicher ist. In solchen Fällen empfiehlt sich die sog. Kurzmethode oder *Methode des angenommenen Mittelwertes* zur Herstellung des arithmetischen Mittels. Am gleichen Datenmaterial soll – in Gegenüberstellung zu der soeben besprochenen Methode – deren Anwendung expliziert werden. Dazu wird folgende Formel benötigt:

$$M = \mu = aM + i\,\frac{\sum fx'}{N}$$

Hierbei bedeuten:

aM = angenommener oder hypothetischer Mittelwert,
i = Intervallbreite,
$\sum fx'/N$ = Korrekturwert beziehungsweise
$\sum fx'$ = Summe in bezug auf die Abweichungen vom hypothetischen Mittelwert,
N = Samplegröße.

Berechnung

a) Zunächst wird wieder die Häufigkeitsverteilung erstellt.
b) Dann wird der hypothetische Mittelwert festgelegt: Nach einer Faustregel wird man diesen in der Maßzahlklasse mit den höchsten Frequenzen annehmen, in unserem Rechenbeispiel also beim Mp = 30 (siehe *rechte* Tabelle oben).

c) Nun wird die Rohwertskala (ganz links in der Tabelle) durch eine sog. Substitutionsskala ersetzt (Spalte: x'), wobei der Wert 0 immer den aM markiert; darunter vs. darüber werden die Abweichungen vom aM unter Beachtung der Vorzeichen notiert.
d) Schließlich werden die f-Werte mit den *neuen* Maßzahlen der Substitutionsskala x' multipliziert, deren Summe (Summe der positiven und der negativen Abweichungen vom aM) durch N dividiert – bei Intervallbildung ist ferner noch i zu berücksichtigen! – und als Korrekturwert zum aM addiert oder von ihm abgezogen. Der Korrekturwert fällt um so kleiner aus, je näher wir mit unserer Mittelwertschätzung (aM) am tatsächlichen M liegen und umgekehrt.

In bezug auf unser Rechenbeispiel ermitteln wir somit (s. Tabelle auf S. 109):

$$M = aM + i\,\frac{\sum fx'}{N} = 30 + 3\left(-\frac{12}{50}\right) = 30 - 0.72 = 29.28 \text{ RP}$$

Das Ergebnis ist also nach beiden Berechnungsmethoden dasselbe, doch ist die Kurzmethode zumindest bei umfangreicheren Datenmengen das rationellere Verfahren.

3.3.4. Exkurs: Gewogener arithmetischer Mittelwert

Zuweilen liegen bereits Mittelwerte aus verschieden großen Untersuchungsgruppen vor, die zu einem einzigen – für alle (Unter-)Gruppen gültigen – Mittelwert M umgerechnet werden sollen. Zweifellos würde man dabei folgenschwere Fehler begehen, wenn man – die unterschiedlichen Samplegrößen außer acht lassend – hier einfach arithmetisch mitteln würde, z. B.:

Gruppe 1: $N_1 = 20$ $M_1 = 15$
Gruppe 2: $N_2 = 50$ $M_2 = 25$

$$\text{Also: } M = \frac{15 + 25}{2} = 20$$

(*falsche* Methode!)

Bei der Zusammenfassung zweier oder mehrerer Mittelwerte muß vielmehr dem Mittelwert der größeren Gruppe(n) auch ein größeres Gewicht zugemessen werden. Deshalb dient der Berechnung des gewogenen arithmetischen Mittelwertes (Mgew) folgende allgemeine Beziehung:

$$Mgew = \frac{N_1 M_1 + N_2 M_2 + N_3 M_3 \ldots + N_k M_k}{N_1 + N_2 + N_3 \ldots + N_k}$$

Anwendungsregeln

a) Das arithmetische Mittel darf nur bei Daten auf Intervall- oder Rationalskalenniveau verwendet werden.
b) Die Beobachtungs- bzw. Meßdaten sollen normal oder wenigstens in etwa symmetrisch verteilt sein. Im Zweifelsfalle wäre der Chi²-Test zur Überprüfung der Normalität (vgl. Kap. 5.4.1.3) einzusetzen.
c) Sind die genannten Voraussetzungen erfüllt, sollte man das arithmetische Mittel verwenden, da diese Methode das präziseste Verfahren zur Ermittlung und Beschreibung der zentralen Tendenz darstellt.
d) Sind die genannten Voraussetzungen – ganz oder teilweise – nicht erfüllt, so muß man auf den voraussetzungsärmeren Medianwert ausweichen. Dies wird insbesondere bei kleineren Stichproben häufiger der Fall sein.

Übungsbeispiele zu Kap. 3.3

Nr. 3/3
Aus den in der ungeordneten Urliste auf Seite 105 (Nr. 3/1) aufgeführten Rohwertdaten sollen folgende Mittelwerte errechnet werden: (grober) Mod, Med und M (unter Verwendung beider Berechnungsmethoden).

Nr. 3/4
Aus den Daten nachstehender Frequenztabelle sind Med und M zu berechnen sowie die fehlenden cumf- und cumf%-Werte in die Tabelle einzutragen.

$\bar{X}$	f	cumf	cumf%
50–51	1		
52–53	3		
54–55	2		
56–57	4		
58–59	5		
60–61	7		
62–63	6		
64–65	4		
66–67	3		
68–69	2		
70–71	2		
	N = 39		

3.4. Maße der Streuung (Variabilitätsmaße)

Mit Hilfe der verschiedenen Mittelwerte lassen sich Häufigkeitsverteilungen schon recht gut statistisch beschreiben. Freilich können sich empirische Verteilungen trotz gleicher Mittelwerte immer noch beträchtlich unterscheiden (vgl. Abb. 10). Für eine vollständige Charakterisierung beobachteter Häufigkeiten sind also weiterhin Angaben über die Streuung oder Dispersion notwendig. Wir beginnen mit unserer Erörterung der Streuungsmaße wieder mit den voraussetzungsärmeren Methoden.

3.4.1. Absolute Dispersion (Dispersionsspanne oder Streuungsbreite)

Die absolute Dispersion spielte früher schon eine Rolle (s. S. 95). Damit ist die Differenz zwischen dem niedrigsten (X_{min}) und dem höchsten (X_{max}) Wert einer Verteilung angesprochen. Dieses absolute Dispersionsmaß ist jedoch stark von der Gruppengröße abhängig; je größer die Stichprobe ist, desto größere Variabilität ist zu erwarten. Ferner ist die absolute Dispersion sehr empfindlich in bezug auf Verteilungsextreme. Deshalb ist es wünschenswert, ein relatives Streuungsmaß zur Hand zu haben, das die genannten Nachteile vermeidet und somit bessere Vergleichbarkeit gewährleistet.

3.4.2. Interquartildifferenz

Die Bestimmung der Interquartildifferenz erfolgt über die Quartile Q_1 und Q_3. Diese können graphisch (s. Abb. 14) oder rechnerisch – analog zur Md-Berechnung (s. S. 106f.) – ermittelt werden. Während der Md oder Q_2 mit dem 50. Perzentil zusammenfällt, sind die Quartile Q_1 mit dem 25. Perzentil und Q_3 mit dem 75. Perzentil identisch. Die entsprechenden *Interpolationsformeln* lauten hier:

$$Q_1 = \text{unt. Gr.} + \frac{N/4 - fo}{fi} \cdot i; \quad Q_3 = \text{unt. Gr.} + \frac{3 \cdot N/4 - fo}{fi} \cdot i$$

Für unser Rechenbeispiel ermitteln wir somit – unter Verwendung der kumulierten Frequenzverteilungstabelle auf S. 107 – folgende Werte:

$$Q_1 = 24.0 \text{ RP} \quad \text{und} \quad Q_3 = 34.32 \text{ RP}$$

Die Differenz zwischen den beiden Punkten ergibt nun die sog. *Interquartildifferenz* (abgekürzt: IQD):

$$\text{IQD} = 34.32 - 23.0 = 11.32 \text{ RP}$$

Dies bedeutet, daß die mittleren 50% unserer Stichprobe über 11.32 RP-Einheiten verteilt sind. Sie haben nicht weniger als 23 und nicht mehr als rd. 34 RP erzielt. Je größer die IQD ausfällt, desto heterogener ist demnach eine Gruppe und umgekehrt.

3.4.3. Mittlerer Quartilabstand (Quartilabweichung)

Ein weiteres Streuungsmaß auf Ordinalskalenniveau ist die Quartilabweichung oder das sog. mittlere Quartil, auch mittlerer Quartilabstand genannt (abgek.: Q). Die Quartilabweichung ist definiert als der durchschnittliche Abstand der beiden Quartilwerte vom Median. Die *Berechnungsformel* lautet:

$$Q = \frac{(Q_3 - \text{Md}) + (\text{Md} - Q_1)}{2} \quad \text{oder} \quad Q = \frac{Q_3 - Q_1}{2}$$

Die – häufig verwendete – Quartilabweichung wird auch als „wahrscheinliche Abweichung“ bezeichnet, weil die Wahrscheinlichkeit gleich groß ist, daß ein zufällig herausgegriffenes Individuum innerhalb der Grenzen von Q_1 und Q_3 liegt (50%) oder außerhalb davon (je 25% oder zusammen wieder 50%).
Bei unserem Rechenbeispiel ist Q = 5.66 RP.

Voraussetzungen und Einschränkungen

a) Die Verwendung der Quartilmaße setzt mindestens Ordinalskalenniveau der Erhebungsdaten voraus.
b) Da die Quartilabweichung (Q) auf den Median – und nicht auf das arithm. Mittel – bezogen ist, kann das Streuungsmaß Q auch bei anormaler (bimodaler, schiefer usw.) Verteilung berechnet werden.
c) Die Anwendung von Q ist auch bei Maßzahlklassen (Intervallbildung) möglich, ohne daß nennenswerte Informationsverluste zu erwarten sind.

3.4.4. Mittlere Variation (Durchschnittliche Abweichung)

Ausgangspunkt dieser und der folgenden Methode(n) ist der arithmetische Mittelwert. Somit erfordert die Verwendung der sog. mittleren Variation oder mittleren Varianz (MV) in bezug auf die Meßdaten mindestens Intervallskalenniveau. Ferner wird – im Gegensatz zu den Quartilmaßen – Normalverteilung oder wenigstens symmetrische Datenverteilung vorausgesetzt.

Berechnung

MV berechnet sich nach folgender Formel:

$$MV = \frac{\sum |X - M|}{N} \quad \text{bzw.} \quad MV = \frac{\sum (f \cdot |X - M|)}{N},$$

wobei man häufig – vereinfachend – statt X – M auch x schreibt. Die „Absolut"-Zeichen (senkrechten Striche) besagen, daß es sich bei den einzelnen Abweichungen von M um absolute Beträge handelt, also die Vorzeichen *nicht* berücksichtigt werden. Anhand unseres Datenbeispiels (s. S. 95 f. bzw. 109) sei das Vorgehen im einzelnen aufgezeigt (vgl. nachstehende Tabelle).

Mp	f	$\|X-M\|$	$f \cdot \|x\|$
9	1	20.3	20.3
12	2	17.3	34.6
15	1	14.3	14.3
18	3	11.3	33.9
21	3	8.3	24.9
24	5	5.3	26.5
27	6	2.3	13.8
30	9	0.7	6.3
33	8	3.7	29.6
36	4	6.7	26.8
39	3	9.7	29.1
42	2	12.7	25.4
45	2	15.7	31.4
48	1	18.7	18.7
	N = 50		$\sum f\|x\| = 335.6$

$$MV = \frac{\sum (f \cdot |x|)}{N} = \frac{335.6}{50} = 6.71 \text{ RP}$$

Das Ergebnis bedeutet, daß die durchschnittliche Abweichung der Individualwerte (Testleistung der einzelnen Pbn) vom arithm. Mittelwert 6,71 RP beträgt.

Voraussetzungen und Einschränkungen

a) Mindestens Intervallskalenniveau und Normal- bzw. symmetrische Verteilung der Meßdaten bilden die Voraussetzung zur Anwendung der MV.
b) Das heute relativ selten verwendete Streuungsmaß der MV ist weniger zuverlässig als das der Varianz und Standardabweichung. Es ist jedoch informationsreicher als die nur Rangskalenniveau erfordernden Q-Maße.

c) Im Zusammenhang unserer Methodenerörterung ist die MV vor allem aus didaktischen Gründen – zur Einführung in die Standardmaße der Variabilität (Varianz und Standardabweichung) – interessant.

3.4.5. Varianz

Neben der sog. Standardabweichung (σ bzw. s) ist die Varianz (σ^2 bzw. s^2) das wohl berühmteste Streuungsmaß. Die *Varianz,* abgekürzt σ^2 (Parameter) bzw. s^2 (Statistik) bezeichnet, ist definiert als die durchschnittliche Abweichung der *quadratischen* Differenzen vom arithm. Mittelwert.

Berechnung

Zur Berechnung der Varianz (σ^2 bzw. s^2) wird meistens folgende Formel benutzt:

$$\sigma^2 = s^2 = \frac{\sum (X - M)^2}{N}$$

bzw. (bei Frequenzverteilung)

$$\sigma^2 = s^2 = \frac{\sum f \cdot (X - M)^2}{N}$$

Diese Summe bildet die Varianz einer Verteilung. Sie ist vor allem im Hinblick auf die sog. Varianzanalyse von Bedeutung (vgl. Kap. 5.3.4).

Voraussetzungen und Einschränkungen

a) Die Berechnung der Varianz setzt Intervall- oder Rationalskalenniveau der Meßdaten voraus.
b) Ebenso verlangt die Anwendung des Varianzmaßes Normalverteilung, zumindest unimodale und symmetrische Verteilung der Daten.
c) Die vielfach als nachteilig empfundene verzerrte (quadratische, nicht lineare) Angabe der Abweichungen kann auch positiv interpretiert werden: Größere Abweichungen vom arithm. Mittelwert erhalten auf diese Weise größeres Gewicht.

3.4.6. Standardabweichung

Wegen der genannten Verzerrung ist es wünschenswert, die Varianz auf ein lineares Maß zurückzuführen. Dies ist durch Radizieren möglich.

Berechnung

Die *Standardabweichung,* abgekürzt mit σ (Parameter) bzw. s (Statistik) bezeichnet, ist folgendermaßen definiert:

$$\sigma = s = \sqrt{\frac{\sum (X - M)^2}{N}}$$

bzw. (bei Frequenzverteilung)

$$\sigma = s = \sqrt{\frac{\sum f \cdot (X - M)^2}{N}}$$

Bei größeren Datenmengen bzw. Frequenzverteilung ist die Benutzung obiger Berechnungsformeln u. U. (z. B. ohne die Hilfe einer Rechenmaschine) sehr aufwendig. Für solche Fälle empfiehlt sich – analog zur Methode des aM (angenommenen Mittelwertes) – wiederum eine *Kurzmethode* zur Bestimmung von σ bzw. s, wozu folgende Formel dient:

$$\sigma = s = i\sqrt{\frac{\sum f \cdot x'^2}{N} - \left(\frac{\sum f x'}{N}\right)^2}$$

Anhand unseres Rechenbeispiels sei nun das Vorgehen im einzelnen dargestellt (s. nachstehende Tabelle):

Mp	f	X − M = x	x^2	$f \cdot x^2$
9	1	20.3	412.09	412.09
12	2	17.3	299.29	598.58
15	1	14.3	204.49	204.49
18	3	11.3	127.69	383.07
21	3	8.3	68.89	206.67
24	5	5.3	28.09	140.45
27	6	2.3	5.29	31.74
30	9	0.7	0.49	4.41
33	8	3.7	13.69	109.52
36	4	6.7	44.89	179.56
39	3	9.7	94.09	282.27
42	2	12.7	161.29	322.58
45	2	15.7	246.49	492.98
48	1	18.7	349.69	349.69
N = 50				$fx^2 = 3718.10$

Mp	f	x'	x'^2	$f \cdot x'$	$f \cdot x'^2$
9	1	−7	49	−7	49
12	2	−6	36	−12	72
15	1	−5	25	−5	25
18	3	−4	16	−12	48
21	3	−3	9	−9	27
24	5	−2	4	−10	20
27	6	−1	1	−6	6
30	9	0	0	0	0
33	8	1	1	8	8
36	4	2	4	8	16
39	3	3	9	9	27
42	2	4	16	8	32
45	2	5	25	10	50
48	1	6	36	6	36
N = 50				$\sum fx' = -12$	$\sum fx'^2 = 416$

$$s = \sqrt{\frac{\sum f \cdot (X - M)^2}{N}}$$

$$s = \sqrt{\frac{3718.10}{50}} = \sqrt{74.36} = 8.63 \text{ RP}$$

$$s = i \cdot \sqrt{\frac{\sum f \cdot x'^2}{N} - \left(\frac{\sum f \cdot x'}{N}\right)^2}$$

$$s = 3 \cdot \sqrt{\frac{416}{50} - \left(\frac{-12}{50}\right)^2} = 3 \cdot \sqrt{8.32 + 0.0576}$$

$$s = 3 \cdot \sqrt{8.3776} = 3 \cdot 2.89 = 8.67 \text{ RP}$$

Entsprechend ist $s^2 = 74.36$ RP.

Die Ergebnisse der beiden Methoden sind also praktisch identisch, jedoch ist der Rechenaufwand – ohne die Hilfe einer Rechenmaschine – nach der Kurzmethode wesentlich geringer. Die Effizienz der Kurzmethode erweist sich bei umfangreicherem Datenmaterial noch deutlicher als bei unserem Rechenexempel.

Voraussetzungen und Einschränkungen

a) Es gelten die bereits auf Seite 114 für die Varianzberechnung genannten Kautelen zu Pkt. a) und b).

b) Bezüglich der Kurzmethode zur Berechnung von σ (Sigma) bzw. s müssen prinzipiell dieselben Regeln beachtet werden wie schon bei der analogen aM-Methode (s. S. 109 f.).

3.4.7. Variabilitätskoeffizient

Der Variabilitäts- oder Variationskoeffizient dient dem von den jeweiligen Maßeinheiten unabhängigen Vergleich mehrerer Streuungswerte. So kann man etwa nicht ohne weiteres die Varianz der Körpergröße und die Varianz des Körpergewichts bei 10jährigen Schulkindern miteinander vergleichen, da ja den entsprechenden Meßdaten unterschiedliche Meßskalen (kg vs. cm) zugrunde liegen. Anderes Beispiel: Man will entscheiden, ob sich Erwachsene hinsichtlich bestimmter Merkmale (z.B. Größe, Gewicht u.ä. oder Intelligenz, Konzentration, Leistung u.dgl.m.) stärker ähneln als Kinder oder Jugendliche einer bestimmten Altersstufe. Solche Vergleiche sind nur mit Hilfe des Variabilitätskoeffizienten durchführbar. Der Variabilitätskoeffizient (V) ist folgendermaßen definiert:

$$V = \frac{\sigma}{M} \cdot 100$$

Der Variabilitätskoeffizient bringt also in relativierter Form zum Ausdruck, wieviel (%) des arithm. Mittelwertes (M) die Standardabweichung (σ oder s) ausmacht.

Rechenbeispiel

Es soll untersucht werden, ob 10jährige Schüler stärker hinsichtlich der Körpergröße oder hinsichtlich des Körpergewichts variieren. Folgende Statistiken sind bekannt: $M_{\text{Größe}} = 115$ cm, $s_{\text{Größe}} = 5.6$ cm; $M_{\text{Gewicht}} = 25$ kg, $s_{\text{Gewicht}} = 3.0$ kg.

Ausrechnung: $V_{\text{Größe}} = \frac{6.5}{115} \cdot 100 = 5.6 \quad V_{\text{Gewicht}} = \frac{3.0}{25} \cdot 100 = 12.0$

Demnach sind die Vpn variabler im Hinblick auf das Gewicht, d.h., die Körpergröße wirkt stärker homogenisierend als das Gewicht. Noch präziser wird unsere Auskunft, wenn wir einen Quotienten bilden:

$$\frac{5.6}{12.0} \cdot 100 = 47$$

Die Variabilität der Schülergröße beträgt demnach 47% der Variabilität bezüglich des Schülergewichts, ist also nur etwa halb so groß.

Voraussetzungen und Einschränkungen

a) Der Variabilitätskoeffizient (V) darf strenggenommen nur bei Daten auf Rational- oder Verhältnisskalenniveau berechnet werden.
b) In den Sozialwissenschaften wird V jedoch auch dann berechnet, wenn kein absoluter, sondern nur ein konventionell akzeptierter Nullpunkt vorliegt, z.B. bei Intelligenztests, in der Leistungsmessung usw.

Übungsbeispiele zu Kap. 3.4

Nr. 3/5
Aus den auf Seite 105 (Nr. 3/1) angeführten Rohwertdaten sind folgende Streuungsmaße zu berechnen: absol. Dispersion, IQD und Q sowie MV, s^2 und s.
Nr. 3/6
Aus den Daten der auf Seite 111 (Nr. 3/4) wiedergegebenen Frequenztabelle sollen Q und s berechnet werden.

3.5. Das Normenproblem

Jedem Vergleich liegt mehr oder weniger explizit eine Normenmatrix, d.h. ein Maßstab zugrunde. Anschließend an die Erörterungen zur deskriptiven Statistik wollen wir hierauf noch kurz eingehen.
Im Hinblick auf die pädagogisch-psychologische Praxis, etwa im Rahmen diagnostischer Fragestellungen (vgl. Heller 1973), sind zwei Arten von Normenskalen von hervorragender Bedeutung: die *Perzentil-* oder *Prozentrang*skala und die sog. *Standardwert*skalen, deren Prototyp die *z-Skala* darstellt.

3.5.1. Prozentränge (Perzentile)

Wir haben bereits einige Perzentile kennengelernt, nämlich die Quartile Q_1 (= 25. Perzentil), Q_2 bzw. Md (= 50. Perzentil) und Q_3 (= 75. Perzentil). Die Perzentilskala ist eine Rangskala und reicht von 0 bis 100 Punkten. Allgemein kann jeder beliebige Punkt auf der Prozentrangskala, d.h. jedes beliebige Perzentil, analog zu den Quartilen folgendermaßen berechnet werden:

$$P_p = \text{unt. Gr.} + \left(\frac{pN - fo}{fi}\right) \cdot i$$

Hierbei bedeuten:
unt. Gr. = die untere Grenze des kritischen Intervalls (s. S. 107),
N = Anzahl der Fälle,
p = die betr. Proportion, also beim 10. Perzentil (P) gleich 0.1, beim 20. Perzentil gleich 0.2 usw.,
fo = Anzahl der Fälle „oberhalb“ (graphisch gesehen) vom kritischen Intervall,
fi = Anzahl der Fälle innerhalb des kritischen Intervalls,
i = Intervallbreite.

Anwendung und Voraussetzung

a) Über cum f und cum f% (s. S. 106f.) läßt sich jederzeit eine Rohwertskala bzw. Rohpunktverteilung in die Prozentrangskala überführen. Die fehlenden, d.h. in der empirischen Verteilung nicht aufgetretenen, Punkte können durch Interpolation oder graphisch bestimmt werden (vgl. Abb. 14).
b) Eine solche Rohwerttransformation setzt lediglich voraus, daß die Daten bzw. Vpn rangmäßig angeordnet werden können.
c) Normalverteilung der Rohdaten ist hierfür nicht erforderlich.

3.5.2. Standardwertskala (z-Skala)

Die Prozentrangskala (PR) ist eine sehr anschauliche Normenskala: Wer z.B. PR = 90 erreicht, ist besser als 90% seiner Bezugsgruppe vs. schlechter als (nur) 10%, wer PR = 35 erzielt, ist nur noch besser als 35% seiner Bezugsgruppe, also deutlich unterdurchschnittlich. Prozentränge repräsentieren freilich keine Intervallskalen, die Abstände der PR-Punkte sind also nicht gleich. Während sich im mittleren Skalenbereich relativ viele Einheiten zusammendrängen, erscheint die PR-Skala an den Enden überdehnt. M.a.W.: Große PR-Differenzen im mittleren Skalenbereich repräsentieren relativ kleine tatsächliche (Leistungs-)Unterschiede, an den Skalenenden indizieren bereits sehr kleine PR-Differenzen schon verhältnismäßig große tatsächliche Differenzen. Auch lassen sich keine Durchschnittswerte auf der Basis von Perzentilen berechnen. Deshalb ist es oft wünschenwert, Normenskalen auf Intervallskalenniveau zur Verfügung zu haben. Der Prototyp einer solchen *Standardwert*skala ist die sog. z-Skala, die durch $M_z = 0$ (engl. *zero*) und $s_z = 1$ definiert ist. Die z-Skala gibt an, um wieviel Einheiten der Standardabweichung σ bzw. s irgendein Individual*score* (RP) über oder unter dem Mittelwert M seiner Bezugsgruppe liegt (s. S. 178, Abb. 26). Siehe auch nachstehende Berechnungsformel.
Eine solche Aussage entspricht dem statistischen Normbegriff besser als die PR-Skala, da hier M und s Berücksichtigung finden. Im Anhang dieses Buches sind nun die einzelnen z-Werte tabelliert zusammen mit den entsprechenden Flächenprozenten beigegeben (vgl. Tab. II sowie die entsprechenden Erläuterungen dazu auf S. 176ff.). Mit Hilfe der Beziehung

$$z = \frac{X - M}{\sigma} = \frac{x}{s}$$

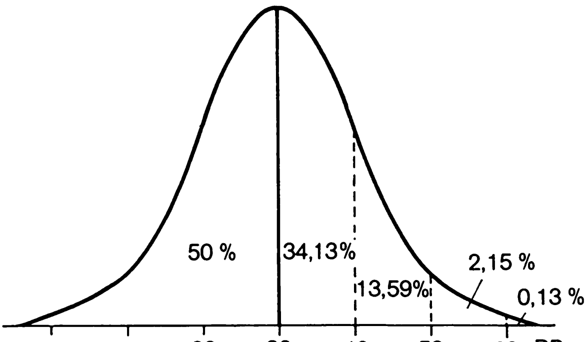
kann jeder Rohwert in einen z-Wert – unter der Voraussetzung der *Normalverteilung* – umgewandelt werden. Ferner setzt die Verwendung von z- bzw. Standardwerten immer Rohwerte auf *Intervallskalenniveau* oder darüber voraus.

3.5.3. Transformation einer Rohwertskala in Prozentränge und Standardwerte

Beispielhaft wird in Abb. 15 – entnommen Heller 1973, S. 77 – die Umwandlung von Rohpunkten (RP) in Prozentränge (PR) sowie in die Standardwertskala z dargestellt. Außerdem können durch lineare Transformation der z-Werte weitere Standardwerte qua Normenskalen auf Intervallebene zum Vergleich herangezogen werden. Gängige Normenskalen dieser Art sind z.B.:

T-Werte ($M_T = 50$, $s_T = 10$)
C-Werte oder Centile ($M_C = 5$, $s_C = 2$)
(Abweichungs-)IQ-Werte ($M_{IQ} = 100$, $s_{IQ} = 15$)

Das Beispiel veranschaulicht, daß PR = 50 und z = 0 bzw. T = 50, C = 5 und IQ = 100 identische Leistungs- oder Merkmalsgrade (durchschnittliche Testwerte von RP = 30) repräsentieren, analog deuten PR = 98 und z = + 2 bzw. T = 70, C = 9 und IQ = 130 überdurchschnittlich gute vs. PR = 2 und z = – 2 bzw. T = 30, C = 1 und IQ = 70 sehr schwache Testleistungen an. Siehe dazu ausführlicher Heller 1973, S. 75ff., 114ff. u. 222ff.



0	10	20	30	40	50	60	RP
0	2	16	50	84	98	100	PR
−3	−2	−1	0	+1	+2	+3	z
20	30	40	50	60	70	80	T
−1	1	3	5	7	9	11	C
55	70	85	100	115	130	145	IQ

Abb. 15: Transformation einer Rohwertskala in Prozentränge und Standardwerte

3.5.4. Normalisierung anormaler Verteilungen (Flächentransformation)

Häufig sind in der pädagogischen oder psychologischen Diagnostik die Voraussetzungen für die Verwendung von Standardwertnormen (z. B. *Normal*verteilung der Testrohpunkte, der Zensuren oder anderer Lehrerurteile) nicht erfüllt. In diesen Fällen muß man entweder auf Prozentränge als Normmaßstab ausweichen, oder man muß den – allerdings rechenaufwendigen – Weg der Flächentransformation und somit die Möglichkeit der Normalisierung anormaler Verteilungen einschlagen. Voraussetzung dafür ist, daß die ursprüngliche (empirische) Verteilung annähernd symmetrisch und unimodal ist, also keine allzu starken Unregelmäßigkeiten aufweist. Bei der Normalisierung macht man sich die in Kap. 5.1.2.5 aufgewiesenen Zusammenhänge zwischen z-Wert und Flächenanteil unter der Normalkurve zunutze, wobei die *Frequenzen* der empirischen Verteilung die Ausgangsbasis für die Berechnung abgeben. Anhand eines Beispiels sei wiederum das Vorgehen im einzelnen erläutert. Hierbei handelt es sich um RP-Ergebnisse von sog. Probearbeiten im Fach Rechnen, wie sie beispielsweise in Baden-Württemberg im Rahmen der Übertrittsauslese jährlich durchgeführt werden.

X	f	cum f	$\text{cum f}-\frac{f}{2}$	$\frac{\text{cum f}-\frac{f}{2}}{N}\,100$	z	T	z*
2	4	4	2.0	0.7	−2.46	25	−2.95
3	10	14	9.0	3.0	−1.88	31	−2.32
4	12	26	20.0	6.7	−1.50	35	−1.70
5	25	51	38.5	12.8	−1.14	39	−1.07
6	42	93	72.0	24.0	−0.71	43	−0.44
7	65	158	125.5	31.8	−0.47	45	−0.20
8	105	263	210.5	70.2	0.53	55	0.83
9	20	283	273.0	91.0	1.34	63	1.46
10	7	300	296.5	99.0	2.33	73	2.09

N = 300 M = 6.69 RP s = 2.74 RP

Anm.: Zum Vergleich wurden auch die nichtnormalisierten z-Werte z* (aufgrund von M und s berechnet!) mit angegeben.

Berechnung

a) Zunächst wird wieder die Frequenzverteilung tabelliert.
b) Dann wird cumf in der bekannten Weise gebildet. Bezieht man die cumf-Werte auf die Intervallmitten, so muß man jeweils die Hälfte der fi-Werte abziehen (vgl. 4. Tabellenspalte).
c) Schließlich werden die kumulierten Frequenzen prozentuiert (vgl. 5. Tabellenspalte).
d) Die so erhaltenen Perzentile werden nun mit Hilfe von Tab. II im Anhang dieses Buches in z-Werte transformiert. Das Ziel hierbei ist der *flächentransformierte* z-Wert, so daß man die betr. Perzentile jeweils zum mittleren Fall (50. PR) in Beziehung setzen, d. h. im kleineren (C) oder größeren (B) Flächenstück aufsuchen muß.
e) Die dazugehörigen z-Werte werden abgelesen und in die obige Tabelle eingetragen. Abschließend kann man die flächentransformierten z-Werte in T-Standardwerte transformieren, wobei man sich folgender Formel bedient:

$$T = z \cdot \sigma_T + M_T = z \cdot 10 + 50$$

Der Effekt dieser etwas aufwendigen Prozedur wird aus der graphischen Darstellung in Abb. 16 ersichtlich. Die schiefe Verteilung der originalen Werte wurde der Gaußschen Kurve angepaßt, d. h., die flächentransformierten Werte repräsentieren jetzt eine Normalverteilung im Sinne der Gaußschen Kurve. Dies geschah auf der Basis der *Häufigkeiten:* Die alte und die neue Häufigkeitsverteilung unterscheiden sich nur hinsichtlich der Säulenhöhe der einzelnen Maßzahlklassen, *nicht* hinsichtlich der Säulenflächen, die konstant blieben. Die Säulen unseres Histogramms wurden also unter Wahrung des Gesamtflächenanteils dem Verlauf der Glockenkurve angepaßt, wozu sie mehr oder weniger stark gestaucht oder gestreckt werden mußten. Die Flächenanteile unter der Kurve sind aber jeweils konstant geblieben.

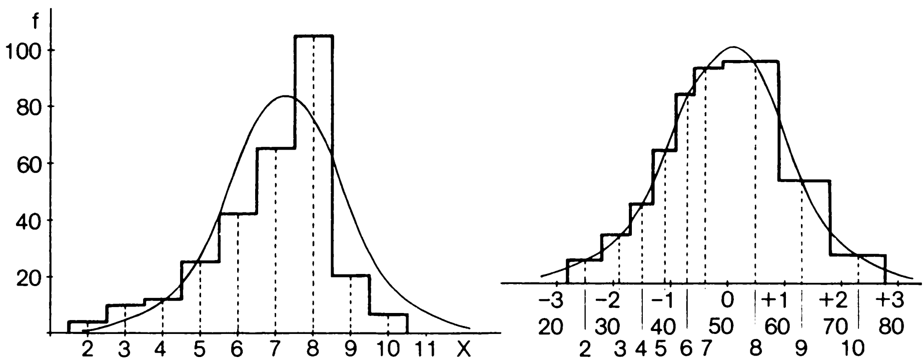


Abb. 16a: Verteilung der Originaldaten (X)

Abb. 16b: Normalisierte Verteilung (T-Werte)

Abb. 16: Transformation einer schiefen RP-Verteilung in eine Normalverteilung (Flächentransformation nach McCall)

Übungsbeispiel

Nr. 37

Es sei folgende Rohwertverteilung gegeben, die via Flächentransformation in eine Normalverteilung übergeführt werden soll:

X (Mp)	10	15	20	25	30	35	40	45	50	55
f	6	8	30	20	10	8	6	6	4	2

4. Korrelation und Regression

4.1. Methoden der Korrelationsrechnung

Ziel der Korrelationsstatistik ist es, Maße für den Grad des Zusammenhangs zwischen zwei Merkmalen (Variablen) zu liefern. Beide Merkmale werden an einer Gruppe von Versuchspersonen (Stichprobe) gemessen. Es existieren also für jede Versuchsperson zwei Meßwerte bzw. für die Stichprobe zwei Meßwertereihen (bivariable Verteilung). Ein numerisches Maß für den Grad des Zusammenhangs dieser Meßwertereihen ist für Merkmale, die auf Intervall- bzw. Ordinalskalenniveau gemessen werden, der Korrelationskoeffizient. Der Zusammenhang zwischen nominalskalierten Merkmalen wird durch einen Kontingenzkoeffizienten beschrieben. Der Korrelationskoeffizient kann einen Wert von -1 bis $+1$ annehmen. Die numerische Größe des Koeffizienten gibt an, wie eng der Zusammenhang zwischen zwei Merkmalen ist, während das Vorzeichen die Richtung des Zusammenhanges ausdrückt. Ein Korrelationskoeffizient von $+1.00$ bedeutet also einen vollständigen und gleichsinnigen Zusammenhang beider Merkmale. Hat eine Versuchsperson bei dem einen Merkmal einen hohen Wert, dann weist sie auch bei dem anderen einen hohen Wert auf. Ein Korrelationskoeffizient von .00 besagt, daß zwischen den beiden Merkmalen kein Zusammenhang besteht, während ein Koeffizient von -1.00 einen negativen Zusammenhang ausdrückt. Hier wird ein hoher Wert bei dem einen Merkmal zusammen mit einem niedrigen Wert bei dem anderen Merkmal auftreten und umgekehrt.
Gewöhnlich werden Korrelationskoeffizienten wie folgt interpretiert:

$$
\begin{aligned}
.00 \quad & = \text{kein Zusammenhang} \\
.00 \text{ bis} \leqq .40 & = \text{niedriger Zusammenhang} \\
.40 \text{ bis} \leqq .70 & = \text{mittlerer Zusammenhang} \\
.70 \text{ bis} < 1.00 & = \text{hoher Zusammenhang} \\
1.00 \quad & = \text{vollständiger Zusammenhang}
\end{aligned}
$$

Der Begriff des Zusammenhangs zwischen zwei Merkmalen darf aber nicht falsch aufgefaßt werden. Er meint *nicht*, daß zwischen den zwei Merkmalen (z. B. x und y) ein kausaler Zusammenhang besteht. Der Korrelationskoeffizient an sich läßt keine Aussage über ein Ursache-Wirkungs-Verhältnis zu, etwa in dem Sinne, daß das Merkmal x das Merkmal y bewirke. So kann man beispielsweise einen hohen positiven Korrelationskoeffizienten zwischen Intelligenz und Schulleistung nicht ohne weiteres so interpretieren, daß eben die Schulleistung aufgrund der Intelligenz zustande komme. Ein solcher Koeffizient besagt zunächst einmal nur, daß ein Schüler mit hoher (niedriger) Intelligenz auch meist eine gute (schlechte) Schulleistung zeigen wird. Für das Zustandekommen dieses Zusammenhanges gibt es aber eine Reihe alternativer Erklärungen, z. B. mag eine dritte Variable dafür verantwortlich sein.

Welches Verfahren man zur Ermittlung des Zusammenhanges zweier Merkmale anwendet, hängt davon ab, auf welchem Skalenniveau die Merkmale gemessen werden (s. folgende Tabelle).

	*Intervall*skala	*Ordinal*skala	*Nominal*skala
Intervall-skala	Produkt-Moment Korrelation (s. S. 123ff.) Partial-Korrelation (s. S. 130ff.)		
Ordinal-skala		Rangreihen-Korrelation (s. S. 137ff.) Konkordanz-Koeffizient (s. S. 140ff.)	
Nominal-skala	Zweizeilen-Korrelation (punktbiseriale, biseriale) (s. S. 133ff.)	Biseriale Rangkorrelation (s. S. 143ff.)	Punkt-Vierfelder-Korrelation (s. S. 146ff.) Kontingenz-Koeffizient (s. S. 148ff.) Cramér-Koeffizient (s. S. 151f.)

Abb. 17: Überblick über die behandelten Korrelationsmethoden

Liegen für beide Merkmale metrische Daten vor, haben sie also Intervallskalenniveau, dann wird die Produkt-Moment-Korrelation berechnet. Haben beide Merkmale Ordinalskalenniveau, liegen also Rangplätze vor oder lassen sich die Daten in Rangplätze verwandeln, dann können u.a. die Rangreihenkorrelation (Spearman) und der Konkordanzkoeffizient (Kendall) errechnet werden. Haben beide Merkmale nur Nominalskalenniveau, dann bestimmt man etwa die Punkt-Vierfelder-Korrelation oder den Kontingenzkoeffizienten. Welche Methoden angezeigt sind, wenn die beiden Merkmale verschiedenes Skalenniveau haben, läßt sich der Tabelle entnehmen.

4.1.1. Die Produkt-Moment-Korrelation (r) nach Pearson

(1) *Untersuchungsproblem*
Welcher Zusammenhang besteht zwischen zwei quantitativen (kontinuierlichen) Merkmalen, die an einer Stichprobe von Merkmalsträgern gemessen werden?
Es existieren also für jede Versuchsperson der Stichprobe zwei Meßwerte: ein

Meßwert bezüglich des Merkmals 1, ein zweiter bezüglich des Merkmals 2. Der Grad des Zusammenhanges zwischen den beiden Merkmalen ist zu ermitteln. Graphisch läßt sich dies Problem mit Hilfe einer Korrelationstabelle *(scatter diagram)* so darstellen:
Die in den Abb. 18 bis 20 eingetragenen Punktschwärme entstehen auf folgende Weise. Man trägt den Meßwert, den eine Versuchsperson bezüglich des Merkmals 1 hat, auf der x-Achse ab und errichtet in dem entsprechenden Punkt das Lot. Entsprechend trägt man auf der y-Achse den Meßwert derselben Versuchsperson bezüglich des Merkmals 2 ab und errichtet wiederum das Lot. Der Schnittpunkt der beiden Lote charakterisiert die Lage der Versuchsperson hinsichtlich der beiden Merkmale. Jeder Punkt des Punkteschwarmes stellt einen solchen Schnittpunkt zweier Koordinaten dar.

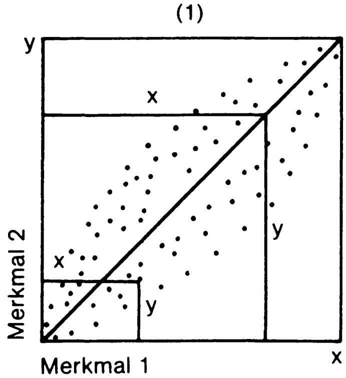


Abb. 18: Punkteschwarm bei einer positiven Korrelation

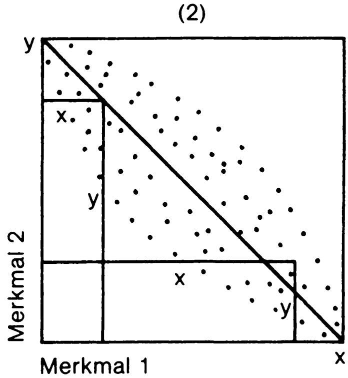


Abb. 19: Punkteschwarm bei einer negativen Korrelation

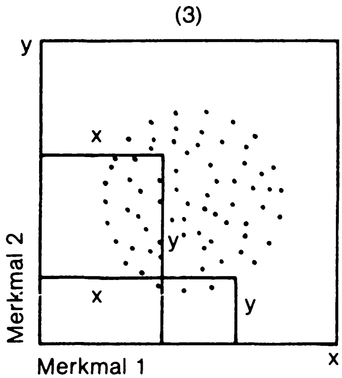


Abb. 20: Punkteschwarm bei fehlender Korrelation

Aus Abb. 18 kann man nun folgendes entnehmen. Hat die Versuchsperson einen niedrigen Wert bezüglich des Merkmals 1, dann hat sie auch einen niedrigen Wert bei Merkmal 2. Eine Versuchsperson mit einem hohen Wert bei Merkmal 1 hat auch einen hohen Wert bei Merkmal 2. Man spricht hier von einem positiven

Zusammenhang bzw. einer positiven Korrelation zwischen den beiden Merkmalen. Eine extrem hohe positive Korrelation (+1) liegt vor, wenn alle Koordinatenschnittpunkte auf die eingezeichnete Gerade (Regressionsgerade; s. Kap. 4.2) fallen.
Abb. 19 veranschaulicht eine negative Korrelation. Hat die Versuchsperson einen hohen Meßwert bei Merkmal 1, dann besitzt sie einen niedrigen Meßwert bei Merkmal 2. Ein niedriger Meßwert bei Merkmal 1 fällt zusammen mit einem hohen Meßwert bei Merkmal 2. Fallen alle Koordinatenschnittpunkte auf die eingezeichnete Gerade, dann liegt eine extrem negative Korrelation (−1) vor.
Abb. 20 zeigt den Fall einer fehlenden Korrelation. Bei einer Versuchsperson kann ein hoher Meßwert bei Merkmal 1 zusammen mit einem niedrigen bei Merkmal 2 auftreten, bei einer anderen Versuchsperson kann es genau umgekehrt sein.

(2) *Beispiele*

Beispiel A:

a) Versuchsplan
10 zufällig ausgewählten Realschülern wurde ein Intelligenztest und ein Rechentest vorgelegt. Es sollte untersucht werden, ob ein Zusammenhang besteht zwischen Intelligenzgrad und Rechenleistung.
Stichprobe: 10 zufällig ausgewählte männliche Schüler der 4. Klasse einer Realschule
Merkmal 1: Intelligenz (x)
Merkmal 2: Rechenleistung (y)

b) Hypothesen
Nullhypothese (H_0):
Zwischen Intelligenzgrad und Rechenleistung besteht kein Zusammenhang.
Alternativhypothese (H_1):
Zwischen Intelligenzgrad und Rechenleistung besteht ein von Null verschiedener Zusammenhang.
Signifikanzniveau: .05

c) Ergebnisse

N = 10

Schüler	Intelligenz x	Rechenleistung y	x^2	y^2	xy
1	112	48	12544	2304	5376
2	110	59	12100	3481	6490
3	115	50	13225	2500	5750
4	118	47	13924	2209	5546
5	120	60	14400	3600	7200
6	107	51	11449	2601	5457
7	99	42	9801	1764	4158
8	128	53	16384	2809	6784
9	114	45	12996	2025	5130
10	106	46	11236	2116	4876
	$\sum x = 1129$	$\sum y = 501$	$\sum x^2 = 128059$	$\sum y^2 = 25409$	$\sum xy = 56767$

d) Berechnung

Für die Berechnung des Produkt-Moment-Korrelationskoeffizienten (r_{xy}) gibt es eine Reihe von Formeln. Bei kleineren Stichproben und vor allem bei Verwendung einer Rechenmaschine eignet sich folgende Formel:

$$r_{xy} = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

Diese Formel geht direkt von den Meßwerten aus. In die Tabelle in Abschnitt c wurden bereits die notwendigen Größen eingetragen.

Der Korrelationskoeffizient schwankt zwischen -1 und $+1$. Setzen wir unsere Werte in die Formel ein, so ergibt sich:

$$r_{xy} = \frac{(10 \cdot 56767) - 1129 \cdot 501}{\sqrt{[10 \cdot 128059 - 1129^2][10 \cdot 25409 - 501^2]}} = .477$$

Zwischen Intelligenzgrad und Rechenleistung besteht also ein mittlerer positiver Zusammenhang.

e) Signifikanzprüfung

aa) Zur Prüfung des Produkt-Moment-Korrelationskoeffizienten gegen die Nullhypothese verwendet man die t-Verteilung (Tab. III). Der t-Wert berechnet sich nach folgender Formel:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}; \; df = N - 2$$ [1]

Es bedeuten:

r = Produkt-Moment-Korrelationskoeffizient
N = Stichprobengröße (Zahl der Meßwertpaare)

Man vergleicht den errechneten t-Wert mit dem in Tab. III unter Berücksichtigung des Signifikanzniveaus und der Zahl der Freiheitsgrade abgelesenen. Ist der errechnete Wert gleich oder größer als der abgelesene, dann können wir die Nullhypothese zurückweisen.

bb) Eine einfachere Prüfung auf Signifikanz ermöglicht Tab. VI. Dort sind die Höchstwerte für die Produkt-Moment-Korrelationskoeffizienten bei 1% und 5% Irrtumswahrscheinlichkeit aufgeführt, unter Berücksichtigung der Freiheitsgrade.

Für unsere Beispiele haben wir folgende t-Werte berechnet:

Beispiel A: $t = 1.536$, $p > .05$, $df = 8$
Beispiel B: $t = 8.12$, $p < .001$, $df = 118$

Der errechnete t-Wert im Beispiel A ist kleiner als der in Tab. III abgelesene. Die Nullhypothese muß also beibehalten werden. Die Rechnung für Beispiel B ergibt, daß die Nullhypothese zurückgewiesen werden kann.

Analoge Schlußfolgerungen können wir auch aufgrund der Tab. VI ziehen.

1 Zum Begriff „Freiheitsgrad" *(degree of freedom)*: Die Summe von vier unbekannten Zahlen sei 50, also $x_1 + x_2 + x_3 + x_4 = 50$. Wir können nun für die Werte x_1, x_2 und x_3 beliebige Zahlen frei wählen, etwa $x_1 = 15$, $x_2 = 12$, $x_3 = 20$. Damit ist die vierte Zahl (x_4) bereits festgelegt ($x_4 = 3$), sie ist also nicht mehr frei wählbar. Die Zahl der Freiheitsgrade beträgt hier df = 3. Freiheitsgrad meint also die Zahl frei verfügbarer Meßwerte, Maßzahlen usw. Die Zahl der Freiheitsgrade, die von der Zahl der zu schätzenden Parameter abhängig ist, wird für die verschiedenen Prüfverfahren jeweils angegeben.

(3) *Anwendungsregeln*

a) Die Meßwerte der Versuchspersonen bezüglich des Merkmals 1 (x-Werte) und des Merkmals 2 (y-Werte) werden in eine Tabelle so eingetragen, daß für jede Versuchsperson die beiden Meßwerte nebeneinanderstehen.
b) Die x- und y-Werte werden quadriert und (bei Handrechnung) in die Tabelle eingetragen.
c) Für jedes Meßwertepaar wird das Produkt xy gebildet.
d) Die Werte in den Spalten x, y, x^2, y^2 und xy werden jeweils addiert, und man erhält: $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ und $\sum xy$.
e) Die erhaltenen Größen werden in die Formel

$$r_{xy} = \frac{N \sum xy - (\sum x)(\sum y)}{\sqrt{[N \sum x^2 - (\sum x)^2][N \sum y^2 - (\sum y)^2]}}$$

eingetragen. Dabei ist auf den Unterschied zwischen $\sum x^2$ und $(\sum x)^2$ zu achten.

Beispiel B:

Steht keine Rechenmaschine zur Verfügung, dann kann der Produkt-Moment-Korrelationskoeffizient mit Hilfe eines Verfahrens berechnet werden, das auf der Methode des angenommenen Mittelwertes (s. Kap. 3.3) beruht.

a) Versuchsplan (nach Lienert 1961)
Ein neu konstruierter Gedächtnistest soll an einem bereits bewährten Gedächtnistest validiert werden. Beide Tests werden an 120 Versuchspersonen durchgeführt. Es soll nun festgestellt werden, in welchem Grade die Gedächtnisleistungen der Versuchspersonen in beiden Tests korrelieren.

b) Hypothesen
H_0: Es besteht kein Zusammenhang zwischen den Punktwerten im alten und neuen Gedächtnistest.
H_1: Es besteht ein bedeutsamer Zusammenhang zwischen den Punktwerten im alten und neuen Gedächtnistest.
Signifikanzniveau: .05

c) Ergebnisse
Die Ergebnisse werden in eine Korrelationstabelle eingetragen. Es empfiehlt sich im übrigen immer, auch bei Maschinenrechnung, solche Korrelationstabellen anzulegen, da sie einen ersten Überblick über den zu erwartenden Zusammenhang und die Verteilung der Meßwerte liefern.

X / Y	100 -109	110 -119	120 -129	130 -139	140 -149	150 -159	160 -169	170 -179	f_y	Y'	f_yY'	$f_y (y')^2$	Σf_{xy} X'y'
72.5								(12) / 1	1	3	3	9	12
70.5			(−2) / 1	/// 3	(6) /// 3	(16) //// 4	(12) // 2	(24) /// 3	16	2	32	64	56
68.5			(−4) //// 4	~~////~~ ~~////~~ / 11	(6) ~~////~~ / 6	(6) /// 3	(6) // 2	(8) // 2	28	1	28	28	22
66.5		// 2	~~////~~ //// 9	~~////~~ ~~////~~ / 11	~~////~~ /// 8	// 2	/ 1		33	0	0	0	0
64.5	(3) / 1	(10) ~~////~~ 5	(7) ~~////~~ // 7	~~////~~ ~~////~~ 10	(−3) /// 3				26	−1	−26	26	17
62.5	(6) / 1	// 2	(14) ~~////~~ // 7	/ 1	(−4) // 2				13	−2	−26	52	24
60.5	(9) / 1	(6) / 1	0	/ 1					3	−3	−9	27	15
f_x	3	10	28	37	22	9	5	6	Σf_y Σf_x =120		2	206	146
x'	−3	−2	−1	0	1	2	3	4			$\Sigma f_y y'$	$\Sigma f_y (y')^2$	$\Sigma\Sigma f_{xy} x'y'$
f_x x'	−9	−20	−28	0	22	18	15	24	22 Σf_x x'				
f_x $(x')^2$	27	40	28	0	22	36	45	96	294 $\Sigma f_x (x')^2$				
Σf_{xy} x'y'	18	24	15	0	5	22	18	44	146 $\Sigma\Sigma f_{xy}$ x'y'				

x = Meßwert im neuen Gedächtnistest
y = Meßwert im alten Gedächtnistest

d) Berechnung
Ein Blick auf die Korrelationstabelle zeigt bereits, daß eine positive Korrelation zu erwarten ist (Verlauf des „Punkteschwarmes" von links unten nach rechts oben).
Der Korrelationskoeffizient r wird dann nach folgender Formel berechnet:

$$r_{xy} = \frac{N\sum\sum f_{xy}x'_i y'_i - \sum f_x x'_i \sum f_y y'_i}{\sqrt{[N\sum f_x x'^2_i - (\sum f_x x'_i)^2]\,[N\sum f_y y'^2_i - (\sum f_y y'_i)^2]}}$$

Es bedeuten: x' und y' die neuen Maßzahlen (aufgrund des angenommenen Mittelwertes)

f_x = die Häufigkeit der x'_i
f_y = die Häufigkeit der y'_i
f_{xy} = die Häufigkeit, mit der der i-te x'-Wert und der i-te y'-Wert gepaart auftreten.

Die für die Formel notwendigen Größen können der Tabelle in Abschnitt c entnommen werden. Setzen wir ein, dann ergibt sich:

$$r_{xy} = \frac{120 \cdot 146 - 22 \cdot 2}{\sqrt{[120 \cdot 294 - 22^2]\,[120 \cdot 206 - 2^2]}} = .60$$

Zwischen den Leistungen im alten und neuen Gedächtnistest besteht also ein positiver Zusammenhang mittlerer Größe.

e) Signifikanzprüfung (s. S. 126)

(4) *Anwendungsregeln (Handrechnung)*

a) Man erstellt eine Korrelationstabelle. In der Kopfzeile werden die Meßwerte des einen Merkmals (x) eingetragen, von links nach rechts steigend. In die linke Randspalte trägt man die Meßwerte des zweiten Merkmals (y), von unten nach oben steigend, ein. Wenn nötig, werden die Meßwerte zu Klassen zusammengefaßt.
In den Zellen strichelt man nun aus, wie oft bestimmte Meßwertkombinationen auftreten.

b) Für jede Zeile wird f_y, für jede Spalte f_x gebildet.
Die Summe der Spalte f_y ($\sum f_y$) muß gleich sein der Summe der Zeile f_x ($\sum f_x$). Beide müssen gleich N sein ($\sum f_x = \sum f_y = N$).

c) Eine Spalte mit einem möglichst hohen f_x-Wert bezeichnen wir mit $x' = 0$, die davon rechts stehenden Spalten mit $x' = 1$; $x' = 2$ usw. (es sind dies Werte über dem angenommenen Mittelwert –134.5–). Die Meßwerte unter dem angenommenen Mittelwert (links von $x' = 0$) erhalten ein negatives Vorzeichen $x' = -1$; $x' = -2$; usw.

d) Entsprechend verfahren wir in den Zeilen. Die Zeile mit dem angenommenen Mittelwert (66.5) wird mit $y' = 0$ benannt. Den benachbarten Zeilen werden die Werte $y' = 1$; $y' = 2$ usw., $y' = -1$; $y' = -2$ usw. zugeordnet.

e) $f_x x'$ erhält man durch die Multiplikation von f_x mit x' für jede Spalte (entsprechend für y' in den Zeilen). Durch Summierung ergibt sich $\sum f_x x'$ bzw. $\sum f_y y'$.

f) Die Multiplikation von $f_x x'$ mit x' ergibt $f_x (x')^2$. Die Summe ist $\sum f_x (x')^2$. Entsprechendes gilt für y'.

g) $f_{xy} x'y'$ erhalten wir, indem für jede besetzte Zelle x' mit y' multipliziert wird und das Produkt mit der in der Zelle stehenden Häufigkeit multipliziert wird (Kreuzprodukte).
Das Ergebnis trägt man (mit Vorzeichen) in Klammern in die rechte obere Ecke der betreffenden Zelle ein.
Beispiel: Die Zelle für $x' = -1$ und $y' = 1$ hat die Häufigkeit
$f_{xy} = 4$. $f_{xy} x'y'$ ist also $4(-1)(1) = -4$.

h) Für jede Spalte und jede Zeile summieren wir die Klammerwerte und erhalten für jede Spalte bzw. Zeile $\sum f_{xy} x'y'$.

i) Durch Summierung der Spalte bzw. Zeile mit der Bezeichnung $\sum f_{xy} x'y'$ ergibt sich $\sum \sum f_{xy} x'y'$. Spaltensumme und Zeilensumme müssen gleich sein (Kontrolle!).

j) Die errechneten Zwischenwerte werden in die Formel

$$r_{xy} = \frac{N \sum \sum f_{xy} x'y' - \sum f_x x' \sum f_y y'}{\sqrt{[N \sum f_x x'^2 - (\sum f_x x')^2]\,[N \sum f_y y'^2 - (\sum f_y y')^2]}}$$

eingesetzt.

(5) *Voraussetzungen und Einschränkungen*

a) Die Beziehung zwischen den beiden Merkmalen x und y muß linear sein (Überprüfung anhand der Korrelationstabellen; Linearitätsprüfung s. Kap. 4.2).
b) Die Verteilung soll symmetrisch sein (möglichst Normalverteilung). Guilford (1965) hält die Normalverteilung der Meßwerte allerdings nicht für unbedingt erforderlich. Die Prüfung auf Normalverteilung kann z.B. erfolgen mit dem Chi^2-Anpassungstest (s. Kap. 5.4.1.3).
c) Die Meßwerte müssen Intervallskalenniveau haben.

(6) *Übungsbeispiele*

Nr. 4/1
10 Schüler wurden einem Wortverständnistest (x) und einem Test, der die Fähigkeit zum logischen Denken (y) prüft, unterzogen. Es ergaben sich folgende Meßwerte:

x:	22	8	19	32	13	24	22	35	18	13	53	15	34	15	27
y:	11	5	6	8	2	5	4	1	7	10	23	9	18	2	4

Der Produkt-Moment-Korrelationskoeffizient ist zu berechnen.

Nr. 4/2
Für folgende Meßwertepaare ist eine Korrelationstabelle zu erstellen und der Korrelationskoeffizient zu berechnen.

x	*y*	*x*	*y*	*x*	*y*
24	29	32	40	29	49
22	40	24	37	50	55
44	36	42	58	76	43
72	32	54	54	40	38
25	46	42	44	32	56
30	47	67	48	61	45
38	49	58	48	56	67
54	53	57	33	61	42
37	51	49	47	17	44
61	50	87	52	61	48
56	45	14	48		
42	48	38	46		
30	25	32	33		
42	48	52	40		
28	28	60	49		

4.1.2. Partialkorrelation

4.1.2.1. Partialkorrelation erster Ordnung

(1) *Untersuchungsproblem*

Die Korrelation zweier Variablen wird durch eine dritte Variable beeinflußt. Das Problem besteht darin, den Einfluß dieser dritten Variablen auszuschalten und den „reinen“ Zusammenhang der beiden Variablen zu ermitteln.

(2) *Beispiel* (nach M. A. May, zit. bei Hofstätter & Wendt 1967)

a) Versuchsplan

An einer Stichprobe von Studenten wurden erhoben: der Studienerfolg, die Intelligenz und die Stundenzahl des Heimstudiums. Es soll der Grad des Zusammenhangs zwischen Intelligenz und Studienerfolg ermittelt werden, wobei der Einfluß der häuslichen Lernarbeit ausgeschaltet werden soll.

Merkmal 1: Studienerfolg
Merkmal 2: Intelligenz
Merkmal 3: Stundenzahl des Heimstudiums

b) Ergebnisse

Je zwei der Merkmale wurden miteinander korreliert. Es ergaben sich folgende Korrelationskoeffizienten:

Studienerfolg/Intelligenz	$r_{12} = .60$
Studienerfolg/Heimstudium	$r_{13} = .32$
Intelligenz/Heimstudium	$r_{23} = -.35$

c) Berechnung

Es ist die Frage gestellt nach dem Grad des Zusammenhanges von Intelligenz und Studienerfolg, wenn man den Einfluß einer dritten Variablen, nämlich die Dauer des Heimstudiums, ausschaltet. Will man einen solchen „reinen" Zusammenhang zwischen Intelligenz und Studienerfolg ermitteln, dann kann man so vorgehen, daß man in seine Stichprobe nur solche Studenten aufnimmt, die zu Hause gleich viel arbeiten.

War ein solches Vorgehen nicht möglich, dann befinden sich in der Stichprobe Studenten, die viel Zeit für ihr Heimstudium aufwenden, und solche, die zu Hause sehr wenig arbeiten. Der Zusammenhang zwischen Intelligenz und Studienerfolg kann nun dadurch verfälscht werden, daß weniger intelligente Studenten etwa ihren Intelligenzmangel durch vermehrte Hausarbeit ausgleichen und einen hohen Studienerfolg aufweisen. Mit Hilfe der Partialkorrelation ist es nun möglich, diesen „störenden" Einfluß der dritten Variablen auszuschalten. Dazu müssen die Korrelationen zwischen je zwei der drei Variablen berechnet werden (s. Abschnitt b). Diese Korrelationskoeffizienten setzt man in folgende Formel ein:

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Es bedeutet $r_{12.3}$: die Korrelation zwischen den Variablen 1 und 2 unter Konstanthaltung der Variablen 3.

Setzen wir unsere Werte ein, dann ergibt sich:

$$r_{12.3} = \frac{.60 - (.32)(-.35)}{\sqrt{(1 - .32^2)(1 - (-.35)^2)}} = .80$$

Es zeigt sich also, daß sich der Korrelationskoeffizient zwischen Intelligenz und Studienerfolg erhöht, wenn man den Einfluß der häuslichen Lernarbeit eliminiert.

In gleicher Weise kann man den Einfluß je einer der beiden anderen Variablen ausschalten.

Betrachtet man den Unterschied zwischen r_{12} (= .60) und $r_{12.3}$ (= .80), dann kann man feststellen, welchen Einfluß die Heterogenität der dritten Variablen (der unterschiedliche häusliche Arbeitsaufwand) auf die Korrelation der Variablen 1 und 2 hat.

d) Signifikanzprüfung

Die Prüfung des Partial-Korrelationskoeffizienten gegen die Nullhypothese erfolgt:

1. Über die z'-Transformation (s. z. B. Fröhlich u. Becker 1971, Mc Nemar 1962),
2. bei kleinen Stichproben auch über die Beziehung:

$$t = \frac{r_{12.3}}{\sqrt{\frac{1 - r_{12.3}^2}{N - 3}}}; \quad df = N - 3$$

Tab. III wird in gewohnter Weise herangezogen.

(3) *Anwendungsregeln*

a) Es werden zunächst die Korrelationen zwischen je zwei der drei Variablen berechnet:
   r_{12} – Korrelation der Var. 1 mit der Var. 2
   r_{13} – Korrelation der Var. 1 mit der Var. 3
   r_{23} – Korrelation der Var. 2 mit der Var. 3
b) Die errechneten Korrelationskoeffizienten werden in die entsprechende Formel eingesetzt.
   Es gilt bei Konstanthaltung der *Variablen 3* die Formel:

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

   Bei Konstanthaltung der *Variablen 2* gilt:

$$r_{13.2} = \frac{r_{13} - (r_{12})(r_{23})}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

   Bei Konstanthaltung der *Variablen 1* gilt:

$$r_{23.1} = \frac{r_{23} - (r_{12})(r_{13})}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

c) Die Signifikanzprüfung erfolgt über die z'-Transformation oder über die t-Verteilung.

(4) *Voraussetzungen und Einschränkungen*

Es gelten die Voraussetzungen für den Produkt-Moment-Korrelationskoeffizienten.

4.1.2.2. Partialkorrelation zweiter Ordnung

Soll der Einfluß zweier Variablen auf die Korrelation zweier anderer Variablen ausgeschaltet werden, dann geschieht dies analog dem oben beschriebenen Vorgehen nach der Formel:

$$r_{12.34} = \frac{r_{12.3} - (r_{14.3})(r_{24.3})}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}}$$

Wie aus der Formel zu ersehen ist, müssen zunächst drei Partial-Korrelationskoeffizienten erster Ordnung ($r_{12.3}$; $r_{14.3}$; $r_{24.3}$) berechnet werden. $r_{12.34}$ liefert dann den Zusammenhang zwischen den Merkmalen 1 und 2 unter Ausschaltung des Einflusses der Merkmale 3 und 4.

(5) *Übungsbeispiel*

Nr. 4 3
Bei einer Stichprobe von 12- bis 19jährigen Jungen ergaben sich folgende Korrelationen:

Größe (1)	und Gewicht	(2)	$r_{12} = .78$
Größe	und Alter	(3)	$r_{13} = .52$
Gewicht	und Alter		$r_{23} = .54$

Es soll die Korrelation zwischen Größe und Gewicht unter Ausschaltung des Alterseinflusses bestimmt werden.

4.1.3. Die Zweizeilenkorrelation

4.1.3.1. Punktbiseriale Korrelation

(1) *Untersuchungsproblem*

Welcher Zusammenhang besteht zwischen zwei Merkmalen, von denen ein Merkmal kontinuierlich, das andere ein Alternativmerkmal ist?
Es existieren also ein normal oder zumindest symmetrisch verteiltes Merkmal 1 und ein Merkmal 2, das nur zwei Klassen aufweist (echtes Alternativmerkmal).

(2) *Beispiel (fiktiv)*

a) Versuchsplan
50 Schüler hatten in einer festgesetzten Zeit möglichst viele Rechenaufgaben zu lösen. Ferner mußten diese Schüler eine Denksportaufgabe lösen. Es wird gefragt nach dem Zusammenhang von Rechengeschwindigkeit und Problemlöseverhalten.
Stichprobe: 50 Schüler der 3. Klasse eines Gymnasiums
Merkmal 1 (kontinuierlich): Zahl der gelösten Aufgaben (x)
Merkmal 2 (alternativ): Denksportaufgabe richtig (y_1),
falsch (y_2).

b) Hypothesen
H_0: Zwischen der Rechengeschwindigkeit und der Fähigkeit, Probleme zu lösen, besteht kein Zusammenhang.
H_1: Es besteht ein bedeutsamer Zusammenhang zwischen Rechengeschwindigkeit und der Fähigkeit, Probleme zu lösen.
Signifikanzniveau: .05

c) Ergebnisse

Zahl der gelösten Rechenaufgaben x	Denksportaufgabe richtig y_1	falsch y_2	f_{tot}
1	1	2	3
2	2	2	4
3	3	3	6
4	6	4	10
5	9	2	11
6	5	1	6
7	4	1	5
8	3	0	3
9	1	0	1
10	1	0	1
	35	15	50

$M_p = 5.26$
$M_q = 3.60$
$s_t \ = 2.045$
$p \ = 35/50 = .70$
$q \ = 15/50 = .30$

Die Eintragungen in der Tabelle bedeuten:
Beispiel Reihe 1: Von 3 Schülern, die eine Rechenaufgabe gelöst haben, hat einer die Denksportaufgabe gelöst, 2 haben sie nicht gelöst.

d) Berechnung

Der punktbiseriale Korrelationskoeffizient (r_{p-bis}) berechnet sich nach folgender Formel:

$$r_{p-bis} = \frac{M_p - M_q}{s_t} \sqrt{pq}$$

Es bedeuten:

M_p = der Mittelwert der Meßwerte x_i aller Versuchspersonen, die in die eine Klasse des Alternativmerkmals (y_1) fallen

M_q = der Mittelwert der Meßwerte x_i aller Versuchspersonen, die in die andere Klasse des Alternativmerkmals (y_2) fallen

s_t = die Standardabweichung aller N Meßwerte x_i

p = Anzahl der Versuchspersonen, die in die eine Klasse des Alternativmerkmals (y_1) fallen, dividiert durch die Gesamtzahl N der Vpn

q = Anzahl der Versuchspersonen, die in die andere Klasse des Alternativmerkmals (y_2) fallen, dividiert durch die Gesamtzahl N der Vpn

$p + q = 1$

N = Größe der Stichprobe

Setzen wir unsere Werte in die Formel ein, so erhalten wir:

$$r_{p-bis} = \frac{5.26 - 3.60}{2.045} \sqrt{.70 \cdot 30} = .373$$

e) Signifikanzprüfung

Die Prüfung von r_{p-bis} gegen die Nullhypothese erfolgt über die Beziehung:

$$t = \frac{M_p - M_q}{s_t}; \ df = N - 2$$

Es ergibt sich ein t-Wert von

$$t = .812$$

Dieser Wert ist kleiner als der in Tab. III unter Berücksichtigung des Signifikanzniveaus und der Freiheitsgrade abgelesene.
Die Nullhypothese ist daher beizubehalten. Zwischen Rechengeschwindigkeit und der Fähigkeit zum Lösen von Problemen besteht kein signifikanter Zusammenhang.

(3) *Anwendungsregeln*

a) Es wird eine Tabelle erstellt, in die die Maßzahlen des kontinuierlichen Merkmals (x) eingetragen werden. In den Spalten des Alternativmerkmals (y_1 und y_2) werden die in jede Klasse fallenden Versuchspersonen unter Berücksichtigung ihres beim Merkmal x erzielten Meßwertes eingetragen.
b) Es werden die Mittelwerte M_p und M_q in bezug auf x berechnet (M_p für die Häufigkeit in Spalte y_1, M_q für die Häufigkeit in Spalte y_2).
c) Die Standardabweichung s_t aller N Meßwerte wird ermittelt.
d) p und q werden errechnet, da $p + q = 1$, ist $p = 1 - q$ und $q = 1 - p$;

$$p = \frac{\text{Häufigkeit in Spalte } y_1}{\text{Gesamt-N}}; \quad q = \frac{\text{Häufigkeit in Spalte } y_2}{\text{Gesamt-N}}$$

e) Der Korrelationskoeffizient wird berechnet nach der Formel:

$$r_{p-bis} = \frac{M_p - M_q}{s_t} \sqrt{pq}$$

($\sqrt{pq}$ kann in Tab. VIII abgelesen werden.)
f) Die Signifikanzprüfung erfolgt über die t-Verteilung.

(4) *Voraussetzungen und Einschränkungen*

a) Das Merkmal x soll normal verteilt sein.
b) Das Merkmal y soll ein echtes Alternativmerkmal sein.

4.1.3.2. Biseriale Korrelation

(1) *Untersuchungsproblem*

Welcher Zusammenhang besteht zwischen zwei Merkmalen, die beide Intervallskalenniveau haben, von denen jedoch ein Merkmal in Alternativen aufgeteilt wurde?
Im Unterschied zur punktbiserialen Korrelation handelt es sich hier beim zweiten Merkmal nicht um ein echtes Alternativmerkmal, sondern um ein an sich kontinuierliches Merkmal, das in zwei Klassen aufgeteilt wurde.

(2) *Beispiel (fiktiv)*

a) Versuchsplan
50 Schüler werden zwei Tests unterzogen. Der erste Test ist ein Intelligenztest, der zweite ein Kreativitätstest. Von beiden Merkmalen wird angenommen, daß sie sich normal verteilen. Das Merkmal Kreativität wird in zwei Klassen aufgeteilt: über dem Median: „hoch kreativ“, unter dem Median: „niedrig kreativ“.

Stichprobe: 50 Schüler der 4. Klasse einer Hauptschule
Merkmal 1 (kontinuierlich): Intelligenz (x)
Merkmal 2 (kontinuierlich, in zwei Klassen aufgeteilt): Kreativität (y)

b) Hypothesen

H_0: Zwischen Intelligenzgrad und Kreativität besteht kein Zusammenhang.
H_1: Zwischen Intelligenzgrad und Kreativität besteht ein signifikanter Zusammenhang.

Signifikanzniveau: .01

c) Ergebnisse

Intelligenz-quotient x	hoch kreativ y_1	niedrig kreativ y_2	f_{tot}
100	0	1	1
106	1	1	2
111	2	4	6
114	3	6	9
115	5	3	8
118	6	2	8
119	7	1	8
120	2	1	3
122	2	1	3
125	2	0	2
	30	20	50

$M_p = 117.33$
$M_q = 113.80$
$s_t = 4.63$
$p = .60$
$q = .40$
$\frac{pq}{y} = .6212$

d) Berechnung

Die Berechnung des biserialen Korrelationskoeffizienten erfolgt nach folgender Formel:

$$r_{bis} = \frac{M_p - M_q}{s_t} \cdot \frac{pq}{y}$$

Die Symbole bedeuten das gleiche wie beim r_{p-bis}. y ist die Höhe der Ordinate in einer Standardnormalverteilung (mit der Fläche $p + q = 1$) an der Stelle der Aufteilung in p und q. Der Quotient pq/y läßt sich für verschiedene p- oder q-Werte der Tab. VII entnehmen.

Setzen wir also unsere Werte ein, dann erhalten wir:

$$r_{bis} = \frac{117.33 - 113.80}{4.63} \cdot .6212 = .474$$

e) Signifikanzprüfung

Die Signifikanzprüfung erfolgt analog dem Vorgehen beim Produkt-Moment-Korrelationskoeffizienten (über die t-Verteilung oder mit Hilfe der Tab. VI).
Aus der Tab. VI entnehmen wir, daß der berechnete Koeffizient $r_{bis} = .474$ nicht signifikant ist. Die Nullhypothese ist also beizubehalten.

(3) *Anwendungsregeln*

a) Es ist in der gleichen Weise vorzugehen wie bei der punktbiserialen Korrelation.
b) Der Wert für den Quotienten pq/y kann in Tab. VII abgelesen werden.

(4) *Voraussetzungen und Einschränkungen*

a) Beide Variablen müssen Intervallskalenniveau haben, wobei eine Variable dichotomisiert (in zwei Klassen aufgeteilt) wurde.
b) Es muß eine symmetrische Verteilung vorliegen.
c) Die Beziehung zwischen den beiden Variablen soll linear sein.

(5) *Übungsbeispiel*

Nr. 4/4
In einer Schulklasse werden die Durchschnittszensuren (x) aller Schüler ermittelt. Weiterhin wird festgestellt, in welchem Grade sich die Eltern jedes Schülers an den Elternversammlungen des laufenden Schuljahres beteiligen (y).
Es soll der Zusammenhang zwischen der Beteiligung der Eltern an den Elternabenden und den Schulzensuren ihrer Kinder untersucht werden.

Durchschnitts-zensur x	Beteiligung am Elternabend anwesend y_1	nicht anwesend y_2	f_{tot}
1.00 - 1.49	1	0	1
1.50 - 1.99	4	0	4
2.00 - 2.49	1	0	1
2.50 - 2.99	5	6	11
3.00 - 3.49	4	5	9
3.50 - 3.99	1	1	2
	16	12	28

4.1.4. Die Rangreihen-Korrelation (rho) nach Spearman

(1) *Untersuchungsproblem*

Welcher Zusammenhang besteht zwischen zwei Merkmalen, deren Meßwerte als Rangdaten vorliegen oder denen Rangplätze zugeordnet werden können?

(2) *Beispiel*

a) Versuchsplan
11 Schüler absolvieren je eine Probearbeit für die Fächer Deutsch und Französisch. Es soll geprüft werden, welcher Zusammenhang zwischen den Leistungen in den beiden Fächern besteht.
Stichprobe: 11 zufällig ausgewählte Schüler einer 5. Realschulklasse
Merkmal 1: Leistungen im Fach Deutsch (Punktwerte)
Merkmal 2: Leistungen im Fach Französisch (Punktwerte)

b) Hypothesen
H_0: Zwischen den Leistungen im Fach Deutsch und dem Fach Französisch besteht kein Zusammenhang.
H_1: Es besteht eine signifikant von Null verschiedene Korrelation zwischen den Leistungen in den beiden Fächern.
Signifikanzniveau: .05

c) Ergebnisse

Schüler	Punktwerte Deutsch	Franz.	R_1 (Deutsch)	R_2 (Franz.)	d	d^2
1	12	20	3	7	−4	16
2	24	18	7	5	2	4
3	16	15	5	2	3	9
4	30	24	10.5	9	1.5	2.25
5	15	19	4	6	−2	4
6	10	14	1	1	0	0
7	29	17	9	4	5	25
8	30	22	10.5	8	2.5	6.25
9	25	28	8	11	−3	9
10	11	16	2	3	−1	1
11	19	26	6	10	−4	16
			$\sum R_1 = 66.0$	$\sum R_2 = 66.0$	$\sum d = 0$	$\sum d^2 = 92.50$

d) Berechnung
Den Meßwerten der Versuchspersonen werden Rangplätze zugeordnet, und zwar getrennt für beide Meßwertverteilungen. Der niedrigste Meßwert erhält den Rang 1, der höchste den Rang N (oder umgekehrt). Existieren Werte von gleicher Größe, so wird jedem Meßwert ein mittlerer Rang zugeordnet (s. ausführliches Beispiel in Kap. 5.4.3). Die Ränge für die Meßwerte „Deutsch“ werden in die Spalte „R_1“ eingetragen, die für Französisch in die Spalte „R_2“.
Die richtige Zuordnung der Ränge läßt sich wie folgt überprüfen: Addiert man die Ränge in den Spalten R_1 und R_2, dann muß sich für jede Spalte $\frac{N(N+1)}{2}$ ergeben.

$$\frac{N(N+1)}{2} = 66; \quad \sum R_1 = 66; \quad \sum R_2 = 66$$

Nun bildet man für jede Versuchsperson die Differenz zwischen den beiden Rangplätzen R_1 und R_2 (Spalte „d“), quadriert diese Differenzen (Spalte „d^2“) und bildet die Summe der Quadrate. Die Summe aller Rangplatzdifferenzen muß Null sein ($\sum d = 0$). Die Berechnung des Rang-Korrelationskoeffizienten Rho (P) oder rho (ϱ) erfolgt nach der Formel:

$$\text{rho}\ (\varrho) = 1 - \frac{6 \sum d^2}{N(N^2-1)}$$

Es bedeuten:

$\sum d^2$ = Summe der quadrierten Differenzen
N = Größe der Stichprobe

Der Rang-Korrelationskoeffizient kann Werte zwischen −1 und +1 annehmen.
Im Beispiel ergibt sich:

$$\text{rho} = 1 - \frac{6 \cdot 92.50}{11(11^2-1)} = 1 - \frac{555}{1320} = 1 - .42 = .58$$

e) Signifikanzprüfung

aa) Für kleine Stichproben

Für kleine Stichproben ($N \leqq 30$) erfolgt die Signifikanzprüfung mit Hilfe der Tab. IX. Ist der errechnete Wert für Rho gleich oder größer als der Wert, der bei dem entsprechenden N unter Berücksichtigung des Signifikanzniveaus in der Tabelle abgelesen wurde, dann ist dieser Rho-Wert als signifikant anzusehen. Die Tabelle gilt für einseitige Tests. Bei einem zweiseitigen Test – die Korrelation unterscheidet sich signifikant von Null – ist das Signifikanzniveau zu verdoppeln. Der in unserem Beispiel errechnete Rho-Wert (rho = .58) ist also signifikant von Null verschieden. Zwischen den Leistungen in den Fächern Deutsch und Französisch besteht ein mittlerer positiver Zusammenhang. Die Nullhypothese wird zurückgewiesen.

bb) Für größere Stichproben ($N > 30$)

Hier erfolgt die Signifikanzprüfung mit Hilfe der Beziehung

$$t = \text{Rho}\sqrt{\frac{N-2}{1-\text{Rho}^2}}; \quad df = N - 2$$

Es bedeuten:

Rho = der Rangkorrelationskoeffizient
N = die Stichprobengröße

(3) *Anwendungsregeln*

a) Den Meßwerten beider Merkmale werden Rangplätze zugeordnet (R_1 und R_2), wenn nicht schon Rangdaten vorliegen.
Meßwerte gleicher Größe erhalten mittlere Ränge.

b) Für jede Versuchsperson werden die Differenzen zwischen den beiden Rangplätzen gebildet (d). Diese Differenzen werden quadriert (d^2), und die Quadrate werden addiert ($\sum d^2$).

c) Der Rangkorrelationskoeffizient wird berechnet:

$$\text{Rho} = \frac{6\sum d^2}{N(N^2-1)}$$

d) Die Signifikanzprüfung erfolgt
für kleine Stichproben über die Tabelle VIII,
für größere Stichproben über die t-Verteilung.

(4) *Voraussetzungen und Einschränkungen*

a) Die Werte müssen Rangdaten sein oder sich in Rangdaten umwandeln lassen.

b) Die Maßzahlen (Rangplätze) sollen normal verteilt sein.

c) Die Abstände zwischen den Rangplätzen sollten für beide Rangreihen gleich sein (Äquidistanz).

d) Die Stichprobe darf nicht kleiner als $N = 5$ sein.

e) Sind diese Voraussetzungen nicht erfüllt, empfiehlt sich die Berechnung des Rangkorrelationskoeffizienten „tau" (τ) von Kendall (s. Siegel 1956).

(5) *Übungsbeispiel*

Nr. 4/5

15 Schüler absolvieren je eine Probearbeit in den Fächern Mathematik und Physik. Sie erhielten folgende Punktwerte.

Es soll festgestellt werden, welcher Zusammenhang zwischen den Leistungen in den beiden Fächern besteht.

Mathematik	Physik
47	75
71	79
52	85
48	50
35	49
35	59
41	75
82	91
71	102
56	87
59	70
73	92
60	54
55	75
41	68

4.1.5. Der Konkordanzkoeffizient von Kendall

(1) *Untersuchungsproblem*

Wie groß ist die Übereinstimmung mehrerer Rangreihen, die von mehreren (k) Personen hinsichtlich mehrerer (N) Beurteilungsgegenstände aufgestellt wurden? War der Rho-Wert ein Maß für den Zusammenhang zweier Rangreihen, so wird hier nach einem Maß für die Übereinstimmung mehrerer Rangreihen gesucht. Diese Problematik entsteht immer dann, wenn mehrere Beurteiler mehrere Objekte oder Individuen zu beurteilen haben.

(2) *Beispiel*

a) Versuchsplan
5 Lehrer sollen 11 Schüler hinsichtlich ihrer künstlerischen Fähigkeiten beurteilen. Jeder Lehrer bringt die Schüler in eine Rangreihe bezüglich ihrer künstlerischen Fähigkeiten. Der am meisten befähigte Schüler erhält den Rang 1, der am wenigsten befähigte den Rang 11.
Es soll nun geprüft werden, in welchem Maße die Lehrer in ihren Einstufungen übereinstimmen.

b) Hypothesen
H_0: Zwischen den Rangreihen der Lehrer besteht keine Übereinstimmung.
H_1: Zwischen den Rangreihen der Lehrer besteht eine bedeutsame Übereinstimmung.
Signifikanzniveau: .05

c) Ergebnisse

Schüler (N = 11)	Rangreihen der (k = 5) Lehrer I	II	III	IV	V	Rangsummen R_j	$\left(R_j - \frac{\sum R_j}{N}\right)^2$
1	4	2	2	1	4	13	289
2	6	5	3	4	3	21	81
3	8	7	10	7	6	38	64
4	1	3	1	2	5	12	324
5	5	1	5	6	2	19	121
6	10	4	8	10	9	41	121
7	11	6	11	11	10	49	81
8	2	10	9	8	7	36	36
9	9	8	7	9	11	44	196
10	7	9	6	5	8	35	25
11	3	11	4	3	1	22	64

$\sum R_j = 330$ $\quad \sum\left(R_j - \frac{\sum R_j}{N}\right)^2 =$

$\frac{\sum R_j}{N} = 30$ $\quad 1402 \rightarrow (\text{QUSR})$

d) Berechnung

Es werden zunächst die Rangsummen in jeder Reihe gebildet (R_j), z.B. Reihe 1: $R_1 = 4+2+2+1+4 = 13$. Diese Rangsummen (R_j) werden addiert und durch N(= 11) dividiert. $\frac{\sum R_j}{N}$ ist die durchschnittliche Rangsumme.

Nun wird von jeder Rangsumme (R_j) die durchschnittliche Rangsumme subtrahiert $\left(R_j - \frac{\sum R_j}{N}\right)$ und jedes Ergebnis quadriert.

Die Quadrate werden addiert, die Summe $\left(\sum\left(R_j - \frac{\sum R_j}{N}\right)^2\right)$ wird als Quadratsumme (QUSR) bezeichnet.

Wir haben damit alle Werte gewonnen, die wir zur Berechnung des Konkordanzkoeffizienten „W" benötigen:

$$W = \frac{12\,\text{QUSR}}{k^2(N^3 - N)}$$

Es bedeuten:

$$\text{QUSR} = \text{Quadratsumme} = \sum\left(R_j - \frac{\sum R_j}{N}\right)^2$$

k = Zahl der Beurteiler bzw. der Rangreihen

N = Zahl der beurteilten Objekte oder Individuen

W kann nur Werte zwischen 0 und +1 annehmen. Es ist damit nur ein Maß für die Stärke des Zusammenhanges (der Übereinstimmung), eine Richtung gibt er nicht an.

Setzen wir unsere Werte ein, dann ergibt sich:

$$W = \frac{12 \cdot 1402}{5^2 \cdot (11^3 - 11)} = \frac{16824}{33000} = .51$$

e) Signifikanzprüfung

aa) Für kleine Stichproben ($N < 8$; k von 3 bis 20)

Die Signifikanzprüfung geschieht mit Hilfe der Tab. X. Ist die errechnete Quadratsumme (QUSR) gleich oder größer als der Wert, der in der Tabelle bei dem in Frage kommenden N bzw. k unter Berücksichtigung des Signifikanzniveaus abgelesen wurde, dann kann H_0 zurückgewiesen werden.

bb) Für größere Stichproben ($N \geqq 8$)

Hier kann zur Prüfung von W gegen die Nullhypothese die Chi-Quadrat-Verteilung benutzt werden.

$$Chi^2 = k(N-1)\,W; \quad df = N-1$$

Wir setzen unsere Werte in diese Formel ein und berechnen den Chi^2-Wert. Dieser errechnete Wert wird mit dem in Tab. XI unter Berücksichtigung des Signifikanzniveaus und der Freiheitsgrade ($df = 11 - 1 = 10$) abgelesenen Wert verglichen. Ist der errechnete Wert gleich oder größer als der Tabellenwert, dann weisen wir H_0 zurück.

Der Chi^2-Wert in unserem Beispiel beträgt $5 \cdot (11-1) \cdot .51 = 25.50$. Er ist größer als der Tafelwert. Die Nullhypothese ist zurückzuweisen. Zwischen den Lehrern besteht eine signifikante Übereinstimmung bei der Beurteilung der Schüler.

(3) *Anwendungsregeln*

a) Die Ergebnisse werden in eine Tabelle mit k Spalten (Zahl der Beurteiler bzw. der Rangreihen) und N Reihen (Zahl der beurteilten Objekte oder Individuen) eingetragen.

b) Für jede Reihe wird die Rangsumme (R_j) gebildet.

c) Alle Rangsummen werden addiert und durch N dividiert: $\frac{\sum R_j}{N}$. Das Ergebnis ist die durchschnittliche Rangsumme.

d) Von jeder Rangsumme wird die durchschnittliche Rangsumme subtrahiert, das Ergebnis quadriert $\left(R_j - \frac{\sum R_j}{N}\right)^2$ und die Summe der Quadrate gebildet. Man erhält die Quadratsumme $\left(QUSR = \sum \left(R_j - \frac{\sum R_j}{N}\right)^2\right)$.

e) Der Konkordanzkoeffizient W wird berechnet:

$$W = \frac{12\,QUSR}{k^2(N^3 - N)}$$

f) Die Prüfung auf Signifikanz erfolgt bei

kleinen Stichproben ($N < 8$; k von 3 bis 20)

mit Hilfe der Tab. X.

Ist die errechnete Quadratsumme gleich oder größer als der entsprechende Tabellenwert, kann H_0 zurückgewiesen werden;

größeren Stichproben ($N \geqq 8$)

über die Chi-Quadrat-Verteilung nach der Beziehung

$Chi^2 = k(N-1)\,W; \quad df = N - 1.$

Ist der errechnete Chi^2-Wert gleich oder größer als der in Tab. X abgelesene, dann kann H_0 zurückgewiesen werden.

(4) *Voraussetzungen und Einschränkungen*

a) Der Konkordanzkoeffizient W nimmt nur Werte von 0 bis +1 an. Er ist nicht mit r_{xy} vergleichbar.
b) Bei einer größeren Zahl gleich großer Werte *(ties)* kann zur Berechnung von W eine korrigierte Formel verwendet werden (vgl. Siegel 1956).

(5) *Übungsbeispiel*

Nr. 4/6
3 Lehrer stuften 6 Schüler hinsichtlich ihrer Mitarbeit während der Unterrichtsstunden ein. Rang 1 erhielt der Schüler mit der besten Mitarbeit, Rang 6 der mit der schlechtesten. Es ist zu überprüfen, inwieweit die Lehrer in ihrem Urteil übereinstimmen.

Schüler		Lehrer	
	I	II	III
1	1	1	6
2	6	5	3
3	3	6	2
4	2	4	5
5	5	2	4
6	4	3	1

4.1.6. Der biseriale Rangkorrelationskoeffizient (Whytfield)

(1) *Untersuchungsproblem*

Welcher Zusammenhang besteht zwischen zwei Merkmalen, von denen das eine Rangskalenniveau hat, während das andere nur zwei Klassen aufweist (Alternativmerkmal)?

(2) *Beispiel* (nach Lienert 1961)

a) Versuchsplan
15 Lehrlinge wurden vor ihrem Eintritt in einen Betrieb einem Eignungstest unterzogen. Den Lehrlingen wurden aufgrund ihrer Testleistung Ränge zugeordnet, wobei der Lehrling mit dem niedrigsten Punktwert den Rang 1 erhielt, der mit dem höchsten den Wert 15. Nach einem halben Jahr wurden von den Ausbildern $N_1 = 10$ Lehrlinge als geeignet, $N_2 = 5$ Lehrlinge als für eine weitere Ausbildung ungeeignet befunden.
Es soll untersucht werden, welcher Zusammenhang besteht zwischen den Testleistungen und dem Urteil der Ausbilder.
Stichprobe: 15 Lehrlinge
Merkmal 1 (ordinal): Leistung im Eignungstest
Merkmal 2 (nominal): Eignungsurteil der Ausbilder
geeignet (+), nicht geeignet (−)
b) Hypothesen
H_0: Zwischen der Leistung im Eignungstest und dem Urteil der Ausbilder besteht kein Zusammenhang.

H_1: Es besteht ein bedeutsamer Zusammenhang zwischen Testwert und Ausbilderurteil.

Signifikanzniveau: .05

c) Ergebnisse

1	2	3	4	5	6	7	8	9	10	11
Lehrlinge	Ränge nach d. Testwerten	Eignungsurteil	Ränge · Minuszeichen	Mittlere Ränge f. Minuszeichen	Ränge für Pluszeichen	Mittlere Ränge f. Pluszeichen	Mittlere Ränge f. Plus- u. Minusz.	Überschreitung	Unterschreitung	Differenzen
1	1	–	1	3			3	10	0	10
2	2	–	2	3			3	10	0	10
3	3	–	3	3			3	10	0	10
4	4	+			6	10.5	10.5	0	2	–2
5	5	+			7	10.5	10.5	0	2	–2
6	6	–	4	3			3	8	0	8
7	7	+			8	10.5	10.5	0	1	–1
8	8	+			9	10.5	10.5	0	1	–1
9	9	–	5	3			3	6	0	6
10	10	+			10	10.5	10.5	0	0	0
11	11	+			11	10.5	10.5	0	0	0
12	12	+			12	10.5	10.5	0	0	0
13	13	+			13	10.5	10.5	0	0	0
14	14	+			14	10.5	10.5	0	0	0
15	15	+			15	10.5	10.5	0	0	0
			$\sum R_2 = 15$	$\frac{\sum R_2}{N_2} = \frac{15}{5} = 3$	$\sum R_1 = 105$	$\frac{\sum R_1}{N_1} = \frac{105}{10} = 10.5$				Sb= 38

d) Berechnung

Die Alternativsymbole (+ und –) des Merkmals 2 sind in zwei Gruppen mit gleichen Rängen zu verwandeln.

Dazu ordnet man zunächst den Minuszeichen Rangplätze zu (z. B. hier die Ränge 1 bis 5; s. Spalte 4). Danach erhalten die Pluszeichen Rangplätze, indem man in der Rangfolge weiterzählt (hier die Ränge 6 bis 15; s. Spalte 6). Die mittleren Ränge ermittelt man, indem man für jede Gruppe das arithmetische Mittel aus den Rängen bildet (z. B. für die Pluszeichen $\frac{\sum R_1}{N_1} = \frac{105}{10} = 10.5$).

Die mittleren Ränge beider Gruppen werden nun untereinandergeschrieben (Sp. 8).

Jetzt zählt man in Spalte 8 *von oben nach unten* aus,

aa) wie oft eine Rangzahl von einer größeren überschritten wird, und trägt die Zahl der Überschreitungen neben diese Rangzahl ein (Sp. 9).
Beispiel: Rang 3 in der ersten Reihe der Spalte 8 wird von 10 größeren Rängen überschritten, nämlich zehnmal von Rang 10.5.

bb) wie oft eine Rangzahl von einer kleineren unterschritten wird, und trägt die Zahl der Unterschreitungen in Spalte 10 ein.
Beispiel: Rang 10.5 in der 4. Reihe der Spalte 8 wird von 2 kleineren Rängen unterschritten, nämlich zweimal von Rang 3.

Nach der Auszählung werden die Differenzen zwischen Überschreitungen und Unterschreitungen gebildet (Sp. 9 minus Sp. 10). Diese Differenzen sind in Spalte 11 einzutragen und zu addieren.
Man erhält die Prüfgröße Sb (= 38).
Außer Sb wird zur Berechnung des biserialen Rangkorrelationskoeffizienten folgender Wert benötigt:

$${}_{corr}Sb - max = \sqrt{\tfrac{1}{4}\{N(N-1)[N(N-1) - N_1(N_1-1) - N_2(N_2-1)]\}}$$

Der biseriale Rangkorrelationskoeffizient tau_b (τ_b) ist zu berechnen aus:

$$tau_b = \frac{Sb}{{}_{corr}Sb - max}$$

Setzen wir unsere Werte ein, so erhalten wir:

$$tau_b = \frac{38}{\sqrt{\tfrac{1}{4}\{15(15-1)[15(15-1) - 10(10-1) - 5(5-1)]\}}} = .52$$

e) Signifikanzprüfung
Zur Signifikanzprüfung verwendet man die Prüfgröße:

$$z = \frac{Sb}{\sigma_{Sb}}$$

Dabei ist

$$\sigma_{Sb} = \sqrt{\tfrac{1}{3}N_1\ N_2(N+1)}$$

Diese Beziehung für σ_{Sb} gilt nur dann, wenn in der ursprünglichen Rangreihe der Versuchspersonen auf dem ordinal gemessenen Merkmal (hier die Rangreihe nach Testwerten – Sp. 2 –) keine gleichen Ränge auftreten. Eine Korrekturformel für gleiche Ränge findet sich bei Lienert (1961).
Setzen wir nun unsere Werte ein, so ergibt sich:

$$z = \frac{38}{\sqrt{\tfrac{1}{3}10.5(15+1)}} = 2.32$$

Der Tab. II können wir entnehmen, daß dieser Wert eine Restwahrscheinlichkeit von $p < .05$ hat. Die Nullhypothese kann verworfen werden. Zwischen Leistung im Eignungstest und Ausbilderurteil besteht ein signifikanter Zusammenhang.

(3) *Anwendungsregeln*

a) Für jede Versuchsperson werden der Rang auf dem ordinal gemessenen Merkmal und die Klasse des Alternativmerkmals (+ ; –) nebeneinander in eine Tabelle eingetragen.
b) Die Alternativsymboie (+ und –) werden in zwei Gruppen gleicher Ränge umgewandelt.
c) Es wird ausgezählt, wie oft jeder der Ränge aus beiden Gruppen über- bzw. unterschritten wird.
d) Die Differenzen zwischen Über- und Unterschreitungen werden gebildet und addiert. Wir erhalten Sb.
e) Der Wert ${}_{corr}Sb - max$ wird nach der angegebenen Formel ermittelt.

f) Der biseriale Rangkorrelationskoeffizient tau_b wird berechnet:

$$tau_b = \frac{Sb}{{}_{corr}Sb - max}$$

g) Zur Signifikanzprüfung dient die Beziehung:

$$z = \frac{Sb}{\sigma_{Sb}}$$

h) In Tab. II wird die Überschreitungswahrscheinlichkeit für z abgelesen. Ist sie größer als unser Signifikanzniveau, dann behalten wir H_0 bei, ist sie kleiner, dann weisen wir H_0 zurück.

4.1.7. Punkt-Vierfelder-Korrelation (Pearson)

(1) *Untersuchungsproblem*

Welcher Zusammenhang besteht zwischen zwei Alternativmerkmalen, die bei einer Gruppe von Merkmalsträgern gemessen wurden?
Ursprünglich war diese Methode für die Berechnung des Zusammenhanges von zwei „echten" Alternativmerkmalen konzipiert (männl.–weibl.; tot–lebendig). Sie findet aber auch Anwendung bei kontinuierlichen Merkmalen, die dichotomisiert (in zwei Klassen aufgeteilt) wurden.

(2) *Beispiel*

a) Versuchsplan
Es soll untersucht werden, welcher Zusammenhang besteht zwischen dem Geschlecht der Schüler und ihrer Zugehörigkeit zu Jugendorganisationen.
Stichprobe: 120 Schüler (60 männl., 60 weibl.) im Alter von 15 Jahren, die ein Gymnasium einer Großstadt besuchen.
Merkmal 1: Geschlecht
Merkmal 2: Zugehörigkeit zu einer Jugendorganisation, ja – nein.

b) Hypothesen
H_0: Es besteht kein Zusammenhang zwischen dem Geschlecht der Schüler und der Zugehörigkeit zu einer Jugendorganisation.
H_1: Es besteht ein bedeutsamer Zusammenhang zwischen beiden Merkmalen.
Signifikanzniveau: .05

c) Ergebnisse
Die Ergebnisse werden in eine Vierfeldertafel eingetragen.

		Geschlecht männl.	Geschlecht weibl.	Σ
Zugehörigkeit	ja	A 78	B 34	112
	nein	C 22	D 66	88
	Σ	100	100	200

d) Berechnung

Die Berechnung des Korrelationskoeffizienten Phi (Φ) oder phi (φ) erfolgt nach der Formel:

$$\text{Phi} = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Der Korrelationskoeffizient Phi kann Werte von -1 bis $+1$ annehmen. Diese Werte werden aber nur erreicht, wenn eine Diagonale der Vierfeldertafel den Wert 0 hat. Zur Berechnung des Schwankungsbereichs für Phi bei gegebener Vierfeldertafel siehe Clauß & Ebner (1971). Durch Einsetzen unserer Werte erhalten wir:

$$\text{Phi} = \frac{78 \cdot 66 - 34 \cdot 22}{\sqrt{112 \cdot 88 \cdot 100 \cdot 100}} = \frac{4400}{9927.739} = .444$$

Die Beziehung zwischen Phi und Chi² (2 × 2)

Zwischen dem Korrelationskoeffizienten Phi und Chi^2 besteht folgender Zusammenhang:

$$\text{Phi} = \sqrt{\frac{\text{Chi}^2}{N}} \quad \text{(s. a. Kap. 5.4.2, Abschn. 4)}$$

Wurde der Chi^2-Wert für die Vierfeldertafel bereits berechnet, so läßt sich Phi leicht ermitteln. Durch Umformung erhält man

$$\text{Phi}^2 \cdot N = \text{Chi}^2$$
$$.444^2 \cdot 200 = 39.60$$

Wurde Phi berechnet, so erhält man über diese Beziehung rasch den Chi^2-Wert. Der Chi^2-Wert wird benötigt für die Signifikanzprüfung. Ist nämlich der Chi^2-Wert aus einer Vierfeldertafel signifikant, dann ist auch der entsprechende Phi-Koeffizient signifikant.

e) Signifikanzprüfung

Sie geschieht, wie erwähnt, über die Chi-Quadrat-Verteilung. In Tab. XI lesen wir bei $df = 1$ für den Chi^2-Wert 39.60 eine Restwahrscheinlichkeit von $p < .001$ ab. Damit ist die Annahme eines signifikanten Zusammenhanges zwischen den beiden Merkmalen bestätigt. Die Nullhypothese kann zurückgewiesen werden.

(3) *Anwendungsregeln*

a) Man zählt die Merkmalskombinationen aus (z.B. die Kombination: männl./Zugehörigkeit nein = 22) und trägt sie in eine Vierfeldertafel ein.

		Merkmal 1		
		+	−	
Merkmal 2	+	A	B	
	−	C	D	
				N

(+ ⟶ Merkmal vorhanden, − ⟶ Merkmal nicht vorhanden)

b) Der Korrelationskoeffizient wird berechnet:

$$\text{Phi} = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \quad \text{oder} \quad \text{Phi} = \sqrt{\frac{\text{Chi}^2}{N}}$$

c) Man ermittelt den zugehörigen Chi^2-Wert (wenn nicht schon berechnet):

$$\text{Chi}^2 = N \cdot \text{Phi}^2$$

d) In Tab. XI liest man bei $df = 1$ die Restwahrscheinlichkeit für den errechneten Chi^2-Wert ab. Ist der abgelesene Wahrscheinlichkeitswert kleiner als unser Signifikanzniveau, dann weisen wir H_0 zurück; ist er größer, dann behalten wir H_0 bei.

(4) *Voraussetzungen und Einschränkungen*

a) In keinem der vier Felder darf eine Häufigkeit Null stehen.
b) Die Erwartungswerte (s. a. Kap. 5.4.1.2) dürfen nicht kleiner als 5 sein.
c) Bei kleineren Stichproben ist eine korrigierte Formel zu verwenden:

$$\text{Phi} = \frac{|AD - BC| - \frac{N}{2}}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

d) Bei Stichproben $N < 25$ ist das Verfahren nicht anwendbar.

(5) *Übungsbeispiel*

Nr. 4/7
Im Rahmen der Konstruktion eines Schultests soll festgestellt werden, ob zwischen der Lösung der ersten und zweiten Aufgabe ein Zusammenhang besteht. Die Antworten von 124 Schülern wurden ausgewertet.

		Aufgabe 1: gelöst +	Aufgabe 1: nicht gelöst −
Aufgabe 2	+	20	35
	−	19	50

4.1.8. Der Kontingenzkoeffizient C nach Pearson

(1) *Untersuchungsproblem*

Wie groß ist der Zusammenhang zwischen zwei mehrklassigen Merkmalen, die bei einer Gruppe von Merkmalsträgern gemessen wurden?
Lieferte Phi ein Maß für den Grad des Zusammenhanges von zwei zweiklassigen (Alternativ-)Merkmalen, so wird nun ein Maß für den Zusammenhang solcher Merkmale gesucht, die mehrere Klassen aufweisen. Dabei kann jedes Merkmal beliebig viele Klassen haben.

(2) *Beispiel*

a) Versuchsplan

Es ist gefragt nach der Stärke des Zusammenhanges zwischen Schulbildung und Einstellung gegenüber politischen Minderheiten. Die Befragung wurde bei 200 männlichen Versuchspersonen durchgeführt.

Stichprobe: 200 Vpn (männlich) im Alter von 25–30 Jahren in einem Großbetrieb.

Merkmal 1 ($k = 4$ Klassen): Schulbildung: Hauptschule, Realschule, Gymnasium, Hochschule

Merkmal 2 ($r = 3$ Klassen): Einstellung: positiv, neutral, negativ.

b) Hypothesen

H_0: Zwischen Schulbildung und Einstellung gegenüber politischen Minderheiten besteht kein von Null verschiedener Zusammenhang.

H_1: Zwischen beiden Merkmalen besteht ein Zusammenhang.

Signifikanzniveau: .05

c) Ergebnisse

Die Merkmalskombinationen werden in eine Mehrfeldertafel eingetragen.

$k = 4$

		Hauptschule	Realschule	Gymnasium	Hochschule	
$r = 3$	positiv	16	19	17	18	70
	neutral	21	24	20	25	90
	negativ	8	12	11	9	40
		45	55	48	52	200

d) Berechnung

Der Grad des Zusammenhanges zwischen den beiden Merkmalen wird ausgedrückt durch den Kontingenzkoeffizienten C:

$$C = \sqrt{\frac{Chi^2}{N + Chi^2}}$$

wobei $Chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(0_{ij} - E_{ij})^2}{E_{ij}}$

Chi^2 wird nach der in Kapitel 5.4.1.2 beschriebenen Weise berechnet. Für unser Beispiel ergibt sich ein Chi^2-Wert von .626.

Die weitere Rechnung ist hier eigentlich überflüssig. Wir setzen aber trotzdem ein und erhalten

$$C = \sqrt{\frac{.626}{200 + .626}} = .064$$

C kann nur Werte zwischen 0 und +1 annehmen, wobei die maximale Größe für C von der Größe der Mehrfeldertafel abhängt. Welchen Wert man für C höchstens erreichen kann, läßt sich nach folgender Formel berechnen (sie gilt aber nur für *quadratische* Mehrfeldertafeln):

$$C_{max} = \sqrt{\frac{k-1}{k}}$$

(k = Anzahl der Zeilen oder Spalten.)

Es ergeben sich folgende Höchstwerte für C bei quadratischen Mehrfeldertafeln (nach Guilford 1965):

Anzahl der Zeilen oder Spalten k =	2	3	4	5	6	7	8	9	10	∞
C_{max} =	.707	.816	.866	.894	.913	.926	.935	.943	.949	1

Für *rechteckige* Tafeln kann man C_{max} schätzen durch Mittelung der Werte für die Spalten bzw. Zeilen.

Beispiel: Bei einer 5 × 7-Tafel ist $C_{max} \approx \frac{.894 + .926}{2} = .910$

Beziehung zwischen C und Chi²

Zwischen dem Kontingenzkoeffizienten C und Chi^2 besteht folgende Beziehung:

$$Chi^2 = \frac{N\,C^2}{1 - C^2}; \quad df = (k-1)(r-1)$$

e) Signifikanzprüfung

Wir müssen nun prüfen, ob der gefundene Zusammenhang signifikant von Null verschieden ist.

Zu dieser Signifikanzprüfung verwenden wir den errechneten Chi^2-Wert mit den Freiheitsgraden $(k-1)(r-1)$. Hierbei bedeutet k die Anzahl der Spalten, r die Anzahl der Reihen. Wir vergleichen also den gefundenen Chi^2-Wert mit dem bei df = 6 und Signifikanzniveau .05 in Tab. XI abgelesenen Wert. Unser errechneter Wert ist kleiner als der dort aufgefundene, die Nullhypothese ist beizubehalten. Zwischen den beiden Merkmalen besteht kein von Null verschiedener Zusammenhang.

(3) *Anwendungsregeln*

a) Die Merkmalskombinationen werden in eine Mehrfeldertafel eingetragen.
b) Chi^2 wird nach den in Kap. 5.4.1.2 genannten Regeln berechnet (s. a. Kap. 5.4.2).
c) Der errechnete Chi^2-Wert wird in die Formel

$$C = \sqrt{\frac{Chi^2}{N + Chi^2}}$$

eingesetzt.
d) Zur Signifikanzprüfung wird der gefundene Chi^2-Wert herangezogen. Der in Tab. XI bei $df = (k-1)(r-1)$ unter Berücksichtigung des Signifikanzniveaus abgelesene Wert wird mit dem errechneten Chi^2-Wert verglichen. Ist der errechnete Wert gleich oder größer, dann weisen wir H_0 zurück, ist er kleiner, behalten wir H_0 bei. C beschreibt nur die Stärke des Zusammenhanges zwischen beiden Merkmalen. Die Richtung des Zusammenhanges ist dem stets positiven C-Wert nicht zu entnehmen. Die Interpretation muß sich daher auf die Werte in der Mehrfeldertafel stützen.

(4) *Voraussetzungen und Einschränkungen*

a) Es gelten die für die Berechnung von Chi^2 gemachten Angaben (s. Kap. 5.4.1.1 u. 5.4.1.2).

b) Zwei Kontingenzkoeffizienten, die aus verschieden großen Mehrfeldertafeln stammen, sind nicht miteinander vergleichbar (C_{max} hängt bekanntlich von der Größe der Tafeln ab).

c) C ist nicht direkt vergleichbar mit anderen Korrelationskoeffizienten, z. B. mit r (Pearson), rho (Spearman).

(5) *Übungsbeispiel*

Nr. 4/8
Gefragt ist nach dem Zusammenhang von Schulerfolg und Erfolg in der Lehrlingsausbildung. Folgende Werte sind gegeben ($N = 500$):

		Schulerfolg		
		niedrig	mittel	hoch
Ausbildungserfolg	hoch	5	45	50
	mittel	50	110	40
	niedrig	45	145	10

4.1.9. Der Cramérsche Koeffizient

Fröhlich (1971) schlägt vor, anstelle des Kontingenzkoeffizienten C den sogenannten Cramérschen Koeffizienten zu berechnen.
Dieser Koeffizient hat den Vorteil, von 0 bis 1 zu variieren und damit eine größere Aussagekraft zu besitzen. Ferner ist der Koeffizient von Cramér im Gegensatz zu dem Kontingenzkoeffizienten C unabhängig von der Größe der Mehrfeldertafeln. Damit sind Vergleiche zwischen Koeffizienten aus verschieden großen Tafeln möglich.

Der Cramér-Koeffizient wird berechnet nach der Formel:

$$C_C = \sqrt{\frac{Chi^2}{N(L-1)}}$$

Es bedeuten:

N = Stichprobengröße
L = die kleinere Zeilen- oder Spaltenzahl der Mehrfeldertafel
(Im Beispiel für C [Kap. 4.1.8] ist also L = 3; bei quadratischen Tafeln ist L gleich der Zeilen- und Spaltenzahl.)

Zwischen dem Cramér-Koeffizienten und dem Punkt-Vierfelder-Koeffizienten Phi besteht folgende Beziehung:

$$Phi = C_C = \sqrt{\frac{Chi^2}{N}}$$

Die beiden Koeffizienten können also unmittelbar miteinander verglichen werden.
Die Berechnung von C_C erfolgt analog den für den Kontingenzkoeffizienten C gemachten Angaben.

Übungsbeispiel

Nr. 4/9
Für die Werte der Aufgabe in Kap. 4.1.8 ist der Cramér-Koeffizient zu berechnen.

4.2. Regression und Vorhersage

Die Methoden zur Untersuchung von Regressionsproblemen haben nahe theoretische Verwandtschaft zu denen der Korrelationsprobleme: Beide wollen einen statistischen Zusammenhang zwischen numerischen Variablen aufdecken, bei beiden gilt das Interesse der *Form* der Beziehung, nicht nur dem Ausmaß, beide beinhalten dieselben Begriffe der deskriptiven Statistik. Die wesentlichen Unterschiede bestehen hinsichtlich der Stichprobenerhebung, der gemachten Annahmen und der möglichen Schlußfolgerungen oder Vorhersagen.

Bei Korrelationsproblemen:

a) Es gibt keine expliziten Einschränkungen hinsichtlich der Werte beider Variablen. Beide Variablen können jeden beliebigen Wert annehmen.
b) Die zugrundeliegende Frage ist die nach dem Ausmaß des Zusammenhangs der beiden Variablen.
c) Zur Bestimmung des Zusammenhangs wird eine bivariate Verteilung angenommen. Wenn diese Annahme gültig ist, sind Aussagen über den Zusammenhang der beiden untersuchten Variablen auch über die ermittelten Daten hinaus möglich, nicht nur über die gerade in *diesem* Experiment aufgetretenen Ausprägungen der Variablen (wenn die übrigen Bedingungen für die Generalisierbarkeit empirischer Untersuchungsergebnisse erfüllt sind).

Bei Regressionsproblemen:

a) Die Ausprägungsgrade einer der beiden Variablen sind keine Zufallsauswahl aus einer Grundgesamtheit (Population) von möglichen Ausprägungsgraden wie bei der Korrelation, sondern sie sind genau festgelegt, sie bilden selbst die Grundgesamtheit.
b) Die zugrundeliegende Frage ist die, ob eine *bestimmte* Variable, nämlich die in ihren Ausprägungsgraden festgelegte, unabhängige Variable die Variation der abhängigen Variablen bedingt, ob von der unabhängigen Variablen Vorhersagen gemacht werden können auf die abhängige (vgl. Varianzanalyse in Kap. 5.3.4).
c) An die experimentelle Verteilung der unabhängigen Variablen wird keine Forderung gestellt. Für die Verteilung der abhängigen Variablen gilt: die Fehlerverteilung soll innerhalb jedes Ausprägungsgrades der unabhängigen Variablen normal und gleich groß sein, und die Werte innerhalb jedes Ausprägungsgrades der unabhängigen Variablen sollen unabhängig voneinander sein.

4.2.1. Einfache (lineare) Regression

(1) *Untersuchungsproblem*

Es liegen die Daten einer abhängigen Variablen vor, die unter verschiedenen Bedingungen einer unabhängigen Variablen erhoben wurden. Es soll eine lineare Funktion (die Regressionsgerade) gefunden werden, der sich die Daten am besten anpassen (Anpassung des Punkteschwarms, vgl. die Veranschaulichungen zu Kap. 4.1), die die beste Vorhersage von Daten der abhängigen Variablen zuläßt, wenn die Werte der unabhängigen bekannt sind.

(2) *Beispiel*

a) Versuchsplan

unabhängige Variable:	Beruhigungsmittel in 6 Stärken
abhängige Variable:	Problemlösefähigkeit (gemessen durch die Anzahl der richtigen Lösungen)
Versuchspersonen:	Gymnasiasten der 12. Klasse

Eine Zufallsstichprobe gleich intelligenter Gymnasiasten der 12. Klasse wird per Zufall auf die 6 quantitativ unterschiedlichen Beruhigungsmittel-Bedingungen verteilt. Nach einer festgelegten Wirkungszeit des Transquilizers werden alle Vpn einem Problemlösungstest unterzogen. Die Frage ist, inwieweit die Variation der Problemlösefähigkeit durch die Variation der Transquilizerstärke bedingt ist bzw. inwieweit man aufgrund dieses Experimentes die Problemlösefähigkeit eines nicht getesteten Gymnasiasten der 12. Klasse vorhersagen kann, wenn man weiß, ein wie starkes Beruhigungsmittel ihm verabfolgt worden ist.

b) Hypothesen

H_0: Es besteht kein linearer Zusammenhang zwischen den beiden untersuchten Variablen; die experimentell abhängige ist linear unabhängig von der experimentell unabhängigen, es können aufgrund der experimentell unabhängigen Variablen keine Vorhersagen für die experimentell abhängige Variable getroffen werden.

H_1: Es besteht ein statistisch nachweisbarer Zusammenhang zwischen den beiden untersuchten Variablen; es können mit einer bestimmten Wahrscheinlichkeit Vorhersagen für die abhängige Variable aufgrund der unabhängigen gemacht werden. Ein Teil der Varianz der abhängigen Variablen wird durch die unabhängige Variable aufgeklärt.

c) Ergebnisse

Die Ergebnisse zu Regressionsproblemen lassen sich ebenso wie bei Korrelationsproblemen am anschaulichsten in einem sog. *Scatter*diagramm und der unten aufgeführten Tabelle anordnen. Auf der x-Achse des *Scatter*diagramms werden die Kategorien oder Bedingungen der unabhängigen, auf der y-Achse die Werte der dazugehörigen abhängigen Variablen abgetragen. Es ergibt sich dabei ein mehr oder weniger konzentrierter Punkteschwarm (vgl. Kap. 4.1).
Die Ergebnisse unseres Beispiels stellen sich dann wie folgt dar:

Rohdatentabelle

	Beruhigungsmittel (X)					
Stärke	1	2	3	4	5	6
Anzahl der	32	30	27	19	23	20
Problemlösungen	29	30	26	21	19	19
pro Vp	30	26	24	20	20	16
(Y)	28	27	24	20	22	18
	26	26	23	18	18	16

$N = 30$

$\sum\sum x_j = 105$ (Summe aller X-Werte)

$\sum\sum x_j^2 = 455$ (Summe aller quadrierten X-Werte)

$\sum\sum y_{ij} = 697$ (Summe aller Y-Werte)

$\sum\sum y_{ij}^2 = 16813$ (Summe aller quadrierten Y-Werte)

$\sum\sum y_{ij}x_j = 2231$ (Summe aller mit dem entsprechenden X-Wert gewichteten Y-Werte)

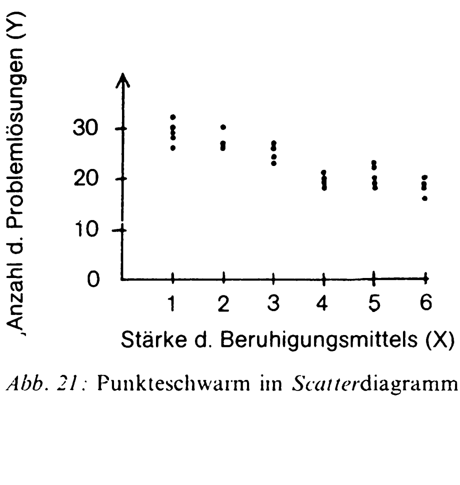


Abb. 21: Punkteschwarm in *Scatter*diagramm (ohne Regressionsgerade)

d) Berechnung
Für diesen Punkteschwarm im *Scatter*diagramm soll nun die bestpassende Gerade gesucht werden, die Funktion, aufgrund deren von der unabhängigen Variablen (X) die Werte der abhängigen Variablen (Y) am genauesten vorhergesagt werden können.
Das nachstehende Diagramm soll das Prinzip veranschaulichen, nach dem bei diesem Problem vorgegangen wird.

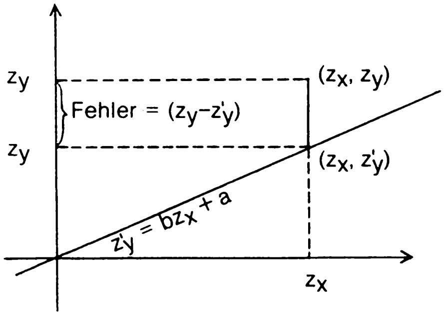


Abb. 22: Punkteschwarm im *Scatter*diagramm (mit Regressionsgerade)

Dabei bedeuten:

z_y = der beobachtete Standardwert von Y
z_x = der beobachtete Standardwert von X
z'_y = der aufgrund der Regressionsgeraden vorhergesagte Standardwert von Y

Die Regressionsgerade bestimmt sich entsprechend der Geradenfunktion $Y = bX + a$ als $z'_y = bz_x + a$. Da wir es im Augenblick noch mit Standardwerten zu tun haben, d.h. die Gerade durch den Ursprung geht, wird $a = 0$ und die Gleichung für die Regressionsgerade lautet einfach:

$$z'_y = bz_x$$

Die Gerade soll so verlaufen, daß der quadrierte Fehler in der Voraussage (die quadrierten Differenzen zwischen z_y und z'_y) ein Minimum beträgt:

$$\frac{\sum (z'_y - z_y)^2}{N} = \text{Minimum}$$

Diese Bedingung kann nur durch die einzige Unbestimmte der Regressionsgleichung, nämlich die Konstante b, erfüllt werden. Aus den bisher aufgestellten Gleichungen läßt sich logisch leicht ein b herleiten, das der Minimum-Bedingung gehorcht:

$$b = \frac{\sum z_x z_y}{N} = \text{Regressionskoeffizient} = \text{Steigung der Regressionsgeraden}$$

Bei Standardwerten entspricht b auch dem Korrelationskoeffizienten r_{xy}, der Korrelation zwischen X und Y, und die Gleichung für die Regressionsgerade lautet nun:

$$z'_y = r_{xy} z_x$$

Wenn wir für $z'_y = \frac{Y' - M_y}{s_y}$ und für $z_x = \frac{X - M_x}{s_x}$ einsetzen, läßt sich die entsprechende Beziehung für die Rohwerte bestimmen, nämlich:

$$Y' = \frac{r_{xy}s_y}{s_x}(X - M_x) + M_y$$

$$b = \frac{r_{xy}s_y}{s_x} = \text{Regressionskoeffizient (Steigung)}$$

Die Berechnung für r_{xy} als der auf die beiden Variablen-Varianzen normierten Kovarianz zwischen den beiden Variablen haben wir bereits in Kap. 4.1 kennengelernt. Hier noch einmal die vereinfachte Rechenformel:

$$r_{xy} = \frac{\left(\sum_i X_i Y_i/N\right) - M_x M_y}{s_x s_y}$$

Bei bekanntem s_x, s_y, M_x und M_y und dem berechneten Korrelationskoeffizienten r_{xy} läßt sich der gesuchte Regressionskoeffizient nun ohne Schwierigkeiten bestimmen, so daß wir durch Einsetzen in die Regressionsgleichung für jeden beliebigen X-Wert den dazugehörigen Y'-Wert schätzen können.

In vielen Fällen ist es jedoch erforderlich, den Regressionskoeffizienten direkt zu bestimmen, ohne vorher r_{xy} berechnet zu haben. Dann ist die Schätzung über folgende Formel vorzunehmen:

$$b_{xy} = \frac{N\sum_i X_i Y_i - NM_x M_y}{\sum_i X_i^2 - NM_x^2}$$

Oder nach der Umformung zur Rechenerleichterung:

$$b_{yx} = \frac{N\sum_i X_i Y_i - \left(\sum_i X_i\right)\left(\sum_i Y_i\right)}{N\sum_i X_i^2 - \left(\sum_i X_i\right)^2}$$

In unserem Beispiel gehen wir nach dieser Formel zur Bestimmung des Regressionskoeffizienten vor und erhalten:

$$b_{yx} = \frac{30(2231) - (105)(697)}{30(455) - (105)(105)} = \frac{-4024}{2625} = -1.53$$

e) Signifikanzprüfung

Die Signifikanzprüfung bei Regressionsproblemen erfolgt über die t-Statistik (s. Kap. 5.3.2).

Der empirische t-Wert berechnet sich nach der Formel:

$$t = \frac{b_{yx}s_x\sqrt{N-2}}{s_{xy}}$$

s_{yx} stellt dabei den Standardschätzfehler dar, der dadurch entsteht, daß die Werte von Y durch die Regressionsgerade vorhergesagt werden, experimentell aber einen Punkteschwarm um diese Gerade herum bilden. Die Formel für s_{yx} lautet:

$$s_{yx} = s_y\sqrt{1 - r_{xy}^2}$$

In Tab. III findet sich entsprechend dem vorher festgelegten Signifikanzniveau und den Freiheitsgraden $N - 2$ der theoretische Vergleichswert.

Wenn der empirische t-Wert kleiner ist als der theoretische, muß H_0 angenommen werden, d.h., es besteht kein signifikant *linearer* Zusammenhang zwischen der abhängigen und der unabhängigen Variablen. Es heißt jedoch nicht, daß gar kein Zusammenhang zwischen beiden Variablen besteht. Es ist durchaus möglich, daß die abhängige Variable eine Funktion höherer Ordnung der unabhängigen ist (Kurvilinearität). Der Signifikanztest zur linearen Regression prüft lediglich die Wahrscheinlichkeit eines *linearen* Zusammenhanges. Wenn der empirische t-Wert größer ist als der entsprechende Tabellen-Wert, dann kann mit der vorher festgelegten Wahrscheinlichkeit angenommen werden, daß die abhängige Variable eine lineare Funktion der unabhängigen ist, d.h., daß die Werte von Y befriedigend genau durch die Kenntnis der X-Werte vorherbestimmt werden können. Ein signifikantes t-Test-Ergebnis besagt jedoch noch nicht, daß die Regressionslinie eine optimale Anpassung an den Punkteschwarm im *Scatter*diagramm darstellt; sie ist die beste *lineare* Anpassung. Funktionen höherer Ordnung lassen möglicherweise eine exaktere Schätzung der Y-Werte aufgrund der X-Werte zu. (Zur Bestimmung kurvilinearer Regressionslinien s.u.a. Hays 1969.)
Für unser Beispiel errechnet sich ein empirischer t-Wert von $t = 6.72$. Bei einem Signifikanzniveau von .05 und den Freiheitsgraden $df = 28$ beträgt der t-Wert in Tab. III $t = 2.05$.

f) Schätzung der Vertrauensintervalle für den Regressionskoeffizienten
Da die Berechnung von b_{yx} nur eine Schätzung der „wahren" Regressionskoeffizienten ist, interessiert uns jetzt die Frage, innerhalb welchen Bereiches um den geschätzten Wert von b_{yx} herum der „wahre" Regressionskoeffizient mit einer vorher festgelegten Wahrscheinlichkeit liegt. Je enger der Vertrauensbereich ist, um so sicherere Vorhersagen können gemacht werden. Der „wahre" Regressionskoeffizient wird mit dem griechischen Buchstaben β bezeichnet. Der Wahrscheinlichkeitsbereich (das Vertrauensintervall) wird wie folgt bestimmt:

$$b_{yx} - \frac{\sigma_{yx} t}{s_x \sqrt{N}} \leqq \beta_{yx} \leqq b_{yx} + \frac{\sigma_{yx} t}{s_x \sqrt{N}}$$

σ wird dabei geschätzt nach der Formel:

$$\sigma_{yx} = \sqrt{\frac{N s_y^2 (1 - r_{xy}^2)}{N - 2}}$$

Der t-Wert entspricht jeweils dem halben vorher festgelegten Signifikanzniveau, da die Vertrauensintervallbestimmung mit einer zweiseitigen Fragestellung verglichen werden kann (der Wahrscheinlichkeitsbereich liegt zwischen zwei Grenzwerten).
In unserem Beruhigungsmittel-Beispiel liegt der „wahre" Regressionskoeffizient mit 95% Wahrscheinlichkeit zwischen den Grenzwerten von − 3.06 und 0.00.

(3) *Anwendungsregeln*

a) Festlegung des Signifikanzniveaus, von dem an ein festgehaltener linearer Zusammenhang zwischen der unabhängigen und der abhängigen Variablen überzufällig sein soll
b) Bestimmung der Regressionsgeraden
c) Berechnung von b_{yx} nach einer der beiden möglichen Formeln
d) Berechnung des empirischen t-Wertes zur Signifikanzprüfung

e) Aufsuchen des theoretischen t-Wertes in Tab. III unter Berücksichtigung der Freiheitsgrade und dem 5%- oder 1%-Signifikanzniveau
f) Vergleich der beiden t-Werte,
wenn $t_{emp} < t_{theo}$: H_0 trifft zu [1]
wenn $t_{emp} \geqq t_{theo}$: H_1 trifft zu
g) Schätzung des Vertrauensintervalls für den „wahren“ Regressionskoeffizienten nach der o.a. Formel

(4) *Voraussetzungen und Einschränkungen*

a) Beide in Frage kommenden Variablen müssen numerisch erfaßbar sein.
b) Innerhalb jeder Stichprobe der X-Werte müssen die Y-Werte normal verteilt sein.
c) Die Varianzen innerhalb aller X-Kategorien müssen gleich sein.
d) Die Stichproben zu allen X-Kategorien müssen unabhängig voneinander sein (experimentelle Bedingung).
e) Voraussagen über Y-Werte können nur aufgrund der in die Regressionsanalyse eingegangenen X-Werte gemacht werden.
f) Es können nur Vorhersagen in eine Richtung getroffen werden, nämlich für die Y-Werte aufgrund der X-Werte, nicht umgekehrt.
g) Die Aussagen über den gerichteten Zusammenhang zwischen abhängiger und unabhängiger Variablen beziehen sich ausschließlich auf einen *linearen* Zusammenhang.

(5) *Übungsbeispiel*

Nr. 4/10
Berechnen Sie die Erwartungswerte im Lesetest (y') für alle 18 Schüler, und vergleichen Sie diese mit den tatsächlich erreichten Lesetestwerten (y)!

Schüler Nr.	IQ x	Lesetest-wert y	x^2	xy	Erwartungswert y'
1	118	66	13924	7788	
2	99	50	9801	4950	
3	118	73	13924	8614	
4	121	69	14641	8349	
5	123	72	15129	8856	
6	98	54	9604	5292	
7	131	74	17161	9694	
8	121	70	14641	8470	
9	108	65	11664	7020	
10	111	62	12321	6882	
11	118	65	13924	7670	
12	112	63	12544	7056	
13	113	67	12769	7571	
14	111	59	12321	6549	
15	106	60	11236	6360	

---
1 t_{emp} = der empirisch ermittelte t-Wert
t_{theo} = der theoretische Tabellenwert für t

Schüler Nr.	IQ x	Lesetest-wert y	x^2	xy	Erwartungswert y'
16	102	59	10404	6018	
17	113	70	12769	7910	
18	101	57	10201	5757	
Summe	2024	1155	228978	130806	

Beispiel nach Ferguson (1971)

4.2.2. Multiple Regression

(1) *Untersuchungsproblem*

Das Untersuchungsproblem ist im Prinzip das gleiche wie bei der einfachen linearen Regression, nur mit dem Unterschied, daß jetzt eine lineare Funktion gefunden werden soll für die Daten einer abhängigen Variablen, die unter verschiedenen Bedingungen *mehrerer* unabhängiger Variablen, gleichzeitig erhoben wurden. Es soll eine lineare Funktion gefunden werden, aufgrund deren die beste Vorhersage für die Werte der abhängigen Variablen bei Kenntnis der Werte mehrerer unabhängiger Variablen getroffen werden kann.

(2) *Beispiel*

a) Versuchsplan
abhängige Variable: Tüchtigkeits-Wert
unabhängige Variable 1: Intelligenztestwert
unabhängige Variable 2: Wert in einem Test zur Kontaktfähigkeit
Versuchspersonen: Bewerber für eine Stelle als Kindergartenleiter(in)
Der zuständige Dezernent möchte aufgrund der beiden Informationen über Bewerber (Intelligenztest-Wert und Wert in dem Kontaktfähigkeits-Test) eine Regressionsgleichung finden, mit deren Hilfe er die bestmögliche Vorhersage für den Berufserfolg (bestimmt durch den Tüchtigkeitswert) des Bewerbers/der Bewerberin treffen kann.

b) Hypothesen
H_0: Es besteht kein linearer Zusammenhang zwischen der abhängigen Variablen einerseits und den beiden unabhängigen Variablen andererseits. Es können aufgrund der beiden experimentell unabhängigen Variablen anhand einer linearen Funktion keine Vorhersagen für die experimentell abhängige Variable getroffen werden.
H_1: Es besteht ein statistisch nachweisbarer linearer Zusammenhang zwischen der abhängigen Variablen einerseits und den beiden unabhängigen Variablen andererseits. Es können mit einer bestimmten Wahrscheinlichkeit Vorhersagen anhand einer linearen Funktionsgleichung getroffen werden, wenn die Wertekombinationen der beiden unabhängigen Variablen bekannt sind.

c) Ergebnisse
Die graphische Darstellung der Ergebnisse ist im Falle der multiplen Regressionsanalyse etwas komplizierter als bei der einfachen Regression, da wir es hier mit

mindestens drei Dimensionen zu tun haben, der Punkteschwarm also nicht mehr auf einer Ebene darzustellen ist. Die tabellarische Darstellung erfordert eine entsprechend größere Datenmatrix. Für unser Beispiel sähe sie z. B. wie folgt aus:

Intelligenztestwert

Kontaktfähigkeitswert	Vp	90	100	110	120
	1	x_{111}	x_{211}	x_{311}	x_{411}
	2	x_{112}	x_{212}	x_{312}	x_{412}
	3	x_{113}	x_{213}	x_{313}	x_{413}
10	4	.	.	.	.
	5	.	.	.	.
	6	.	.	.	.
	7	.	.	.	.
	8	.	.	.	.
	1	x_{121}	x_{221}	x_{321}	x_{421}
	2	x_{122}	x_{222}	x_{322}	x_{422}
	3	x_{123}	x_{223}	x_{323}	x_{423}
20	4	.	.	.	.
	5	.	.	.	.
	6	.	.	.	.
	7	.	.	.	.
	8	.	.	.	.
	1	x_{131}	x_{231}	x_{331}	x_{431}
	2	x_{132}	x_{232}	x_{332}	x_{432}
	3	.	.	.	.
30	4	.	.	.	.
	5	.	.	.	.
	6	.	.	.	.
	7	.	.	.	.
	8	.	.	.	.

d) Berechnung

Da es sich bei der multiplen Regression um mehr als zwei zu untersuchende Variablen handelt, wollen wir statt der bisher üblichen Kennzeichnung X für die unabhängige und Y für die abhängige Variable durchgehend abhängige und unabhängige Variablen mit X kennzeichnen und mit 1 über k bis K indizieren. Dabei steht X_1 jeweils für die abhängige und $X_2 \ldots X_K$ für die unabhängigen bzw. z_1 für den Standardwert der abhängigen und $z_2 \ldots z_K$ für die Standardwerte der unabhängigen Variablen.

Die einfache lineare Regressionsgleichung (nach der neuen Kennzeichnung: $z'_1 = b_{1.2}\, z_2$) erweitert sich im Falle einer multiplen Regression entsprechend zu:

$$z'_1 = z_2(b_{12.3\ldots K}) + z_3(b_{13.2\ldots K}) + \ldots + z_K(b_{1K.2\ldots K-1})$$

Bei drei Variablen sieht die Gleichung z. B. so aus:

$$z'_1 = z_2\, b_{12.3} + z_3\, b_{13.2}$$

Damit die Gleichung die Bedingung der „kleinsten Quadrate“ erfüllt (wonach die Summe der quadrierten Differenzen zwischen tatsächlichen und vorhergesagten Werten die kleinstmögliche sein soll), müssen nun entsprechende Werte für alle b-Gewichte gefunden werden.
Die mathematische Herleitung ihrer Bestimmungsformeln würde im Rahmen dieses Buches zu weit führen, so daß wir hier nur zusammenfassend die Endformeln angeben, und diese auch nur für den Fall von 3 Variablen. Die Berechnung per Hand für mehr als 3 Variablen wäre mit einem zu großen Aufwand verbunden; für diesen Fall wird man sich einer elektronischen Datenverarbeitungsanlage bedienen. Die Gleichungen für 3 Variablen lauten:

$$b_{12.3}(r_{23}) + b_{13.2} = r_{13}$$
$$b_{12.3} + b_{13.2}(r_{23}) = r_{12}$$

Wenn also die Korrelationen der Variablen untereinander bekannt sind, lassen sich die beiden b-Werte leicht berechnen. In unserem Beispiel ergeben sich folgende Werte:

$$r_{12} = .20 \qquad r_{13} = .40 \qquad r_{23} = .30$$

$$\begin{array}{rcrl}
b_{12.3} & + & .30(b_{13.2}) & = .20 \\
.30(b_{12.3}) & + & b_{13.2} & = .40 \\
3.333\ b_{12.3} & + & b_{13.2} & = .667 \\
.300\ b_{12.3} & + & b_{13.2} & = .400 \\
\hline
3.033\ b_{12.3} & & & = .267 \\
b_{12.3} & & & = .088 \\
b_{13.2} & & & = .374
\end{array}$$

Demnach lautet die multiple, lineare Regressionsgleichung:

$$z'_1 = (.088)z_2 + (.374)z_3$$

D.h.: Um eine optimale lineare Schätzung für den Tüchtigkeitswert zu finden, muß der Intelligenzwert mit .088 und der Kontaktfähigkeitswert mit .374 gewichtet werden (gemeint sind jeweils die Standardwerte!).
Wenn nun ein Stellenbewerber einen Standard-Intelligenzwert von 1.5 und einen Standard-Kontaktfähigkeitswert von −.3 hat, so berechnet sich der wahrscheinlichste Tüchtigkeitswert zu $z'_1 = .02$. Wenn aber die Standardwerte für Intelligenz und Kontaktfähigkeit vertauscht sind, ergibt sich ein vorhergesagter Tüchtigkeitswert von $z'_1 = .53$. Daraus läßt sich ableiten, daß die Kontaktfähigkeit für die eingeschätzte Tüchtigkeit gewichtiger ist als die Intelligenz.

(3) *Anwendungsregeln*

a) Die Korrelationen zwischen X_1, X_2 und X_3 werden berechnet.
b) Die Gewichte $b_{12.3}$ und $b_{13.2}$ werden bestimmt.
c) Die berechneten Gewichte werden in die multiple lineare Regressionsgleichung eingesetzt, so daß die jeweils optimale lineare Schätzung für den Wert der abhängigen Variablen bei bekannten Werten der unabhängigen Variablen erfolgen kann.

(4) *Voraussetzungen und Einschränkungen*

Siehe Kap. 4.2.1 (einfache lineare Regression).

(5) *Übungsbeispiele*

Nr. 4/11 a

Es soll der Schulerfolg (Abschlußzensur = X_1) im Fach Deutsch vorhergesagt werden für einen Schüler, der im Intelligenztest (X_2) einen Standardwert von 1.7 und in einem reinen Wortschatztest (X_3) einen Wert von .9 erreicht.
Die Korrelation zwischen Deutschzensur und Intelligenzwert soll .54 betragen, zwischen Deutschzensur und Wortschatztest .70 und zwischen Wortschatztest und Intelligenztest ebenfalls .70.

Nr. 4/11 b

Es soll der Schulerfolg (Abschlußzensur = X_1) im Fach Deutsch vorhergesagt werden für einen Schüler, dessen Werte im Intelligenz- und Wortschatztest genau umgekehrt sind wie die des Schülers im Beispiel Nr. 4/10a. Welcher Test ist für die Vorhersage der Deutschzensur effektiver?

5. Analytische oder Inferenzstatistik (Stichprobenstatistik)

Die in Kap. 3 und 4 behandelten Methoden werden allgemein der *deskriptiven* Statistik zugeordnet. Mit deren Hilfe lassen sich empirisch gewonnene Daten graphisch und/oder numerisch *beschreibend* ordnen. Dagegen ist die *Inferenzstatistik* (als mathematische oder analytische Statistik, unter anderem Aspekt auch als Stichprobenstatistik bezeichnet) die Lehre von den statistischen Schlüssen. Auf der Basis der Wahrscheinlichkeitstheorie lassen sich mit ihrer Hilfe nunmehr Stichprobenergebnisse verallgemeinern. Dies wäre (allein) auf der Grundlage der deskriptiven Statistik nicht möglich bzw. würde die Gefahr grober Fehlschlüsse in sich bergen. Kennzeichnend für die Inferenzstatistik sind keine Wenn-Dann-Beziehungen (vgl. Korrelation und Regression), sondern Regeln, aufgrund deren Anwendung entschieden werden soll, ob ein empirisch gefundenes (Stichproben-) Ergebnis als zufällig vs. überzufällig betrachtet werden kann. Die Gegenstände der sozialen Verhaltenswissenschaften sind ja meistens sehr komplexer Natur, denen eine mehr oder minder große Zahl von Faktoren und somit ein weiter Raum für den Zufall inhäriert. Bei der Feststellung entsprechender Gesetzmäßigkeiten oder Ordnungszusammenhänge ist es deshalb wichtig nachzuweisen, daß die gefundenen Ergebnisse überzufälliger Art sind, d.h. nach der Wahrscheinlichkeit kein Zufall die Ergebnisse bestimmt.

Hierzu dienen nun die verschiedenen Methoden der Inferenz- oder Stichprobenstatistik (vgl. Kap. 5.3 und 5.4). Bevor wir diese ausführlicher darstellen, sei zunächst auf die Axiomatik der Wahrscheinlichkeitstheorie und im Zusammenhang damit auf die wichtigsten Modelle für Zufallsereignisse kurz eingegangen. Hierbei kommt dem Begriff der Normalität oder Normalverteilung fundamentale Bedeutung zu.

5.1. Theoretische Modelle der Häufigkeitsverteilung

5.1.1. Zum Begriff der Normalität

Der Begriff der Normalität ist mehrdeutig. Hofstätter (1957, S. 217ff.) unterscheidet drei Begriffsvarianten: Normalität im statistischen, im idealistischen und im funktionalistischen Sinne.

Der *statistische* Normenbegriff bezieht sich auf die Häufigkeit. Eigenschaften, Verhaltensweisen u.ä., die am häufigsten vorkommen, d.h. im Durchschnittsbereich – verteilungsmäßig gesehen – liegen, werden als „normal“ interpretiert; solche Merkmale, die im (oberen oder unteren) Extrembereich liegen, werden als „anormale“ Erscheinungen gewertet. Zwischen normalen und anormalen Phänomenen bestehen demnach nur graduelle, d.h. quantitative, nicht qualitative Unterschiede. Der statistische Normbegriff ist weit verbreitet, seine Anwendung erstreckt sich genauso auf die Medizin wie auf die Psychologie, Soziologie u.ä. Verhaltenswissen-

schaften. Somatische Erkrankungen, Psychosen, Neurosen, Kriminalität, Suizidalität u.dgl.m. sind anormale Erscheinungen im statistischen Sinne. Der statistische Normbegriff wird freilich immer dann zum Problem, wenn ehemals „anormale" (seltene) Merkmale bzw. Merkmals- und Verhaltensausprägungen sich zu „normalen" Erscheinungen entpuppen. Am Maßstab der statistischen Norm gemessen, wäre beispielsweise Karies nicht mehr als anormales Phänomen (Krankheit) zu deuten, da dieser Befund heute zu den Regelsymptomen gehört, d.h. bei der Mehrheit der zivilisierten Bevölkerung vorkommt. Ähnliche Trends zeichnen sich ab in bezug auf (leichtere) Sehschäden, Haltungsfehler, Überempfindlichkeit u.ä.
Die *ideale* Norm fixiert ein wünschenswertes, wenngleich selten erreichtes Ziel der Vollkommenheit. Idealnormen beeinflussen auf mannigfache Weise unser Handeln, ihnen kommt „Aufforderungscharakter" im Sinne Lewins zu. Entsprechenden Tugendlehren kann eine idealistische vs. eine realistische Wendung inhärieren. In der Ethik Platons bzw. in der Nachfolge bei Augustinus ist die Vollkommenheit i.e.S., also die das Ideal voll erreichte Verhaltensweise, das erklärte Ziel (= idealistische Ethik). So hätte jemand das Ideal der Wahrhaftigkeit (erst) dann erreicht, wenn er *nie* die Unwahrheit sagen oder, genauer bei jeder Gelegenheit die Wahrheit äußern bzw. stets völlig offen, d.h. nie etwas verbergend, handeln würde. Dem Psychologen oder Psychiater, der sich der statistischen Norm verbunden weiß, käme ein solches Individuum höchst verdächtig vor; Übersteigerungen dieser Art sind meistens krankhafte Befunde oder werden wenigstens als solche, z.B. im Rahmen eines Zwangssyndroms, gewertet. Wie alle Erfahrung lehrt, sind idealistisch postulierte Normen Ideale, die kaum jemals i.e.S. erfüllt werden können. Das bedeutet jedoch nicht, daß sie unwirksam wären. Sie üben vielmehr einen starken Einfluß im Sinne der Verhaltenssteuerung aus.
In der Tugendlehre des Aristoteles (der sog. Nikomachischen Ethik) bzw. in der Nachfolge bei Thomas v. Aquin ist der Mensch dann tugendhaft, wenn er die *richtige Mitte* zwischen den Extremen einhält (= *realistische* Ethik). Wahrhaftig ist demnach derjenige, der sich weder naiv bei jeder Gelegenheit offenbart bzw. bloßstellt noch hinterhältig andere über seine Einstellungen, Verhaltensweisen oder Absichten zu täuschen versucht. Entsprechend wäre ein Individuum dann als freigebig zu qualifizieren, wenn es die Waage zwischen Verschwendung und Geiz hält. Oder: Tapferkeit bedeutete ein ausgewogenes Verhältnis (zur Mitte) zwischen Tollkühnheit und Feigheit, das Ideal demokratischer Erziehung läge zwischen autoritärem und Laisser-faire-Verhalten usw.
Idealnormen – in der idealistischen oder realistischen Deutung – erlangen besonders im Bereich der Sozialisation und deren Institutionen (Familie, Schule, Staat u.ä.) Einfluß und somit Macht über das Verhalten des Individuums. „Soziale Systeme jeder Größenordnung bedienen sich idealer Normen zur Steuerung des Verhaltens ihrer Angehörigen, von denen sie in einzelnen Daseinsbereichen absolute, in anderen nur eine weitgehende Konformität verlangen" (Hofstätter 1957, S. 218f.). Nach F. H. Allport manifestiert sich die Wirksamkeit institutioneller Zwänge häufig im Typus der sog. *J-Kurve*, einer für die geschilderten Phänomene charakteristischen Verteilungsform. Am Beispiel der (Un-)Pünktlichkeit von Studenten, die eine Vorlesung besuchen, möge die institutionelle Einengung der „natürlichen" Variationsbreite (vgl. Normalverteilungskurve) – in Anlehnung an das bei Hofstätter zitierte, hier jedoch modifizierte Beispiel – veranschaulicht werden. Analoge Ereignisverteilungen repräsentieren Verhaltensweisen an der Verkehrsampel (bei

Umschaltung auf Rotlicht), das Schülerverhalten lt. Anweisung des Lehrers, z.B. Mitarbeit im Unterricht, Anfertigen vs. Vergessen oder Unterlassen von Hausaufgaben u.ä.

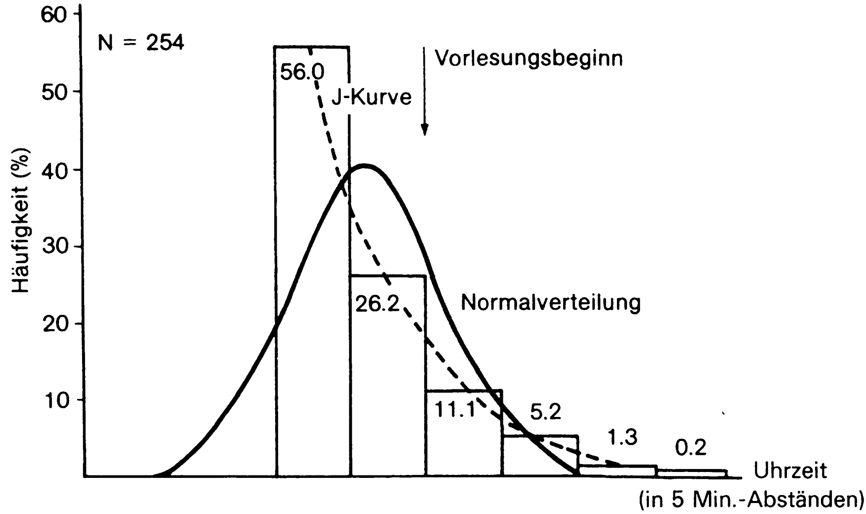


Abb. 23: Die J-Kurve nach F. H. Allport (1934)

Schließlich verweist Hofstätter noch auf die *funktionale* Norm. Diese „verzichtet auf die Herstellung eines Bezuges zwischen den Erscheinungs- und Verhaltensweisen des Individuums und denen der Allgemeinheit einerseits (statistische Norm) oder einer absoluten Wertlehre andererseits (Idealnorm); sie definiert als ‚normal' vielmehr den einem Einzelwesen hinsichtlich seiner Zielsetzungen und Leistungen gemäßen Zustand. Während z.B. die ideale Norm die Feiertagsarbeit untersagt und die statistische Betrachtungsweise deren relativ geringe Häufigkeit feststellen kann, mag es zu den Besonderheiten einer bestimmten Persönlichkeit gehören, daß diese gerade an Feiertagen gern, erfolgreich und ohne ihre Gesundheit zu schädigen schafft, so daß diese Arbeitszeit für sie unter dem funktionalen Aspekt als normal anzusprechen wäre" (a.a.O.).

Psychologisch oder soziologisch vs. pädagogisch bedeutsam ist vor allem die Unterscheidung in statistische und ideale Normen resp. deren quantitative und qualitative Abweichungen (= anormale bzw. abnorme Verhaltensweisen), die freilich interdependente Bezüge aufweisen, worauf wiederum Hofstätter aufmerksam machte. So steht „die Schwere des Verstoßes gegen eine ideale (sittliche) Norm der Seltenheit solcher Abweichungen proportional" gegenüber. Mord und Totschlag werden stärker verabscheut als Zerstörung oder Diebstahl. Andererseits bedeutet die Zunahme bestimmter Delikte nicht selten auch eine Verunsicherung individueller Verhaltensorientierung an der entsprechenden Idealnorm, gegenwärtig recht deutlich bei Eigentumsdelikten zu beobachten. Interdependente Einstellungs- und Verhaltensänderungen ließen sich auch am Beispiel des Drogenmißbrauchs, sexueller Perversionen usw. belegen, ohne daß wir hierauf näher eingehen können. Bei den folgenden Erörterungen werden wir uns ausschließlich auf den statistischen Normalitätsbegriff beziehen.

5.1.2. Normalverteilung (Gaußsche Glockenkurve)

5.1.2.1. Herleitung der Normalverteilung

Für die statistische Schlußfolgerung und somit die Anwendung statistischer Tests im Sinne von Entscheidungstechniken erlangen die theoretischen Verteilungsmodelle für Zufallsereignisse besondere Bedeutung. „Der Begriff *Zufall* bezieht sich zunächst auf ‚Mengen' von Ereignissen oder Fällen, die voneinander unabhängig sind, d.h. bei denen ein Ereignis oder Fall auf kein anderes Ereignis oder keinen anderen Fall einen vorhersagbaren Einfluß ausübt. Die im Zufallsmodell betrachteten Ereignisse sind demnach unverbunden, unabhängig. Die Unabhängigkeit der Ereignisse untereinander ist eine wichtige Bedingung für den Zufall" (Fröhlich & Becker 1971, S. 94).

In der empirischen Forschung gewinnt man die Untersuchungsdaten, also die gewünschten Informationen über den Untersuchungsgegenstand, in der Regel anhand von Stichprobenerhebungen (vgl. Kap. 2.2.5.1). Hierbei dient eine mehr oder minder große Zahl einzelner Ereignisse – im Extrem ein einziger Fall – als Grundlage für Verallgemeinerungen, d.h. Schlußfolgerungen im Hinblick auf die Populationsverhältnisse. Das Fehlerrisiko ist bei solchem Vorgehen um so größer, je kleiner die Stichprobe (Zahl beobachteter Einzelfälle oder Ereignisse) gewählt wurde, vs. um so kleiner, je größer die Stichprobe oder Zahl der beobachteten Ereignisse ist. Dieser Zusammenhang besteht strenggenommen nur für eine echte Zufallsauswahl oder Zufallsstichprobe. Für die Verteilung so gewonnener Zufallsereignisse lassen sich – theoretisch und empirisch – verschiedene Modelle aufstellen, deren wichtigste Form, die sog. Normalverteilung (normale Wahrscheinlichkeitskurve), Ziel unserer folgenden Erörterungen sein soll. Die *Normal*verteilungskurve erlangt zweifellos die größte Bedeutung im Hinblick auf die Stichprobenstatistik, wenngleich es auch andere Verteilungsmodelle gibt.

„Mathematiker haben verschiedene Formen solcher Zufallsmodelle mit den dazugehörigen Gleichungen beschrieben. Wir reden in diesem Zusammenhang deshalb von Modellen, weil sich die dazugehörigen Gleichungen auf bestimmte Idealformen der Verteilung von Zufallsereignissen beziehen. Modelle, die mit den auftretenden Ereignissen oder Ereignisfolgen vollständig übereinstimmen, gibt es allerdings nur in unserer Vorstellung. Eine volle Übereinstimmung im Sinne eines Abbildes des Beobachteten im Modell können wir deshalb nicht voraussetzen. Viel eher werden wir von näher zu bestimmenden, annähernden Übereinstimmungen ausgehen können. *Immerhin erfahren wir durch Anwendung dieser Idealgebilde, die wir Modelle nennen, wie sich eine Menge von Zufallsereignissen unter den ideal gesehenen Bedingungen des Zufalls verteilen*" (Fröhlich & Becker 1971, S. 95).

Zur Herleitung der *Normalverteilung* bzw. *normalen Wahrscheinlichkeitskurve* bedienen wir uns zunächst eines einfachen Gedankenexperiments: des Münzenwurfs. Das Beispiel des Münzenwerfens ist zugleich ein Paradigma für die Häufigkeitsverteilung qualitativer Variablen und besonders im Hinblick auf kleinere Stichproben von Alternativereignissen, z.B. die Erwartung eines Sohnes oder einer Tochter, interessant.

Die hier interessierende Frage lautet: Wie wahrscheinlich ist es, daß beim einmaligen Werfen einer Münze die Ziffernseite (Z) erscheint vs. die Wappenseite (W) nicht erscheint (oder ein Sohn vs. eine Tochter geboren wird)?

Die gesuchte Wahrscheinlichkeit (p) errechnet sich aus dem Quotienten

$$p = \frac{g}{m}$$

bzw. die Gegenwahrscheinlichkeit (q) aus:

$$q = \frac{\text{nicht } g}{m}$$

Hierbei bedeuten:
p = (Zufalls-)Wahrscheinlichkeit – von „probabilitas" – des einen Ereignisses, z.B. für das Erscheinen der Ziffernseite;
q = Gegenwahrscheinlichkeit, d.h. die Zufallswahrscheinlichkeit dafür, daß Z (Ziffernseite) nicht erscheint, sondern daß W (Wappenseite) erscheint;
g = Anzahl der *günstigen* (erwünschten oder erwarteten) Fälle;
m = Anzahl der *möglichen* Fälle.

5.1.2.2. Theoretische und empirische Wahrscheinlichkeit (Definitionen)

Mit der Formel $p = \frac{g}{m}$ haben wir die klassische Definition der *theoretischen* Wahrscheinlichkeit oder Wahrscheinlichkeit a priori – für Alternativereignisse – angegeben.
Davon abzuheben wäre die moderne Definition der *empirischen* Wahrscheinlichkeit oder Wahrscheinlichkeit a posteriori, für die nach R.v. Mises (zit. bei Bartel 1971, S. 54) folgende Formel gilt:

$$p(X) = \lim_{n \to \infty} \frac{f(X)}{n}$$

Hierbei wird die (empirische) Wahrscheinlichkeit p(X) definiert als „Grenzwert aus der Zahl der Fälle, in denen das Ereignis beobachtet wird, dividiert durch die Anzahl der Fälle, in denen es hätte auftreten können" (a.a.O.). In obiger Formel steht der Ausdruck f(X) für *absolute* Häufigkeiten, der Quotient $\frac{f(X)}{n}$ für *relative* Häufigkeiten. Dabei gilt folgende Beziehung: „Je größer die Anzahl n ist, desto mehr nähert sich der Quotient $\frac{f(X)}{n}$ einem Grenzwert, der theoretischen Wahrscheinlichkeit."
Im folgenden wenden wir uns wieder der klassischen Definition der A-priori-Wahrscheinlichkeit zu. Im Beispielfall des (einmaligen) Münzenwerfens bestehen nur zwei Möglichkeiten – Ziffern oder Wappen –, also ist:

$$p_z \text{ bzw. } p = \frac{1}{2} = 0.5 = .5$$

Entsprechend wäre: q_W bzw. $q = \frac{1}{2} = 0.5 = .5$

In allen Fällen ist $p + q = 1$, was man in der Mathematik als „sicheres" Ereignis bezeichnet.
Wählen wir statt der Münze komplexere Beispiele, wie Spielwürfel oder verschieden-

farbige Kugeln in einer Urne, so können wir unsere Fragestellung erweitern, z.B.: Wie groß ist die Wahrscheinlichkeit, daß bei einmaligem Würfeln eine 6 erscheint?

$$p_6 = \frac{1}{6} = 0.167$$

Oder: Wie groß ist p für das Erscheinen (Würfeln) einer 3 vs. einer 1 usw?

$$p_3 = \frac{1}{6} = 0.167 \qquad p_1 = \frac{1}{6} = 0.167 \quad \text{usw.}$$

Entsprechend könnten wir im Urnenbeispiel die theoretische Wahrscheinlichkeit (Wahrscheinlichkeit a priori) vorausbestimmen. Dazu benötigen wir – fiktiv – eine Urne mit beispielsweise 80 Kugeln, und zwar 8 weißen (w), 24 schwarzen (s) und 48 roten (r) Kugeln. Die Kugeln sollen sich nur hinsichtlich der Farbe unterscheiden, ansonsten jedoch gleich groß und gleich schwer sein. Ferner müssen wir eine exakte Zufallswahl annehmen, d.h., jede Kugel muß die gleiche Chance – durch Mischung – für den blinden Zugriff haben. Frage: Wie groß ist die Wahrscheinlichkeit, beim einmaligen blinden Zugriff eine weiße Kugel zu fassen?

$$p_w = \frac{8}{80} = \frac{1}{10} = 0.1 = .1$$

Entsprechend beträgt p für das Fassen einer schwarzen vs. roten Kugel:

$$p_s = \frac{24}{80} = .3 \qquad p_r = \frac{48}{80} = .6$$

Diese a priori bestimmten Wahrscheinlichkeiten könnten wir sehr leicht auf empirischem Wege kontrollieren und somit die A-posteriori-Wahrscheinlichkeit bestimmen. Dazu müßten wir allerdings jeweils die herausgegriffene Kugel wieder in die Urne zurückgeben und den Zugriff-Versuch genügend oft wiederholen, etwa 1000- oder besser 10000mal. Analog wäre beim Würfeln und Münzenwerfen zu verfahren. Nach dem Gesetz der großen Zahl würde die Abweichung der empirisch ermittelten von der theoretisch berechneten Wahrscheinlichkeit bei steigenden Versuchszahlen geringer werden, d.h., mit zunehmend geringerer Abweichung würden wir in 10% der Fälle weiße Kugeln, in 30% der Fälle schwarze und in 60% der Fälle rote Kugeln erwischen. Analog erhielten wir Ziffern und Wappen im Verhältnis (etwa) 50:50, dem ja auch die Verteilung der Geschlechtsvariablen entspricht. Mit Sicherheit würden wir entweder Ziffer oder Wappen erzielen vs. eine weiße oder eine schwarze oder eine rote Kugel fassen usw. Der statistische Begriff der *Sicherheit* steht ausschließlich für 100%ige Wahrscheinlichkeit. Diese ist immer 1, beispielsweise beim Spielwürfel: $p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1$.

5.1.2.3. Zusammengesetzte Wahrscheinlichkeiten (Additions- und Multiplikationssatz der Wahrscheinlichkeitstheorie

Das Würfelbeispiel erweiternd, könnten wir nunmehr fragen: Wie groß ist die Wahrscheinlichkeit, eine 1 *oder* eine 6 zu würfeln? In diesem Falle sind zwei erwartete und sechs mögliche Fälle gegeben, also:

$$p = \frac{2}{6} = \frac{1}{3}$$

Entsprechend bestimmt sich die Wahrscheinlichkeit, eine 1 *oder* eine 3 *oder* eine 6 zu würfeln, nämlich:

$$p = \frac{3}{6} = \frac{1}{2}$$

Auf das Urnenbeispiel angewandt, könnten wir fragen: Wie groß ist die Wahrscheinlichkeit, beim einmaligen Zugriff eine weiße *oder* eine schwarze Kugel zu fassen?

$$p = \frac{8}{80} + \frac{24}{80} = .4$$

Für die Bestimmung der Wahrscheinlichkeit von *Alternativ*ereignissen gilt somit der *Additionssatz* oder das *Entweder-Oder-Gesetz* der Wahrscheinlichkeitslehre: *Die Wahrscheinlichkeit von Alternativereignissen erhält man, indem man die einzelnen Wahrscheinlichkeiten addiert.*

Auch die zusammengesetzte Wahrscheinlichkeit aller möglichen Ereignisse muß wiederum 1 ergeben; so erwische ich mit Sicherheit eine weiße oder schwarze oder rote Kugel beim Zugriff.

Schwieriger ist der folgende Fall, in dem nach der Wahrscheinlichkeit von Ereignis-*Wiederholungen* gefragt wird. Zur Veranschaulichung diene wiederum das Urnenbeispiel, diesmal enthalte die Urne 100 Kugeln: 50 rote und 50 schwarze Kugeln. Somit wäre:

$$p_r = .5 \qquad p_s = .5 \qquad p_r + p_s = 1$$

Wir fragen nun: Wie groß ist die Wahrscheinlichkeit, daß beim blinden Zugriff zweimal hintereinander eine rote Kugel erfaßt wird?

Der erwünschte Fall ist: rr

Folgende Fälle sind möglich: rr, ss, rs, sr

Demnach ist: $p_{rr} = \frac{1}{4} = 0.25$

Nächste Frage: Wie groß ist die Wahrscheinlichkeit, daß dreimal nacheinander eine rote Kugel erfaßt wird?

Der erwünschte Fall ist: rrr

Folgende Fälle sind möglich: rrr, rrs, rsr, srr, rss, srs, ssr, sss

Demnach ist: $p_{rrr} = \frac{1}{8} = 0.125$

Entsprechend errechnet sich die Wahrscheinlichkeit, viermal nacheinander eine rote Kugel zu greifen; dabei sind 16 Fälle möglich:

$$p_{rrrr} = \frac{1}{16} = 0.055$$

Fassen wir die Beispiele zusammen, so ergeben sich folgende Wahrscheinlichkeiten:

1 rote Kugel $= \frac{1}{2} = .5$

2 rote Kugeln $= \frac{1}{4} = .25 = \frac{1}{2} \cdot \frac{1}{2}$

3 rote Kugeln $= \frac{1}{8} = .125 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$

4 rote Kugeln $= \frac{1}{16} = .055 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$

Für die Bestimmung der Wahrscheinlichkeit von Ereignis*wiederholungen* gilt somit der *Multiplikationssatz* oder das *Sowohl-Als-auch-Gesetz* der Wahrscheinlichkeits-

lehre: *Die Wahrscheinlichkeit von Ereigniswiederholungen erhält man, indem man die Einzelwahrscheinlichkeiten multipliziert.* Dieses Theorem gilt auch dann, wenn die Wahrscheinlichkeit der Einzelereignisse nicht gleich ist.
So können wir etwa fragen: Wie groß ist die Wahrscheinlichkeit, beim erstenmal eine 3 und beim zweitenmal keine 3 zu würfeln?

$$p = \frac{1}{6} \cdot \frac{5}{6} = 0.72$$

Ferner gilt der Multiplikationssatz bei simultaner Wirkung mehrerer Faktoren im Hinblick auf die Bestimmung der erwarteten Kombinationen. Beispielsweise werden zwei Münzen gleichzeitig geworfen. Die Wahrscheinlichkeit, daß bei beiden Münzen Z erscheint, ist hier p = .25, denn es gibt vier Kombinationsmöglichkeiten:

Münze	1	zeigt	Z –	Münze	2	zeigt	Z
"	1	"	W –	"	2	"	Z
"	1	"	Z –	"	2	"	W
"	1	"	W –	"	2	"	W

Entsprechend errechnet sich die Wahrscheinlichkeit bei drei simultan geworfenen Münzen für das Erscheinen jeweils der Ziffernseite (p_{ZZZ}) vs. jeweils der Wappenseite (p_{WWW}), wenn wir uns die insgesamt acht möglichen Fälle bzw. Kombinationsmöglichkeiten vorstellen:

Münze	1	zeigt	Z –	Münze	2	zeigt	Z –	Münze	3	zeigt	Z
"	1	"	Z –	"	2	"	Z –	"	3	"	W
"	1	"	Z –	"	2	"	W –	"	3	"	Z
"	1	"	W –	"	2	"	Z –	"	3	"	Z
"	1	"	Z –	"	2	"	W –	"	3	"	W
"	1	"	W –	"	2	"	Z –	"	3	"	W
"	1	"	W –	"	2	"	W –	"	3	"	Z
"	1	"	W –	"	2	"	W –	"	3	"	W

Also ist $p_{ZZZ} = \frac{1}{8}$ bzw. $p_{WWW} = \frac{1}{8}$ oder $p_{ZZW} = \frac{3}{8}$ bzw. $p_{WWZ} = \frac{3}{8}$ usw.

Sicherheit ist wie immer 1.
Schwieriger wird die Aufgabe, wenn wir beispielsweise 7 Münzen simultan werfen und wissen wollen, mit welcher Wahrscheinlichkeit sämtliche 7 Münzen – beim einmaligen Werfen – die Ziffernseite zeigen. Die Herstellung aller möglichen Fälle bzw. Kombinationen wäre nach dem bisherigen Verfahren äußerst langwierig und umständlich. In der Statistik bzw. Mathematik bedient man sich zu diesem Zweck der sog. Bernoullischen Formel und erhält so die Bernoullische oder binomische Verteilung. Die Binomialverteilung geht bei genügend großem N als Exponent (Anzahl der Münzen), etwa bei N > 30 bzw. $N \to \infty$, über in eine Normalverteilung. N bedeutet hier allgemein die Anzahl der voneinander *unabhängigen* Ereignisse oder Zufallsbedingungen, die zusammen die Maßzahl i (Anzahl der Fälle oder Treffer) bestimmen. Siehe Kap. 5.1.2.4.

5.1.2.4. Binomischer Lehrsatz (Bernoullische Formel)

Wir gehen wieder von dem Münzenbeispiel aus. Zunächst nehmen wir nur zwei Münzen. Beim simultanen Wurf von zwei Münzen ermitteln wir die Anzahl der möglichen Fälle mit folgender Formel:

$$(Z + W)^2 = Z^2 + 2\,ZW + W^2$$

Oder untereinandergeschrieben:

1mal Z^2, d.h. *ein*mal zeigen beide Münzen Z
2mal ZW, d.h. *zwei*mal zeigt sich die Kombination ZW
1mal W^2, d.h. *ein*mal zeigen beide Münzen W

Die Anzahl der möglichen Fälle ergibt – einfach aufaddiert: 1mal plus 2mal plus 1mal – somit 4.

Entsprechend lautet die Bernoullische Formel für den simultanen Wurf mit drei Münzen:

$$(Z + W)^3 = Z^3 + 3Z^2W + 3ZW^2 + W^3$$

Oder wieder untereinandergeschrieben:

1mal Z^3, d.h. *ein*mal ZZZ-Kombination
3mal Z^2W, d.h. . . .
3mal ZW^2, d.h. . . .
1mal W^3, d.h. . . .

Somit gibt es im vorliegenden Beispiel 8 Möglichkeiten – beim einmaligen Werfen – und demnach folgende Wahrscheinlichkeiten (s. a. S. 169f. oben):

$$p_{zzz} = \frac{1}{8}$$

$$p_{zzw} = \frac{3}{8}$$

$$p_{zww} = \frac{3}{8}$$

$$p_{www} = \frac{1}{8}$$

Bei n Münzen (n = Anzahl der simultan, einmalig geworfenen Münzen) und zwei Faktoren (Fall 1: In-Erscheinung-Treten des Faktors = p; Fall 2: Nicht-in-Erscheinung-Treten des Faktors = q) kann nun folgende Formel zur Berechnung der Binomialverteilung verwendet werden; nach Mittenecker (1970, S. 36) kann diese Verteilung „auch für Fälle, in denen beide Alternativen (p_{pop}, q_{pop}) nicht die gleiche Populationswahrscheinlichkeit haben, mit Hilfe der Binomialentwicklung nach Bernoulli berechnet werden“:

$$\begin{aligned}(p+q)^n = p^n + np^{n-1} \cdot q &+ \frac{n(n-1)}{1 \cdot 2}\, p^{n-2} \cdot q^2 \\ &+ \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}\, p^{n-3} \cdot q^3 \\ &+ \frac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3 \cdot 4}\, p^{n-4} \cdot q^4 + \ldots q^n\end{aligned}$$

Nach der Formel für Permutationen läßt sich dies wesentlich kürzer so schreiben:

$$n\, p_p^q = \frac{n!}{p!\,q!}$$

Dabei bedeutet $n!$ „n faktorielle" und ist $n \cdot (n-1) \cdot (n-2) \cdot (n-3) \ldots$, wobei $0! = 1$.
Allerdings müssen hierbei – genauso wie bei der Benutzung des sog. Pascalschen Dreiecks unten – beide Alternativen (p und q) jeweils gleiche Wahrscheinlichkeit haben; anderenfalls wäre die Bernoullische Formel oben zur Ermittlung der Zufallsverteilung zu verwenden.
Im Beispiel mit 10 Münzen oder Faktoren gleicher Art können demnach – beim simultanen, einmaligen Wurf – folgende Fälle bzw. Kombinationen auftreten:

$$\begin{aligned}(Z+W)^{10} = {} & \ \ 1 Z^{10} W^0 \\ & + \ 10 Z^9\ W^1 \\ & + \ 45 Z^8\ W^2 \\ & + 120 Z^7\ W^3 \\ & + 210 Z^6\ W^4 \\ & + 252 Z^5\ W^5 \\ & + 210 Z^4\ W^6 \\ & + 120 Z^3\ W^7 \\ & + \ 45 Z^2\ W^8 \\ & + \ 10 Z^1\ W^9 \\ & + \ \ 1 Z^0\ W^{10} \\ \hline & \text{1024 Möglichkeiten}\end{aligned}$$

M.a.W.: Es gibt insgesamt zwei Extremfälle: Z^{10} und W^{10}, wobei alle 10 Münzen die gleiche Seite zeigen, das eine Mal alle Z und das andere Mal alle W. Häufiger ist schon die Kombination 9Z und 1W, nämlich 10mal tritt dieser Fall auf. Die häufigste Kombination ist 5mal Z und 5mal W, diese birgt 252 Möglichkeiten.
Will man nun die Wahrscheinlichkeit einer bestimmten Kombination, z.B. p_{Z7} (d.h., daß der Faktor 7mal erscheint und 3mal nicht erscheint), bestimmen, so müssen wir wieder die 120 möglichen Kombinationen der gewünschten (erwarteten) Art in Beziehung setzen zur Anzahl der Möglichkeiten überhaupt, d.h. der Möglichkeiten, die insgesamt gegeben sind. Bei 10 verschiedenen Faktoren (z.B. 10 Münzen) können bei gleicher Wirkung oder Stärke der Faktoren insgesamt 1024 verschiedene Phänomene (Fälle) auftreten. Die Wahrscheinlichkeit des hier interessierenden Falles ist demnach:

$$p_{Z7} = \frac{120}{1024}$$

Analog kann man für jedes mögliche – aus den 10 Faktoren resultierende – Phänomen die Wahrscheinlichkeit seines Auftretens (p) errechnen:

$$p_{Z10} = \frac{1}{1024}$$

$$p_{Z9} = \frac{10}{1024}$$

$$p_{Z8} = \frac{45}{1024}$$

$$p_{Z7} = \frac{120}{1024} \quad \text{usw.}$$

Binom		Summe der möglichen Fälle (Kombinationen)
	1	
$(p+q)^1$	1 1	2
$(p+q)^2$	1 2 1	4
$(p+q)^3$	1 3 3 1	8
$(p+q)^4$	1 4 6 4 1	16
$(p+q)^5$	1 5 10 10 5 1	32
$(p+q)^6$	1 6 15 20 15 6 1	64
$(p+q)^7$	1 7 21 35 35 21 7 1	128
$(p+q)^8$	1 8 28 56 70 56 28 8 1	256
$(p+q)^9$	1 9 36 84 126 126 84 36 9 1	512
$(p+q)^{10}$	1 10 45 120 210 252 210 120 45 10 1	1024

Pascalsches Dreieck

Für den Fall, daß beide Ereignisse gleich wahrscheinlich sind ($p = q = 0.5$), kann die Verteilung der Zufallswahrscheinlichkeiten – bei nicht allzu großem N – leicht anhand des Pascalschen Dreiecks ermittelt werden. Dabei ergibt sich jede Zahl einer Zeile als Summe der beiden schräg (links und rechts) darüberstehenden Zahlen. Die Quersumme der jeweils einschlägigen Zeile – bei $N = 3$ wäre es die dritte, bei $N = 5$ die fünfte, bei $N = 10$ die zehnte Zeile unten; die oberste 1 nicht mitgerechnet – bildet dann den Nenner für die Wahrscheinlichkeitsberechnung.
So beträgt bei 3 (simultan geworfenen) Münzen die *Wahrscheinlichkeit* für 0 Treffer (d.h. daß keine Ziffer erscheint) 1/8, für 1 Treffer (d.h. daß eine Ziffer erscheint) 3/8, für 2 Treffer (d.h., daß zwei Ziffern erscheinen) 3/8 und für 3 Treffer (d.h., daß alle drei Münzen Ziffern zeigen) 1/8. Entsprechend ließen sich die Wahrscheinlichkeiten bei 10 Münzen ermitteln:

$$p_{Z0} = 1/1024, \quad p_{Z1} = 10/1024, \quad p_{Z2} = 45/1024, \quad p_{Z3} = 120/1024 \text{ usw.}$$

Die geschilderten Verhältnisse kann man auch graphisch darstellen. Dazu werden auf der Ordinatenachse (y-Achse) die Wahrscheinlichkeiten und auf der Abszissenachse (x-Achse) die Möglichkeiten des In-Erscheinung-Tretens der Faktoren abgetragen. Wir erhalten so eine symmetrische Häufigkeitsverteilung, die bei genügend großem N sich der Gaußschen Glockenkurve anpaßt; der x-Achse nähert sich unsere Kurve nur asymptotisch. Eine solche symmetrische *Wahrscheinlichkeitskurve* kommt nur unter folgenden drei *Bedingungen* zustande:
1. wenn alle Faktoren von gleicher Art und gleicher Stärke sind;
2. wenn das Erscheinen vs. Nichterscheinen gleich wahrscheinlich ist;
3. wenn die Faktoren voneinander unabhängig sind.

Zu den Eigenschaften der Normalverteilungskurve (Gaußschen Glockenkurve) s. Kap. 3.5 oben sowie Abb. 24 bzw. S. 176ff.
Obwohl die ideale *Normalverteilungskurve* (theoretische Wahrscheinlichkeitskurve) Wahrscheinlichkeitsverhältnisse a priori repräsentiert, können wir eine mehr oder weniger stark angenäherte empirische Normalverteilungskurve in bezug auf viele Phänomene (Variablen) nachweisen, d.h. solche Kurven a posteriori entstehen lassen. Im Beispielfall des Münzenwerfens wären dazu 1024 Versuche notwendig, also 1024 simultane Würfe von 10 Münzen gleicher Art und Stärke.

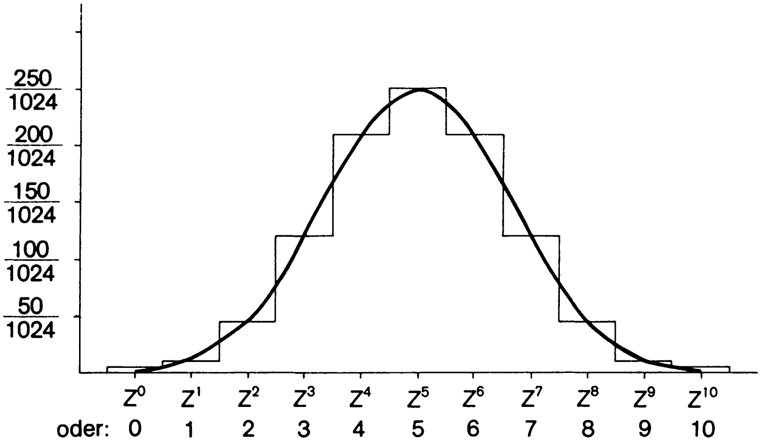


Abb. 24: Die (theoretische) Wahrscheinlichkeits- oder Normalverteilungskurve

Sofern eine genügend große Zahl von Beobachtungen bzw. Messungen vorliegt, ist demnach sehr häufig mit dem Erscheinungsbild der sog. Gaußschen Glockenkurve zu rechnen. Im Hinblick auf pädagogische, psychologische oder soziologische Fragestellungen könnte man auf normal verteilte Phänomene wie Schulleistungen, Begabungen (Schuleignungen), Intelligenz, Aufmerksamkeit, Konzentration, Leistungsmotivation, „persönliche Fehler" u.ä., aber auch auf Variablen wie Kontaktstreben, Aggression, Berufserfolg usw. verweisen. Das zunächst theoretische Modell der Normalverteilung läßt sich anhand mannigfacher Phänomene empirisch „belegen". In diesem Modell, dessen Aufweis bestimmten Bedingungen unterliegt (s. S. 173), hat der mittlere Fall stets die größte Wahrscheinlichkeit.

Das Modell der Binomial- bzw. Normalverteilung dient nicht nur zur Schätzung von Auftretenshäufigkeiten, d.h. der Berechnung von Ereigniswahrscheinlichkeiten, dasselbe Modell ist auch grundlegend für die statistische Schlußfolgerung (Inferenz), d.h. für die Verallgemeinerung von Stichprobenergebnissen. Dabei geht man in der Regel so vor, daß man die empirisch ermittelten Häufigkeiten am Modell der theoretischen Verteilung – dieser oder jener Prüfgröße – mißt, d.h. Statistiken und Parameter miteinander vergleicht (vgl. Kap. 5.2). Unser nächstes Beispiel möge das hierfür notwendige Denken vorbereiten. Zur Veranschaulichung wählen wir wiederum ein einfaches Rechen*beispiel.*

Am Ende einer Unterrichtsstunde möchte der Lehrer überprüfen, wieviel von dem behandelten Stoff behalten wurde. Zu diesem Zweck werden relevante Fragen in der Auswahl-Antwortform auf Kärtchen geschrieben und den Schülern appliziert. Der Einfachheit halber nehmen wir an, daß jeder Schüler nur 10 Aufgaben bzw. Fragen vorgelegt bekommt, deren Lösungsvorschlag (vorgegebene Antwort) er mit „Stimmt" (+) oder „Stimmt nicht" (−) bewerten soll.

Ein Schüler habe nun 8 Fragen richtig beantwortet, d.h. in 8 von 10 Fragen eine zutreffende Stellungnahme abgegeben. Ist diese Trefferzahl Zufall oder ein überzufälliges Ergebnis? Oder anders formuliert: Hat der Schüler bloß geraten oder tatsächliches Wissen reproduziert? Weiterhin könnten wir fragen, wie viele Schüler der Grundgesamtheit (Population) ein solches Ergebnis erzielten, d.h., ob 8 von 10 möglichen Treffern oder Punkten ein überdurchschnittliches, durchschnittliches oder unterdurchschnittliches Ergebnis repräsentieren. Diese oder ähnliche Fragen

lassen sich mit Hilfe der Wahrscheinlichkeitsrechnung beantworten. Zunächst wenden wir uns der ersten Fragestellung (zum zweiten Frageaspekt vgl. Kap. 5.1.2.5 unten) zu.
In unserem Fallbeispiel handelt es sich um Alternativereignisse (+ vs. −) mit zwei Entscheidungsmöglichkeiten. Die Wahrscheinlichkeit für richtige vs. falsche Reaktionen beträgt demnach $p = .5$ bzw. $q = .5$. Sofern ein Schüler seine Entscheidung ausschließlich nach dem Zufall (z. B. durch blindes Raten) treffen würde, hätte er insgesamt $2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^{10}$ Wahlmöglichkeiten. Mit Hilfe des Pascalschen Dreiecks (s. S. 173) läßt sich nun ermitteln, daß von den $2^{10} = 1024$ Kombinationsmöglichkeiten – theoretisch – eine zu 10 richtigen (und 0 falschen) Reaktionen führt, zehn zu 9 richtigen (und 1 falschen) Reaktion(en), fünfundvierzig zu 8 richtigen (und 2 falschen) Reaktionen usw. führen. Die Zufallswahrscheinlichkeit für 10 Treffer beträgt somit 1/1024, für 9 Treffer 10/1024, für 8 Treffer 45/1024 usw. In einer Serie von 10 Ereignissen (Aufgaben) sind demnach 8 Zufallstreffer relativ unwahrscheinlich, wenngleich immerhin $p = 45/1024 = 0{,}0449$ oder 4.49% beträgt. Fragen wir allerdings, wie wahrscheinlich es ist, daß unser Kandidat 8 oder mehr (9 oder gar 10) Treffer erzielt, dann beträgt die Zufallswahrscheinlichkeit dafür nunmehr 45/1024 plus 10/1024 plus 1/1024 = 56/1024 = 0.055 oder 5.5%. In über 5 von 100 Fällen wäre somit ein Ergebnis von 8 oder mehr Treffern per Zufall zu erwarten. Mit der gleichen Wahrscheinlichkeit wären nur 2 oder weniger Treffer per Zufall zu erwarten, während der Kombination 5 Treffer und 5 Fehler die größe Auftretenswahrscheinlichkeit zukäme, nämlich $p = 0.246$ oder 24.6%. In einem wissenschaftlichen Versuch müßte man bereits das Ergebnis mit 5.5% Auftretenswahrscheinlichkeit (8 und mehr Treffer) zugunsten der H_0-Hypothese, d.h. hier als Zufallsergebnis und nicht als Wissensreproduktion, interpretieren. Da das Ergebnis jedoch „hart an der Signifikanzgrenze" liegt (vgl. Kap. 5.2.2), würde sich im vorliegenden Fall eine Vergrößerung der Stichprobe, z. B. Erhöhung der Aufgabenzahl auf $N = 15$, empfehlen, um so gegebenenfalls doch noch die Nullhypothese verwerfen zu können.
Eine andere Möglichkeit, das Zufallsrisiko zu verringern, bestünde in der Umwandlung der zunächst dichotomisch abgefaßten (Ja/Nein-)Antwortform in einen sog. Mehrfachwahl-Antwortmodus, etwa mit 5 Auswahlantworten und somit 5 Wahlmöglichkeiten bei jedem der 10 Items. Bei einem gleich großen Aufgabenpaket ($N = 10$) reduzierte sich jetzt die Zufallswahrscheinlichkeit der Treffer pro Aufgabe von $p = .5$ (1/2) auf $p = .2$ (1/5). Durch Einsetzen der bekannten Daten in die Bernoullische Formel $(1/5 + 4/5)^{10}$ läßt sich wiederum für jede beliebige Trefferquote die Zufallswahrscheinlichkeit bestimmen.

Übungsbeispiele:

Nr. 5/1
Unter der Voraussetzung, daß im Populationsdurchschnitt gleich häufig Jungen und Mädchen geboren werden, wie groß ist dann die Wahrscheinlichkeit, daß in einer Familie mit 4 Kindern 3 Jungen und 1 Mädchen sind?

Nr. 5/2
Wie groß ist die Wahrscheinlichkeit bei insgesamt 10 Testaufgaben und jeweils 2 Antwortmöglichkeiten (+ vs. −), sieben oder mehr richtige Lösungen per Zufall zu erzielen?

Nr. 5/3
Wie groß wäre die Zufallswahrscheinlichkeit für 7 Treffer bei 10 Testaufgaben und jeweils 4 Wahlmöglichkeiten?

5.1.2.5. Beziehungen zwischen der Normalkurve und der z-Skala (Standardnormalverteilung)

Mit steigendem N erfordert die Benutzung der Bernoulli-Formel einen ziemlichen Rechenaufwand. In solchen Fällen nähert sich jedoch die diskrete (unstete) Binomialverteilung immer mehr dem Modell der Gaußschen Glockenkurve, die durch die Parameter $\mu_i = N \cdot p_{pop}$ und $\sigma_i = \sqrt{N \cdot p_{pop} \cdot q_{pop}}$ definiert ist; zur mathematischen Definition der Normalkurve siehe unten S. 176f. Mit μ ist hier der Mittelwert der Population bzw. der theoretischen (Normal-)Verteilung der Trefferzahlen i symbolisiert. Im früheren Beispiel der Vorgabe von 10 Alternativ-Items ($N = 10$) wären demnach bei kontinuierlicher (stetiger) Verteilung der Zufallstreffer $M = 10 \cdot 0.5 = 5$ Treffer (Richtiglösungen) am häufigsten, entsprechende Abweichungen vom Mittelwert (M) mit zunehmender Extremität – symmetrisch gleich häufig in beiden Richtungen – immer seltener zu erwarten. Die Häufigkeit der Abweichungen vom Mittelwert kann man nun in Maßeinheiten der Standardabweichung (σ) angeben gemäß der Formel $\sigma = \sqrt{N \cdot p \cdot q}$. Auf das nämliche Beispiel bezogen wäre $\sigma_i = \sqrt{10 \cdot 0.5 \cdot 0.5} = 1.581$. Dieser Wert läßt sich mit Hilfe der Formel

$$z = \frac{X - \mu}{\sigma}$$

in den Standardwert z transformieren, d.h. ausdrücken, um wieviel Einheiten der Standardabweichung dieser oder irgendein anderer Rohwert über oder unter dem Mittelwert der Bezugsgruppe liegt (s. Kap. 3.5.2 oben). Für den Anwendungsfall von 8 Treffern bzw. einem noch günstigeren Ergebnis ergäbe sich ein z-Wert von:

$$z = \frac{7.5 - 5}{1.581} = 1.58$$

Da jeder Punkt auf einer kontinuierlichen Skala bekanntlich einen Bereich darstellt, erstreckt sich der Rohwert 8 von 7.5 bis 8.5; im Beispielfall ist deshalb die untere Bereichsgrenze (7.5 RW = X) in obige z-Formel einzusetzen. Was besagt nun dieser Wert $z = 1.58$ für unser Rechenbeispiel? Bevor wir dieser Frage weiter nachgehen, sei in einem Exkurs die Funktion der Normalkurve – ergänzend zu den früheren Ausführungen auf S. 166ff. – etwas ausführlicher dargestellt.

Die mathematische Definition der Normalkurve geht auf entsprechende Wahrscheinlichkeitsberechnungen von A. de Moivre (1667–1754) und K. F. Gauß (1777–1855) zurück. Wie Abb. 25 erkennen läßt und bereits mehrfach ausgeführt wurde, gruppieren sich im Modell der Normalkurve die häufigsten Ereignisse um den Mittelwert (μ), während zunehmend seltenere Ereignisse mit zunehmend größerer Entfernung vom Mittelwert korrespondieren.

Die *Funktionsgleichung für Y* (Ordinatenhöhe) lautet.

$$Y = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{1}{2}\left(\frac{X - \mu}{\sigma}\right)^2}$$

Dabei sind:

X = Maßzahl (Wert der Zufallsvariablen X)

μ = arithm. Mittel der Verteilung von X

σ = Standardabweichung der Verteilung von X

π = 3.14159 (Konstante: Verhältnis des Kreisumfangs zum Kreisdurchmesser)

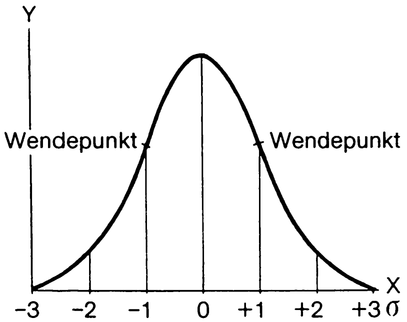


Abb. 25: Eigenschaften der Standardnormalverteilung (Normalkurve)

e = 2.71828 (Konstante: Basis der *natürlichen* Logarithmen ln, auch *E*ulersche Zahl genannt)

Aus der vorstehenden Abbildung wird deutlich, daß die Abszisse asymptotisch verläuft, d.h. ins Unendliche geht. Ferner ist ersichtlich, daß man für jeden beliebigen X-Wert den zugehörigen Y-Wert bestimmen kann.

Durch senkrechte Striche rechts und links vom Mittelwert (oberhalb vs. unterhalb vom Durchschnitt) lassen sich nun einzelne Abschnitte der Fläche unter der Normalkurve abgrenzen, die mit Hilfe der erwähnten Funktionsgleichung als Prozentanteile der Gesamtfläche (100%) berechenbar sind. Diese Werte liegen – für die Standardnormalverteilung in z-Einheiten ausgedrückt – bereits in tabellierter Form vor (vgl. Tab. II im Anhang dieses Buches). Sofern man nämlich die Abweichung der Maßzahlen in Sigma-Werten (Standardabweichung) zum Ausdruck bringt, läßt sich mit Hilfe der z-Transformationsformel jeder beliebige Meßwert (Rohwert) in den Standardwert z transformieren. Geschieht dies, d.h., werden die Abweichungswerte X in die Standardwerte z übertragen, erhalten wir die sog. *Standardnormalverteilung*. Anhand von Tab. II (s. S. 255ff.) läßt sich jetzt praktisch für jeden z-Wert der zugehörige Flächenabschnitt und somit die gesuchte Erwartungshäufigkeit (Zufallswahrscheinlichkeit) ermitteln.

Im obigen Rechenbeispiel (s. S. 176) ermittelten wir einen z-Wert von 1.58. Zu diesem Wert $z = 1.58$ können wir nunmehr anhand der Tab. II die gesuchte Zufallswahrscheinlichkeit sehr leicht folgendermaßen bestimmen: Wir lesen für $z = 1.58$ in Tab. II (S. 258) ab, daß zwischen μ und z (Flächenstück A) 44.29% aller Fälle liegen bzw. im größeren Teil (Flächenstück B) 94.29% und im kleineren Teil (Flächenstück C) 5.71%; siehe dazu auch die Erläuterungsskizzen auf S. 254. Sinngemäß auf unser Rechenbeispiel angewendet bedeutet dies, daß in 5.7% aller Fälle eine Trefferquote von 8 oder mehr Punkten (9 oder 10 Punkten) per Zufall erwartet werden kann, denn: Zwischen $z = -3$ und $z = 0$ (Mittelwert) liegen 50% aller Fälle in der Normalverteilung, zwischen $z = 0$ und $z = 1.58$ weitere 44.29%, zusammen also 94.29%. Die Wahrscheinlichkeit, per Zufall 8 (eigentlich 7.5) oder mehr Treffer zu erzielen, beträgt demnach 100% − 94.29% = 5.71%. Nach der genaueren, aber umständlicheren Wahrscheinlichkeitsberechnung via Binomialverteilung erhielten wir einen Wert von 5.5% (s. S. 175); die Differenzen sind praktisch unerheblich. Bei größeren Stichproben, etwa $N > 30$ bzw. $N > 100$, bedienen wir uns deshalb zweckmäßig der ökonomischeren z-Werte (Standardnormalverteilungsfunktion). Zur Sicherung der Anwendungstechnik diene noch folgendes *Beispiel.*

Eine Stichprobe von N = 100 Schülern sei mit einem Schulleistungstest untersucht worden. Die Statistiken M = 32 RW und s = 10 RW sind bekannt. Das hier interessierende Individuum habe eine Testleistung von RW = 22 erzielt. Wie wahrscheinlich ist ein solches Ergebnis?

Gemäß der Beziehung $z = \frac{X - M}{s}$ (im Beispielfall handelt es sich ja um empirische Werte einer Stichprobe oder Statistiken, weshalb hier M und s als Symbole für Mittelwert und Standardabweichung stehen) entspricht RW = 22 einem

$$z = \frac{22 - 32}{10} = -1 \quad \text{(s. Abb. 26).}$$

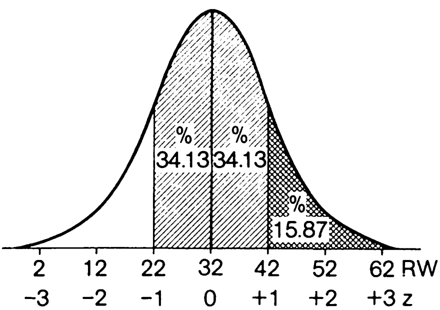


Abb. 26: Veranschaulichungsskizze zum Schulleistungstest-Beispiel

Wir sehen: Zwischen M ± 1 z scharen sich 2mal 34.13%, zusammen also 68.26% aller Fälle. Zwischen z = +1 und z = +3 liegen noch einmal 15.87%, ebenso zwischen z = −1 und z = −3. Hier interessiert nur der „untere“ Kurvenabschnitt, da dem RW = 22 der Standardwert z = −1 entspricht, der ja unterhalb vom Mittelwert (Gruppendurchschnitt) liegt. In rund 16 von 100 Fällen oder einer Wahrscheinlichkeit von 16% (genau: p = 15.87%) ist demnach mit einer Leistung von RW = 22 im Test zu rechnen, d.h., 84% der Schüler erzielten bessere Leistungen (RW 22) als der herausgegriffene Fall. Dieselbe Beziehung können wir aus Tab. II (s. S. 257) ablesen: für das Flächenstück C = 15.87% bzw. für B = 84.13%. Bei der Ermittlung der Flächenprozente sind die in Tab. II angegebenen Werte der Spalten A, B und C jeweils mit der Konstanten 100 zu multiplizieren.

Legt man an die Normalkurve Tangenten an, so können diese entweder innerhalb oder außerhalb der Kurve verlaufen. An einer bestimmten Stelle, dem sog. *Wendepunkt,* schneidet eine Tangente die Normalkurve. Diese Stelle liegt genau bei z = ± 1, weshalb die z-Werte gewissermaßen natürliche Werte darstellen. Dadurch wird die konventionelle Festlegung der z-Skala in besonderer Weise unterstrichen.

Übungsbeispiele (für die Handhabung der z-Tabelle):

Nr. 5/4a

Wieviel Prozent der Fälle liegen – unter der Voraussetzung der Normalverteilung – zwischen μ und z = 2?

Nr. 5/4b
Wieviel Prozent der Fälle unter der Normalkurve verteilen sich auf den Bereich zwischen $z = -2$ und $z = +2$?

Nr. 5/5a
Wieviel Prozent der Fälle in der Normalverteilung liegen über dem Punkt $z = +1.5$?

Nr. 5/5b
Welcher z-Wert markiert jene Bereichsgrenzen, innerhalb und außerhalb deren gleiche Auftretenshäufigkeit ($p = 50\%$) besteht?

Nr. 5/5c
Bei welchem z-Wert liegen in einer Normalverteilung die oberen 80% (die oberen 20%)?

Nr. 5/6a
Es seien 35 Viertklaßkinder mit einem Intelligenztest untersucht worden. Dabei wurden $M = 12$ RP und $s = 4$ RP ermittelt. Unter der Voraussetzung der Normalverteilung des gemessenen Merkmals (Intelligenz) sei nach der Wahrscheinlichkeit dafür, daß ein Schüler nicht weniger als 8 und nicht mehr als 16 Rohpunkte (RP) erzielt, gefragt. M.a.W.: Wie groß ist p, daß im Test irgendein Wert zwischen 7.5 und 16.5 RP erzielt wird? Ggf. Veranschaulichungsskizze anfertigen!

Nr. 5/6b
Wie groß ist die Wahrscheinlichkeit, daß in der Intelligenzuntersuchung ein zufällig herausgegriffener Proband eine Leistung von mehr als 18 RP erzielt?

Nr. 5/6c
Innerhalb welcher Grenzen liegen die mittleren 75% der Fälle bei einer normal verteilten Untersuchungsvariablen, wenn $M = 16$ RP und $s = 4$ RP sind? Die Grenzen sollen sowohl in z-Einheiten als auch in Rohwerten ausgedrückt werden.

5.1.3. Poisson-Verteilung (Modell für seltene Ereignisse)

Handelt es sich um relativ seltene Ereignisse (z. B. Kriegsausbrüche, Verkehrs- oder Arbeitsunfälle, Sterbefälle im mittleren Alter, Drillingsgeburten usw. oder Schulschwänzer, Prüfungsversager, Lehrerstreik u.ä.), dann ist auch bei sehr großen Stichproben (N 10.000 oder N 100.000) keine Normalverteilung, sondern eine auffällig *unsymmetrische* Verteilungsform zu erwarten. Poisson berechnete hierfür folgende Formel:

$$f(X) = \frac{\mu^x}{e^\mu \cdot X!} = \frac{\mu^x}{X!} \cdot e^{-\mu},$$

wobei wiederum $e = 2.71828$ (Basis der natürlichen Logarithmen) und

$$e^{-\mu} = e^{-2} = 0.13534 \quad \text{ist.}$$

Die sog. Poisson-Verteilung, deren Varianz dem Mittelwert μ entspricht, ist demnach allein durch den Mittelwert gekennzeichnet. Die Benutzung nachstehend aufgeführter Tabellenwerte erspart zeitraubende Berechnungen, indem die Auftretenshäufigkeiten qua Zufallswahrscheinlichkeiten in Spalte f (X) für seltene Ereignisse unter Berücksichtigung der interessierenden Häufigkeit X einfach abgelesen werden.

X	f(X)
0	0.13534
1	0.27068
2	0.27068
3	0.18045
4	0.09027
5	0.03609
6	0.01203
7	0.00343
.	....
.	....

Beispiel:

In einer Millionenstadt sterben im Jahresdurchschnitt täglich 2 Personen durch Verkehrsunfall (M = 2 Verkehrstote). Wie wahrscheinlich ist es nun, daß an einem beliebigen Tag 4 oder mehr Verkehrsunfalltote registriert werden?
Die *Gesamt*wahrscheinlichkeit dafür, daß an einem einzigen Tag 0 oder 1 oder 2 oder 3 Verkehrsopfer zu beklagen sind, ist gleich der Summe der in obiger Tabelle aufgeführten *Einzel*wahrscheinlichkeiten, nämlich p = 0.13534 plus 0.27068 plus 0.27068 plus 0.18045 = 0.85715, d.h. 85.715%. Die Wahrscheinlichkeit für 4 oder mehr Unfalltote an einem einzigen Tag ist nunmehr gleich der Gegenwahrscheinlichkeit 1 − 0.85715 = 0.14285, d.h. 14.285%. Die Wahrscheinlichkeit für genau 4 Unfalltote an einem einzigen Tag wäre p = 0.09027, also p ~ 9%.

5.2. Spezielle Probleme der Stichprobenstatistik

Die analytische oder Stichprobenstatistik ist die Lehre von den statistischen Schlüssen oder sog. Inferenzstatistik. Mit ihrer Hilfe lassen sich Stichprobenergebnisse verallgemeinern. In der empirischen Forschung ist man fast immer auf Untersuchungen anhand von Stichproben angewiesen, da der Untersuchung von Grundgesamtheiten (Populationen oder Kollektiven) zeitlich-ökonomische Gründe entgegenstehen. Die aufgrund einer Stichprobenerhebung gewonnenen Ergebnisse (z.B. %-Zahlen, Mittelwerte, Streuungsgrößen, Korrelationskoeffizienten u.ä.), die man auch *Statistiken* nennt, können nicht ohne weiteres verallgemeinert, d.h. als gültige Werte für die Population betrachtet werden. Hierzu bedarf es der Anwendung bestimmter Prüfmethoden, die uns die Inferenzstatistik zur Verfügung stellt. Erst dann kann entschieden werden, ob die Statistiken generalisiert werden dürfen, also für die Population gültig sind. Statistische Kenndaten (%-Quoten, Mittelwerte usw.), die sich auf die Population beziehen, nennt man in der Fachterminologie *Parameter*. Diese sind allerdings in den seltensten Fällen bekannt, weshalb man hier auf Parameterschätzungen angewiesen ist. Mit den verschiedenen Schätzmethoden befassen sich nun die folgenden Ausführungen.

5.2.1. Parameterschätzung

5.2.1.1. Stichprobenverteilung und Standardfehler des Mittelwertes

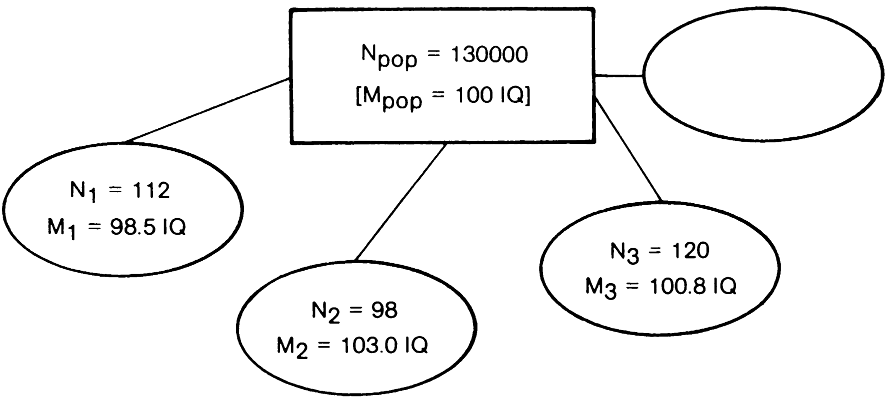
Zunächst gehen wir wieder von einem einfachen Beispiel aus. Uns interessiere die Frage, wie intelligent die Grundschüler der 4. Klasse von Baden-Württemberg sind. Somit sind alle Viertklaßkinder Baden-Württembergs (N = 130.000) als Untersuchungs*population* definiert. Daraus sei eine Stichprobe von N = 112 entnommen, wobei es sich um eine Zufallsstichprobe handeln soll. Also:

$$N_{pop} = 130000$$
$$N_1 \;\; = 112$$

Die Intelligenztestauswertung ergibt einen Stichprobenmittelwert von 98.5 IQ, also:

$$M_1 = 98.5\ \mathrm{IQ}$$

Uns interessiert freilich nicht dieser Stichprobenmittelwert, sondern der – unbekannte – Parameter, d.h. der Populationsmittelwert. Wie kommt man nun von der Statistik auf den Parameter?
Folgendes Gedankenexperiment möge das Vorgehen hierbei einsichtig machen. Wir entnehmen nicht nur *eine* Stichprobe, sondern (nachdem die untersuchten 112 Fälle in die Population zurückgegeben worden sind) eine zweite Stichprobe mit dem Ergebnis $M_2 = 103.0$ IQ, eine dritte Stichprobe mit dem Ergebnis $M_3 = 100.8$ IQ usw.



Bei genügend Stichproben darf man annehmen, daß sich die einzelnen Stichprobenmittelwerte im Sinne der Gaußschen Glockenkurve um den „wahren Mittelwert", d.h. den Populationsmittelwert (M_{pop} oder $\bar{X}$) scharen. Die *Stichprobenverteilung* der Mittelwerte sieht dann etwa folgendermaßen aus (s. Abb. 27: breit durchgezogene Kurve).
Bei einer ausreichenden Anzahl von Stichproben wäre der Mittelwert der Stichprobenverteilung identisch mit dem Mittelwert der Population. Unser Gedankenexperiment ist freilich kaum in die Praxis zu übertragen. Ein solches Vorgehen wäre viel zu aufwendig und zeitraubend. In der Regel wird man sich mit einer einzigen Stichprobe begnügen müssen. Läge ihr Mittelwert in der Mitte der Gauß-

schen Kurve, so wäre er gleich dem Populationsmittelwert ($\bar{X}$), läge er rechts von der Mitte, so wäre $\bar{X}$ kleiner, läge er links von der Mitte, so wäre $\bar{X}$ größer. Dabei kann die Stichprobenverteilung verschiedene Formen annehmen (s. Abb. 27).

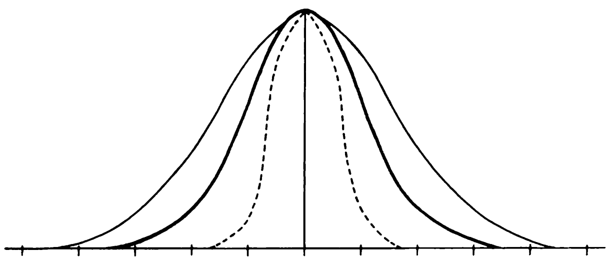


Abb. 27: Stichprobenverteilung der Mittelwerte (verschiedene Formen)

Für die Parameterschätzung wäre die engere Kurve (gestrichelte Kurvenlinie) besser geeignet als die breitere (durchgezogene dünne Kurvenlinie), weil dann der gefundene Stichprobenmittelwert näher am „wahren Mittelwert" läge. Anders ausgedrückt: Der Schluß auf den wahren Wert $\bar{X}$ gelingt um so leichter, je kleiner die Streuung der Stichprobenverteilung ist. Diese Streuung, als Zufallsschwankung um den wahren Wert M_{pop} aufgefaßt, müßte man kennen. Bekanntlich wird die Streuung mit der sog. Standardabweichung $\sigma = s$ ausgedrückt, wobei folgende Formel gilt:

$$s = \sqrt{\frac{\sum (X - M)^2}{N}}$$

Hierbei sei

X = Mittelwert der Stichprobe,
M = Mittelwert aller Stichproben oder „wahrer Mittelwert",
N = Anzahl der Stichproben.

Die nach obiger Formel berechnete Standardabweichung der Stichprobenverteilung nennt man nun „*Standardfehler* des Mittelwertes", abgekürzt σ_M. Im Englischen lautet die Bezeichnung: „*Standard Error (SE) of the Mean*". Warum spricht man hier von „Fehler" und nicht von Abweichung?
Wäre die Verteilung durch eine Verteilung individueller Werte repräsentiert, könnte man für jeden Schüler seine Abweichung vom (Gruppen-)Mittelwert $(X - M)$ berechnen. Diese Abweichung ist natürlich gegeben und stellt keinen Fehler dar. Wenn es sich jedoch um Differenzen zwischen Mittelwerten – Stichproben- und Populationsmittelwerten – handelt, handelt es sich bei der Abweichung um Fehler, da ja alle Stichprobenmittelwerte als Schätzung des wahren Mittelwertes aufgefaßt werden und entsprechende Abweichungen tatsächlich Fehler darstellen.
Nach dieser Umbenennung können wir nunmehr formulieren, daß die Möglichkeit der Parameterschätzung M_{pop} um so günstiger ist, je kleiner der Standardfehler σ_M ist. Wäre er klein, so dürfte der Schluß auf die Population vollzogen werden, wäre er groß, müßte man darauf verzichten. Der Standardfehler der Stichprobenverteilung, d.h. der Verteilung von Stichprobenmittelwerten, läßt sich am besten

mit Hilfe folgender Formel schätzen (zur Ableitung der Schätzformel vgl. z.B. Fröhlich & Becker 1971, S. 172ff.):

$$\sigma'_M = s_{M\,gesch.} = \frac{s}{\sqrt{N-1}}$$

Um den Standardfehler $s_{M\,gesch.}$ zu ermitteln, muß man also a) die Standardabweichung der individuellen Werte der Stichprobe berechnen und b) das Ergebnis durch $\sqrt{N-1}$ dividieren. Was kann man nun mit Hilfe des so geschätzten Standardfehlers über den wahren Mittelwert (M_{pop}) aussagen?
Wo der Mittelwert der Population genau liegt, kann nicht angegeben werden; es läßt sich aber der Bereich ermitteln, in dem er mit einer bestimmten Wahrscheinlichkeit, z.B. 95 : 100, liegen muß, sofern der Standardfehler des arithmetischen Mittels bekannt ist. Dieser Bereich wird Vertrauensbereich oder Vertrauensintervall des (arithmetischen) Mittelwertes genannt.

5.2.1.2. Vertrauensintervall des Mittelwertes

Die *Vertrauensgrenzen* („*Confidential Limits*" oder abgekürzt: CL) bestimmen das sog. *Vertrauensintervall* oder *Konfidenzintervall,* in der deutschsprachigen Literatur auch *Vertrauensbereich* genannt. Das Vertrauensintervall des arithmetischen Mittels kann nun für beliebige Wahrscheinlichkeiten mit Hilfe der z-Verteilung nach folgender Beziehung ermittelt werden:

$$M_{pop} = M_M = M \pm z \cdot \sigma'_M$$

Dabei bedeuten:
M_{pop} = der gesuchte Populationsmittelwert oder Mittelwert aller (Stichproben-) Mittelwerte = M_M;
M = Mittelwert einer Stichprobe, von wo aus das Vertrauensintervall zu berechnen ist;
σ'_M = Standardfehler der Stichprobenverteilung (geschätzt);
z = der mit dem gewünschten Sicherheits- oder Signifikanzniveau korrespondierende z-Wert (vgl. Tab. II im Anhang).

Die einzelnen Schritte zur Bestimmung des Vertrauensintervalls des arithmetischen Mittelwertes sollen wiederum in Anlehnung an das oben beschriebene Beispiel der Intelligenzuntersuchung in der 4. Grundschulklasse erläutert werden. Die Untersuchungsfrage lautet ja jetzt: Innerhalb welcher Grenzen liegt der „wahre Mittelwert", d.h. der IQ-Mittelwert der definierten Population? Die Schätzung des Standardfehlers habe einen Wert von $\sigma'_M = 1.2$ IQ ergeben. Die gewünschte Sicherheit – genauer: Wahrscheinlichkeit für die zu ermittelnden Grenzen betrage 95%, d.h., das Vertrauensintervall darf allenfalls mit einem Fehlerrisiko von 5% behaftet sein. Folgende Schritte sind zur Berechnung des Vertrauensintervalls nunmehr notwendig (siehe auch Erläuterungsskizze in Abb. 28):
1. Zunächst wird der z-Wert ermittelt, über und unter dem jeweils die Hälfte von 95% aller Fälle in der Standardnormalverteilung liegt. Das gewünschte Flächenstück wird in Tab. II (Spalte A) aufgesucht und der zugehörige z-Wert abgelesen (s. S. 259). Der gesuchte z-Wert lautet 1.96. Abb. 28 veranschaulicht noch einmal die Verteilungsverhältnisse: Zwischen – 1.96 z und 0.0 z (μ) liegen 47.5% der Fälle, ebenso liegen zwischen 0.0 z und + 1.96 z 47.5% der Fälle. Also liegen zwischen – 1.96 z und + 1.96 z insgesamt 95% aller Fälle.

2. Die Berechnung der Abweichung vom Mittelwert für $z = 1.96$ in IQ-Punkten ausgedrückt lautet:

$$z = \frac{X - M}{\sigma} = \frac{x}{\sigma} \quad \text{oder}$$

$$x = z \cdot \sigma$$
$$x = 1.96 \cdot 1.2$$
$$= 2.352$$

3. Daraus folgt: Mit einer Wahrscheinlichkeit von 95 : 100 hat der Populationsmittelwert keine größere Abweichung vom erhaltenen Stichprobenmittelwert von $M_1 = 98.5$ IQ als ± 2.35 IQ-Punkte. Demnach bestimmen sich die Vertrauensgrenzen aus 98.5 IQ $\pm$ 2.35 IQ, d.h., sie liegen bei 96.15 IQ (untere Grenze) und 100.85 IQ (obere Grenze). Mit anderen Worten: Mit einer Wahrscheinlichkeit von 95% liegt der „wahre Mittelwert" der Population zwischen rd. 96 IQ und 101 IQ vs. mit nur 5% Wahrscheinlichkeit außerhalb dieser Grenzen.

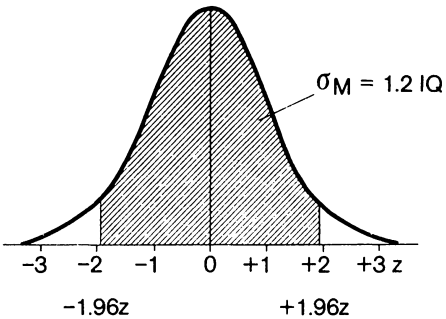


Abb. 28: Erläuterungsskizze zu nebenstehendem Rechenbeispiel für die Ermittlung des Vertrauensbereichs des arithmetischen Mittels

Sofern uns das Fehlerrisiko von 5% noch zu hoch erscheint, können wir die entsprechenden Vertrauensbereiche auf dem (nur noch selten verwendeten) 2%-Niveau oder (dem häufiger verwendeten) 1%-Niveau ermitteln. Laut Tab. II liegen 98% aller Fälle in der Standardnormalverteilung zwischen ± 2.33 z beziehungsweise 99% aller Fälle zwischen ± 2.58 z. Die Berechnung der Abweichung in IQ-Punkten ergibt diesmal $x = z \cdot \sigma$ oder $x = 2.3 \cdot 1.2 = 2.8$ und $x = 2.58 \cdot 1.2 = 3.1$. Entsprechend lauten die Vertrauensbereiche des Mittelwertes mit 98% Sicherheit (oder 2% Fehlerwahrscheinlichkeit) 98.5 IQ $\pm$ 2.8 IQ = 95.7 IQ bis 101.3 IQ und 98.5 IQ $\pm$ 3.1 IQ = 95.4 IQ bis 101.6 IQ.
Es wird ersichtlich, daß mit der Verringerung des Fehlerrisikos der Vertrauensbereich anwächst. Mit der größeren Sicherheit bzw. Wahrscheinlichkeit geht ein gewisser Genauigkeitsverlust einher. Es ist also nicht immer sinnvoll, das kleinstmögliche Fehlerrisiko zu wählen; vielmehr wird man sich je nach der konkreten Problemstellung einmal lieber auf das 5%-Niveau und ein anderes Mal lieber auf das 1%-Niveau festlegen. (Dazu ausführlicher: Kap. 5.2.2.1 unten.)
Hier sei angemerkt, daß legitimerweise die z-Tabelle nur bei größeren Stichproben ($N > 100$) zur Bestimmung der Vertrauensgrenzen herangezogen werden darf. Bei kleineren Stichproben sollte man hierfür die t-Tabelle verwenden (vgl. Tab. III im

Anhang). Außerdem ist zu beachten, daß die t-Werte von der Stichprobengröße abhängig sind, welche durch Berücksichtigung der Freiheitsgrade ($df = N - 1$) in die Berechnung eingeht. Ferner arbeitet die t-Tabelle mit den „Restwahrscheinlichkeiten" und nicht mit der „inneren" Wahrscheinlichkeit wie die z-Tabelle (vgl. Kap. 5.2.2.1 und 5.2.2.2).

5.2.1.3. Vertrauensintervalle anderer Statistiken

Analog zum Standardfehler des arithmetischen Mittels kann der *Standardfehler* für das Vertrauensintervall *des Median* geschätzt werden. Fröhlich & Becker (1971, S. 191 f.) geben hierfür folgende Formel an:

$$\sigma_{Med} = 1.253 \cdot \sigma'_M$$

Nach Süllwold (vgl. Heller 1973, S. 70) kann der Standardfehler des Median auch nach folgender Formel geschätzt werden:

$$\sigma_{Med} = \frac{1.858 \cdot Q}{N}$$

Dabei sind die Zahlen 1.253 und 1.858 Konstanten, N bezieht sich auf die Anzahl der Fälle, und Q ist die Quartilabweichung, also: $Q = Q_3 - Q_1/2$.

Für die Bestimmung des *Standardfehlers von Prozentzahlen* (%) kann nach Süllwold folgende Beziehung verwendet werden:

$$\sigma_{\%} = \sqrt{\frac{P \cdot Q}{N}}$$

Dabei bedeuten:

$\sigma_{\%}$ = Standardfehler der %-Zahl,
P = empirisch ermittelte Prozentzahl,
Q = 100 – P und wiederum
N = Stichprobengröße.

Für die Schätzung des *Standardfehlers der Standardabweichung* geben Fröhlich & Becker (a.a.O.) folgende Definitionsformel an:

$$\sigma_\sigma = 0.71 \cdot \sigma'_M$$

Und für den *Standardfehler der Quartilabweichung* oder (nach Fröhlich) des sog. *mittleren Quartils* kann zur Berechnung folgende Schätzformel benutzt werden:

$$\sigma_Q = 0.786 \cdot \sigma'_M$$

5.2.2. Hypothesenprüfung

Wir kehren wieder zurück zu unseren Ausgangsproblemen, nämlich der Frage nach der Intelligenz aller Viertklaßkinder Baden-Württembergs. Bisher versuchten wir via Parameterschätzung eine Antwort auf diese Frage zu finden. Ein anderer Ansatz bestünde darin, daß man von einer Hypothese ausgeht, nämlich der, daß sich die untersuchten Schüler ($N_1 = 112$) nicht oder nur zufällig von allen anderen unterscheiden. Der mittlere IQ einer unausgelesenen Stichprobe bzw. Population liegt bei IQ = 100. Die Prüfung dieser – oder einer anderen – Hypothese ist in

der experimentellen Praxis oft wichtiger und interessanter als die Ermittlung von Vertrauensgrenzen diverser Parameter.
Im angenommenen Beispielfall lautet die Hypothese:

$$M_{pop} = 100 \text{ IQ}$$

Wenn die Hypothese stimmt, liegt IQ = 100 in der Mitte der Verteilung. Folgende Statistiken sind bekannt:

$$M_1 = 98.5 \text{ IQ}$$

$$\sigma'_M = 1.2$$

Mit Hilfe dieser Daten können wir nun die aufgestellte Hypothese ($M_{pop} = 100$ IQ) überprüfen. Zunächst bilden wir die Differenz zwischen dem hypothetischen und dem empirischen Mittelwert; diese beträgt $100 - 98.5 = 1.5$ IQ und wird auch *Stichprobenfehler* genannt. Der Stichprobenfehler ist durch Zufall entstanden, d.h. dadurch, daß in die kleine Stichprobe relativ mehr unterdurchschnittlich begabte Schüler geraten sind. Die entscheidende Frage lautet nun: Wie wahrscheinlich ist es, daß ein solcher Fehler durch Zufall entstanden ist? Vom Ergebnis der Wahrscheinlichkeitsberechnung wird es abhängen, ob die zu überprüfende Hypothese akzeptiert oder verworfen wird.
Zunächst müssen wir den in IQ-Einheiten ausgedrückten Stichprobenfehler in den Standardwert z umwandeln. Dies geschieht mit Hilfe der Formel:

$$z = \frac{x}{\sigma}$$

Bezogen auf unser Rechenbeispiel ist $z = \frac{1.5}{1.2} = 1.25$, d.h., unser Stichprobenmittelwert weicht vom wahren Populationsmittelwert ($M_{pop} = 100$ IQ) um 1.25 z-Werte ab. Wie wahrscheinlich ist es nun, daß – unter der Voraussetzung normal verteilter Meßwerte – eine solche Abweichung auftritt? Die Frage läßt sich wiederum anhand der z-Tabelle beantworten.

5.2.2.1. Einseitiger und zweiseitiger Test

Bei der Bestimmung der Wahrscheinlichkeit gibt es grundsätzlich drei Möglichkeiten: nur die Wahrscheinlichkeit der positiven Abweichung, nur die Wahrscheinlichkeit der negativen Abweichung, schließlich die Wahrscheinlichkeit der positiven *und* negativen Abweichung.
Dem z-Wert 1.25 entsprechen lt. Tab. II im Anhang 39.44% aller Fälle. Somit liegen zwischen $z = -1.25$ und $z = +1.25$ zweimal $39.44\% = 78.88\%$ aller Fälle. Entsprechend liegen außerhalb von ± 1.25z 21.12% (100 – 78.88%) der Fälle. Der hier ermittelte Stichprobenfehler ist also mit einer Wahrscheinlichkeit von rd. 21% durch Zufall entstanden. Dies war der sog. *zweiseitige* Test, der beide Richtungen der Abweichung vom Mittelwert berücksichtigt.
Hingegen wird beim sog. *einseitigen* Test jeweils nur eine Abweichungsrichtung – nach der positiven oder negativen Seite hin – beachtet. In unserem Beispiel würde sich folgende Wahrscheinlichkeit beim einseitigen Test ergeben:

$$21.12\% : 2 = 10.56\% \quad \text{oder} \quad 50\% - 39.44 = = 10.56\%$$

Der einseitige Test spielt eine wichtige Rolle bei der Interpretation von Mittelwertsdifferenzen, z. B., wenn bereits Vorinformationen über die Richtung der Abweichung bekannt sind. So genügt etwa in der Frage der Schuleignungsermittlung sehr oft die einseitige Problemstellung: Während zur Klärung der Lernbehinderung via IQ-Tests in der Regel allein der „untere" Bereich kritisch wird, ist bei der Erfassung der Gymnasialeignung praktisch nur der „obere" Bereich interessant. Mit dem einseitigen Test sind, wie unser Rechenbeispiel zeigte, Ergebnisse leichter signifikant zu machen als mit dem zweiseitigen Test (vgl. S. 66 f.).

5.2.2.2. Signifikanzniveaus und Fehlerrisiken

Schließlich wäre noch nach den Kriterien, die über die Annahme oder Ablehnung der Prüfhypothese (H_1 bzw. H_0) entscheiden, zu fragen. Diese Kriterien sind durch internationale Konvention heute festgelegt auf dem *5%-Niveau* (für diesen Fall ist $z = \pm 1.96$) und auf dem *1%-Niveau* (für diesen Fall ist $z = \pm 2.58$).
Im Hinblick auf unser Demonstrationsbeispiel gilt nun: Wenn der z-Wert des Stichprobenfehlers 1.96 oder 2.58 erreicht bzw. übersteigt, so ist die untersuchte Hypothese ($M_{pop} = 100$ IQ) abzulehnen. In unserem Fall liegt der z-Wert des Stichprobenfehlers darunter, d. h., die Hypothese wird akzeptiert. Die Wahrscheinlichkeit dafür, daß der Populationsmittelwert im Akzeptationsbereich liegt, ist größer als 5 : 100 (vgl. Abb. 29).
Es kann freilich vorkommen, daß die Hypothese stimmt, obwohl der Stichprobenmittelwert in den Rejektionsbereich fällt. Dies ist denkbar für den Fall, daß man per Zufall eine sehr extreme Stichprobe gezogen hat. Dann wird die Hypothese verworfen, obwohl sie zutrifft – freilich nur mit einer Wahrscheinlichkeit von kleiner als 5% bzw. 1%.

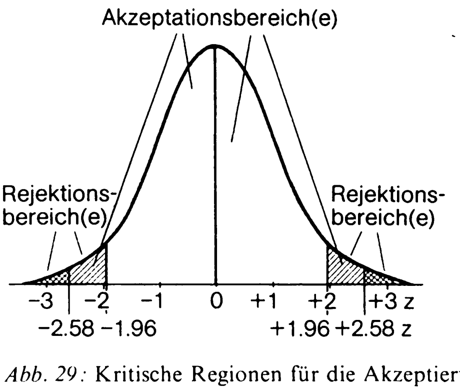


Abb. 29: Kritische Regionen für die Akzeptierung vs. Zurückweisung von H_0

Bei der Entscheidung von Hypothesen geht man also grundsätzlich ein gewisses *Fehlerrisiko* ein, d. h., man unterliegt einmal dem Fehler vom Typ α oder *Risiko erster Art:* Die Hypothese wird verworfen, obwohl sie stimmt. Die Wahrscheinlichkeit dafür beträgt aber nur 5 : 100 oder 1 : 100; und zum andern setzt man sich dem Fehler vom Typ β oder *Risiko zweiter Art* aus: Die Hypothese wird akzeptiert, obwohl sie falsch ist. Die Wahrscheinlichkeit dafür beträgt wiederum nur 5 : 100 oder 1 : 100. Meistens wird die Größe des Fehlers vom Typ α gewählt.

Die Frage, ob das 5%- oder 1%-Niveau im Hinblick auf die Hypothesensicherung günstiger sei, kann nicht generell beantwortet werden. Die allgemeine Meinung geht dahin, daß das Ergebnis um so gesicherter sei, je niedriger das gewählte Sicherheitsniveau ist. Dazu muß bemerkt werden: Je rigoroser man vorgeht, desto weniger wahrscheinlich ist es, daß die richtige Hypothese verworfen wird. Aber um so höher ist auch die Wahrscheinlichkeit, eine falsche Hypothese zu akzeptieren. In der Regel wird aber nur *eine* Hypothese favorisiert, in unserem Beispiel die Hypothese $M_{pop} = 100$ IQ.

Die Schritte zur *Hypothesenprüfung* lassen sich folgendermaßen zusammenfassen:

1. Eine Hypothese wird aufgestellt (vgl. auch Kap. 2.2.4.5 oben).
2. Der Stichprobenfehler wird errechnet.
3. Der Rohwert wird in den Standardwert z umgewandelt nach der Beziehung

   $$z = \frac{x}{\sigma_M},$$

   wobei sich σ_M nach der Formel $\sigma_M = \frac{s}{\sqrt{N-1}}$ bestimmen läßt.
4. Zur Entscheidung über die Akzeptierung oder Ablehnung der Hypothese wird der gefundene z-Wert anhand der kritischen Regionen (± 1.96 bzw. ± 2.58) überprüft, d.h. der empirische z-Wert mit dem theoretischen z-Wert verglichen.
5. Abschließend erfolgt die Wahrscheinlichkeitsbestimmung mit Hilfe der z-Tabelle (Tab. II im Anhang).

Analog wird mit Prüfgrößen anderer Art, z.B. t-Test oder F-Test, dem Chi^2-Test usw., verfahren. Die entsprechenden Tabellen sind ebenfalls im Anhang beigefügt. Für Demonstrationszwecke wurde bisher ausschließlich der z-Test zur Überprüfung der Differenzen zwischen M_{pop} und $M_{Stichpr.}$ erörtert. In gleicher oder ähnlicher Weise lassen sich – auch unter Verwendung anderer Prüftests – Differenzen zwischen zwei bzw. mehreren Stichproben, z.B. zwischen Experimental- und Kontrollgruppe(n), statistisch überprüfen, d.h. auf Zufälligkeit vs. Überzufälligkeit hin kontrollieren. Gewöhnlich prüft man in solchen Fällen die sog. *Nullhypothese* (H_0): Man setzt die Hypothese an, daß die beobachteten Differenzen unbedeutend, d.h. gleich „Null" seien. Erst wenn die Hypothese H_0 zurückgewiesen ist, kann die *Alternativhypothese* (H_1) statistisch als gesichert gelten, also akzeptiert werden. Entsprechende Ergebnisse sind dann „signifikant" oder „sehr signifikant", d.h., die ermittelten Differenzen weichen in diesem Falle signifikant (5%-Niveau) oder sehr signifikant (1%-Niveau) von Null ab. Siehe ergänzend die früheren Ausführungen auf S. 65ff.

In den folgenden Kapiteln werden nun die im Hinblick auf die empirische Forschungspraxis wichtigsten Verfahren der Inferenzstatistik systematisch behandelt. Hierbei wird jede Methode anhand eines konkreten Untersuchungsbeispiels ausführlich dargestellt; darüber hinaus sollen wiederum zahlreiche Übungsbeispiele die Anwendungstechnik der statistischen Tests (Verfahren der Inferenzstatistik) beim Leser bzw. Benutzer dieses Lehrbuches sichern helfen.

5.2.2.3. *Parametrische und nonparametrische Signifikanztests*

Die klassischen Verfahren der Inferenz- oder Stichprobenstatistik setzen voraus, daß die Parameterverteilung bekannt ist, d.h. die untersuchten Merkmalsdaten einer normal verteilten Population entstammen. Einschlägig sind hier z.B. der z-Test, der t-Test, der F-Test u.ä. Prüfverfahren, die hinsichtlich ihrer Anwendbarkeit Normalverteilung voraussetzen, bezeichnet man als *verteilungsgebundene* oder *verteilungsabhängige* oder *parametrische* Signifikanztests.
In der pädagogischen bzw. sozialwissenschaftlichen Forschungspraxis stehen aber häufig keine verläßlichen Informationen über die Verteilung der Meßdaten zur Verfügung. In solchen Fällen ist man auf *verteilungsfreie* oder *verteilungsunabhängige* oder *nonparametrische* Prüftests angewiesen, z.B. den Chi^2-Test, U-Test, H-Test, Wilcoxon-Test u.a. Diese Prüfverfahren setzen *keine* normal verteilten Daten für ihre Anwendung voraus, weshalb ihnen breite Aufmerksamkeit in diesem Buch gewidmet wurde. Viele der aufgeführten nonparametrischen Verfahren zur Hypothesenprüfung sind bereits so weit vervollkommnet, daß bei ihrer Anwendung – im Vergleich zum Einsatz parametrischer Verfahren – nur geringfügige Informationsverluste in Kauf genommen werden müssen. Im Zweifelsfalle, d.h., wenn keine gesicherten Informationen über die Normalverteilung der Untersuchungs-Variablen vorliegen, sollte man deshalb immer nonparametrische Signifikanztests bevorzugen. Diese Kautel gilt logischerweise auch dann, wenn die Meßdaten nicht mindestens Intervallskalenniveau erreichen, also auf Rang- oder gar nur auf Nominalskalenniveau erhoben wurden (vgl. Kap. 2.2.6.2).

5.3. Parametrische Verfahren zur Hypothesenprüfung

Allen im folgenden aufgeführten Verfahren zur Überprüfung der statistischen Bedeutsamkeit von Gruppenunterschieden sind zwei wesentliche Charakteristika gemeinsam:

a) Sie sind nur für Daten auf dem Intervallskalenniveau zulässig, d.h. jeder Messung muß ein Intervallskalenwert zugeordnet werden können, der sich in einer Maßeinheit ausdrücken läßt (z.B. cm, IQ, Anzahl der richtig gelösten Aufgaben...).
b) Sie gründen sich auf einen Vergleich verschiedener Schätzungen von Populations-Parametern (Mittelwert oder Streuung), wobei die Schätzungen über die entsprechenden Kennwerte von Zufallsstichproben erfolgen (daher der Name „parametrische Verfahren"). Als entscheidende Voraussetzung wird die Normalverteilung der Population angenommen, so daß als die Grundbedingung für die Repräsentativität der Stichproben ihre Normalverteilung gilt.
   Ob sich die Daten einer Stichprobe normal verteilen, läßt sich grob anhand eines Häufigkeitendiagramms überprüfen; eine exakte Überprüfung bietet der in Kap. 5.4.1.3 dargestellte Chi^2-Test für die Güte der Anpassung. Als Faustregel gilt: Je größer die Anzahl der untersuchten Fälle (Stichprobe), um so wahrscheinlicher wird die Normalverteilung, sofern die der Stichprobe zugrunde liegende Population normal verteilt ist. Im übrigen konnte empirisch nachgewiesen werden, daß parametrische Verfahren, besonders die Varianzanalyse, gegen

Abweichungen von der Normalität recht unempfindlich sind (z. B. Norton 1952). Bei groben Verletzungen der Normalitätsbedingung sollten jedoch besser keine Datentransformationen durchgeführt werden, um zu einer annähernden Normalverteilung zu gelangen, und es sollten die parametrischen Verfahren auch nicht direkt angewandt werden; in einem solchen Fall muß auf die nonparametrischen Verfahren (siehe Kap. 5.4) zurückgegriffen werden. In dem unten dargestellten Schema geben wir eine Übersicht über die hier besprochenen parametrischen Prüfverfahren.

a) *Mittelwertsvergleiche*

	unabhängige Stichproben		abhängige Stichproben	
	zwei	mehrere	zwei	mehrere
Intervall-skalen-niveau	*z-Test* (wenn der Pop.-Mittelwert bekannt ist) (s. S. 190 ff.) *t-Test* für unabhängige Stichproben (s. S. 192 ff.)	einfache *Varianzanalyse* (s. S. 198 ff.) mit: Newman-Keuls-Test (s. S. 205 ff.) und Omega-Test (s. S. 207 f.)	*t-Test* für abhängige Stichproben (s. S. 194 ff.)	

b) *Varianz-Vergleiche*

	unabhängige Stichproben		abhängige Stichproben	
	zwei	mehrere	zwei	mehrere
Intervall-skalen-niveau	F-Test (s. S. 197 ff.)	F_{max}-Test (s. S. 198 ff.)	t-Test nach Ferguson (s. S. 198)	t_{max}-Test nach Ferguson (s. S. 198 ff.)

Abb. 30: Parametrische Verfahren zur Hypothesenprüfung (Übersichtsschema)

5.3.1. Der z-Test zur Überprüfung von Mittelwertsdifferenzen zwischen einer Population und einer Stichprobe

(1) *Untersuchungsproblem*

Unterscheidet sich der Mittelwert (M) einer Zufallsstichprobe von dem Mittelwert (μ = My) einer Population? Die Mittelwerte beziehen sich auf ein kontinuierliches Merkmal (Intervallskala). Wir wollen also wissen, ob eine Zufallsstichprobe (z. B. Schüler der 4. Klasse einer Grundschule) hinsichtlich eines bestimmten kontinuierlichen Merkmals (z. B. Körpergröße) zu einer uns bekannten Grundgesamtheit (z.B. alle Schüler von 4. Schulklassen in der BRD) gehört oder ob sie sich bedeutsam (signifikant) unterscheidet.

(2) *Beispiel*

a) Versuchsplan
Ein Lehrer in Ostfriesland möchte wissen, ob die Schüler seiner vierten Klasse größer oder kleiner sind als der Durchschnitt aller Viertkläßler der BRD (die Fragestellung in diesem Beispiel ist also zweiseitig).
Population: alle Viertkläßler in der BRD
Stichprobe: Viertkläßler einer Schule in Ostfriesland
Merkmal: Körpergröße (gemessen auf einer kontinuierlichen cm-Skala)

b) Hypothesen
Nullhypothese (H_0):
Der empirische Unterschied zwischen dem Mittelwert der Population und dem der Stichprobe ist rein zufällig.
Alternativhypothese (H_1):
Der Mittelwertsunterschied ist bei einem Signifikanzniveau von .05 überzufällig.

c) Ergebnisse
μ = 149 cm (Mittelwert der Population)
M = 153 cm (Mittelwert der Stichprobe)
σ = 10 cm (Streuung der Population)
N = 36 (Größe der Stichprobe)

d) Berechnung
Zur Überprüfung des Mittelwertsunterschiedes zwischen einer Population und einer Stichprobe lassen sich die sog. z-Werte einer Normalverteilung heranziehen.
Der empirische z-Wert für den Mittelwert der Stichprobe (vgl. Kap. 3.5.2) berechnet sich nach folgender Formel:

$$z = \frac{\mu - M}{\sigma} \sqrt{N}$$

Für unser Beispiel ergibt sich ein empirischer z-Wert für den Stichprobenmittelwert von $z = -2.4$.

e) Signifikanzprüfung
Zum Vergleich wird der theoretische z-Wert in Tab. II unter Berücksichtigung des vorher festgelegten Signifikanzniveaus – zweiseitige Fragestellung – aufgesucht (bei einer Fehlerwahrscheinlichkeit von $p = .05$ ist $z = 1.96$). Wenn nun der empirische z-Wert gleich oder größer ist als der theoretische, muß H_0 abgelehnt werden, d.h., der Unterschied zwischen den Mittelwerten der Stichprobe und der Population ist signifikant. Wenn der empirische z-Wert kleiner ist als der theoretische, muß H_1 abgelehnt werden, d.h., es besteht kein signifikanter Unterschied zwischen den beiden untersuchten Mittelwerten.
In unserem Beispiel ergibt sich eine positive Differenz zwischen theoretischem und empirischem z-Wert, d.h., der Mittelwertsunterschied ist signifikant. Die Kinder der vierten Klasse einer Schule in Ostfriesland unterscheiden sich also in ihrer Größe vom Durchschnitt aller Viertkläßler in der BRD.

(3) *Anwendungsregeln*

a) μ, σ, M und N sind bekannt.
b) Die Fehlerwahrscheinlichkeit wird festgelegt.
c) Der z-Wert, der dem zugelassenen Fehlerbereich bei ein- oder zweiseitiger Fragestellung entspricht, wird nach Tab. II bestimmt.
d) Der empirische z-Wert wird nach der o.a. Formel berechnet.
e) Vergleich der beiden z-Werte:
wenn $z_{emp} < z_{theo}$: H_0 trifft zu
wenn $z_{emp} \geqq z_{theo}$: H_1 trifft zu

(4) *Voraussetzungen und Einschränkungen*

a) Die Daten der Population und der Stichprobe müssen normal verteilt sein.
b) Das untersuchte Merkmal muß stetig und auf einer Intervallskala meßbar sein.
c) Wenn σ unbekannt ist, kann man es entweder über die Formel $\sigma = \frac{s}{\sqrt{N-1}}$ schätzen (das ist jedoch nur bei einem N von mindestens 40 möglich!), oder man verwendet die Prüfstatistik t (siehe nächsten Abschnitt).

(5) *Übungsbeispiele*

Nr. 5/7
Das Alter der Population Volksschullehrer in der BRD beträgt durchschnittlich 38 Jahre bei einer Streuung von 15. Das Land Bremen berechnet für seine 167 Lehrer in Volksschulen ein Durchschnittsalter von 40 Jahren. Unterscheiden sich die Volksschullehrer aus Bremen von ihren Kollegen in der gesamten BRD signifikant auf dem 5%-Niveau hinsichtlich des Alters (einseitige Fragestellung)?
Nr. 5/8
Ein frisch versetzter Lehrer möchte feststellen, ob das Intelligenzniveau der Schüler seiner neuen Klasse über oder unter dem Bundesdurchschnitt liegt.
Er testet 20 Schüler mit dem Hamburg-Wechsler-Intelligenztest für Erwachsene ($\mu = 100$, $\sigma = 15$) und erhält einen Mittelwert von 118.

5.3.2. Der t-Test nach Student zur Überprüfung der Mittelwertsdifferenzen zwischen zwei Stichproben

5.3.2.1. Für unabhängig gewonnene Stichproben (vgl. Kap. 2.2.5) bei homogener Varianz

(1) *Untersuchungsproblem*

Unterscheidet sich der Mittelwert einer Zufallsstichprobe (M_1) von dem einer anderen (M_2), die statistisch unabhängig ist von der ersten?
Die Mittelwerte beziehen sich auf ein stetiges Merkmal (Intervallskala). Wir wollen also wissen, ob zwei unabhängige Zufallsstichproben (z. B. Mütter von Jurastudenten und Mütter von Germanistikstudenten) hinsichtlich eines bestimmten stetigen Merkmals (z. B. Intelligenzquotient) derselben Population angehören (z. B. Mütter von Studenten) oder ob den beiden Stichproben unterschiedliche Populationen zugrunde liegen, d.h. ob sie sich hinsichtlich des Merkmals signifikant unterscheiden.

(2) *Beispiel*

a) Versuchsplan

Ein psychologisches Institut will untersuchen, ob Mütter von Jurastudenten einen höheren oder niedrigeren IQ haben als Mütter von Germanistikstudenten (zweiseitige Fragestellung).

Stichprobe 1: 45 Mütter von Germanistikstudenten

Stichprobe 2: 40 Mütter von Jurastudenten (beide Stichproben werden unabhängig voneinander erhoben)

Merkmal: Intelligenzquotient

b) Hypothesen

H_0: Der Unterschied zwischen den beiden Stichprobenmittelwerten ist rein zufällig.

H_1: Der Mittelwertsunterschied ist bei einem Signifikanzniveau von .05 überzufällig.

c) Ergebnisse

$M_1 = 120$ (Mittelwert der Stichprobe 1)

$s_1 = 9$ (Streuung der Stichprobe 1)

$N_1 = 45$ (Größe der Stichprobe 1)

$M_2 = 117$ (Mittelwert der Stichprobe 2)

$s_2 = 12$ (Streuung der Stichprobe 2)

$N_2 = 40$ (Größe der Stichprobe 2)

d) Berechnung

Die Prüfstatistik ist im Falle eines Mittelwertsvergleichs zwischen zwei kleinen ($N < 40$) Stichproben die t-Statistik. Sie bezieht sich nicht wie der z-Wert auf die (Standard-)Normalverteilung, sondern auf die theoretische t-Verteilung nach Student, ebenfalls eine symmetrische Verteilung, die bei großen Stichproben in die Normalverteilung übergeht, bei kleineren Stichproben mehr oder weniger große Abweichungen von dieser zeigt. Bei kleinen Stichproben repräsentiert die t-Verteilung die Verhältnisse besser als die Gaußsche Normalverteilung.

Der empirische t-Wert berechnet sich nach der Formel

$$t = \frac{M_1 - M_2}{s_{(M_1 - M_2)}}, \text{ wobei bei unabhängigen Stichproben}$$

$$s_{(M_1 - M_2)} = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2} \quad \frac{1}{N_1} + \frac{1}{N_2}}\ ^{1}$$

Für unser Beispiel ergibt sich ein t-Wert von $t = 1.71$.

e) Signifikanzprüfung

Zum Vergleich wird der theoretische t-Wert in Tab. III unter Berücksichtigung des vorher festgelegten Signifikanzniveaus bei ein- oder zweiseitiger Fragestellung und der entsprechenden Freiheitsgrade aufgesucht.

Die Freiheitsgrade berechnen sich nach der Formel:

$$df = (N_1 - 1) + (N_2 - 1) = N_1 + N_2 - 2$$

Für unser Beispiel enthält Tab. III bei einem Signifikanzniveau von .05 und $df = 83$ einen t-Wert von $t = 2.000$.

---

1 $s_{(M_1 - M_2)} = \sigma_{diff}$, d.h. Standardabweichung der Verteilung aller M-Differenzen zwischen den beiden Stichproben oder Standardfehler der Stichprobenmittelwerte.

Wenn nun der empirische t-Wert gleich oder größer ist als der theoretische, so muß H_0 abgelehnt werden, d.h., die Mittelwertsdifferenz zwischen den beiden Stichproben ist signifikant. Wenn aber der empirische Wert kleiner ist als der theoretische, dann muß H_1 abgelehnt werden, d.h., es besteht kein überzufälliger Unterschied zwischen den beiden untersuchten Mittelwerten.
In unserem Beispiel erweist sich $t_{emp} = 1.71$ kleiner als $t_{theo} = 2.00$. Wir können demnach mit einer Sicherheit von 95% sagen, daß sich die Mütter von Jura- und Germanistikstudenten hinsichtlich ihres Intelligenzquotienten höchstens zufällig unterscheiden (fiktives Beispiel; die Daten wurden nicht empirisch gewonnen).

(3) *Anwendungsregeln*

a) M_1, s_1, N_1 und M_2, s_2, N_2 sind bekannt.
b) Das Signifikanzniveau und die Freiheitsgrade werden festgelegt.
c) Der t-Wert, der dem zugelassenen Fehler bei ein- oder zweiseitiger Fragestellung entspricht, wird nach Tab. III bestimmt.
d) Der empirische t-Wert wird nach der o.a. Formel berechnet.
e) Vergleich der beiden t-Werte:
wenn $t_{emp} < t_{theo}$: H_0 trifft zu
wenn $t_{emp} \geqq t_{theo}$: H_1 trifft zu

(4) *Voraussetzungen und Einschränkungen*

Siehe t-Test für abhängige Stichproben, Pkt. 4.

(5) *Übungsbeispiele*

Nr. 5/9
Der mittlere Intelligenzquotient von 14 Schülern eines Villenvorortes beträgt 112 mit einer Standardabweichung von 8, und der mittlere IQ von 16 Schülern einer Mietshaussiedlung derselben Stadt beträgt bei einer Standardabweichung von 10 IQ = 107. Besteht ein signifikanter Unterschied zwischen den beiden Stadtgebieten hinsichtlich der IQs von Schülern auf dem 5%-Niveau?
Nr. 5/10
Bei einem psychologischen Experiment wurde die Lernfähigkeit von neurotischen Personen verglichen mit der von nichtneurotischen Vpn. Als Maß für die Lernfähigkeit diente die Zahl der Fehler bis zu einem Kriterium. Die Gruppe der neurotischen Vpn hatte folgende Werte: 17, 18, 10, 18, 5, 7, 22, 22, 6, 11, 5, 23, 14, 25, 5, 24. Die Gruppe der nichtneurotischen Vpn hatte folgende Werte: 11, 12, 18, 10, 15, 14, 17, 10, 10, 11, 12, 14.
Welche Aussagen kann man aufgrund der Daten über die Lernfähigkeit bei neurotischen verglichen mit nichtneurotischen Vpn machen?

5.3.2.2. Für abhängige Stichproben (vgl. Kap. 2.2.5) bei homogener Varianz

(1) *Untersuchungsproblem*

Unterscheidet sich der Mittelwert einer Datenmenge von dem einer zweiten, die im statistischen Sinne von der ersten abhängig ist? Wir wollen also wissen, ob sich zwei voneinander abhängige Datenstichproben (z.B. die Intelligenzquotienten einer Schülergruppe vor und nach einem Intelligenztraining) signifikant unterscheiden, d.h. ob zwei verschiedene Populationen zugrunde liegen.

(2) *Beispiel*

Eine Lehrerin führt mit ihrer Klasse ein Intelligenzförderungstraining durch und möchte wissen, ob dieses Training tatsächlich einen positiven Effekt auf die Höhe des IQ ausübt oder nicht.

a) Versuchsplan
Die Schüler werden vor der Durchführung des Tests hinsichtlich ihrer Intelligenz getestet. Nach dem Intelligenztraining werden dieselben Schüler ein zweites Mal (mit einer Parallelform des gleichen Tests) untersucht. Die Daten sind also dadurch voneinander abhängig, daß sie an denselben Versuchspersonen gewonnen wurden.

b) Hypothesen
H_0: Der Unterschied zwischen den Mittelwerten der beiden Datenmengen ist rein zufällig.
H_1: Die beiden Mittelwerte unterscheiden sich bei einem Signifikanzniveau von .05 und einseitiger Fragestellung überzufällig.

c) Ergebnisse
$M_1 = 109.3$ (mittlerer IQ *vor* dem Intelligenztraining)
$M_2 = 116.8$ (mittlerer IQ *nach* dem Intelligenztraining)
$s_1 = 3$ (Streuung der IQ *vor* dem Training)
$s_2 = 4$ (Streuung der IQ *nach* dem Training)
$N = 20$ (Anzahl der getesteten Schüler)
$\sum (X_1 - X_2)^2 = 1887$ (Summe der quadrierten Einzeldifferenzen)

d) Berechnung
Die Berechnung des Mittelwertsunterschiedes zwischen zwei abhängigen Stichproben auf Signifikanz hin erfolgt ebenfalls mit Hilfe der Prüfstatistik t und der Formel:

$$t = \frac{M_1 - M_2}{s_{(M_1 - M_2)}}$$

Der Unterschied bei der Berechnung im Vergleich zum t-Test für unabhängige Stichproben liegt in der Bestimmung der Streuung der Mittelwertsdifferenzen: $s_{(M_1 - M_2)}$. Sie erfolgt bei *abhängigen* Stichproben nach der Formel:

$$s_{(M_1 - M_2)} = \sqrt{\frac{\sum (X_1 - X_2)^2}{N(N-1)}}$$

Für unser Beispiel ergibt sich ein t-Wert von $t = 1.51$.

e) Signifikanzprüfung
Siehe t-Test für unabhängige Stichproben.
Zu unserem Beispiel finden wir in Tab. III bei einem Signifikanzniveau von .05 und einseitiger Fragestellung für $df = 19$ einen theoretischen t-Wert von $t = 1.73$. Da $t_{emp} = 1.51$ kleiner ist als t_{theo}, müssen wir H_1 verwerfen. Die Intelligenzquotienten der Schüler unterscheiden sich vor und nach dem Intelligenztraining mit 95% Wahrscheinlichkeit nicht signifikant voneinander.

(3) *Anwendungsregeln*

Siehe t-Test für unabhängige Stichproben, Pkt. 3.

(4) *Voraussetzungen und Einschränkungen*

a) Die Populationen, aus denen die Stichproben stammen, müssen normal verteilt sein (diese Bedingung kann unter Umständen vernachlässigt werden).
b) Die Meßwerte müssen kontinuierlich auf einer Intervallskala meßbar sein.
c) Die Varianzen der beiden Stichproben müssen homogen (gleich) sein (bei gleich großen Stichproben ist diese Bedingung weniger bedeutsam als bei ungleich großen).
   Die Varianzhomogenität läßt sich mit Hilfe der Prüfstatistik F ermitteln, die sich nach R. A. Fisher linksgipflig verteilt (s. Kap. 5.3.3).
   Bei Varianzheterogenität (Ungleichheit der Varianzen) schlagen die meisten Autoren die Anwendung parameterfreier Verfahren vor (s. Kap. 5.4); zwei Näherungsmethoden in Anlehnung an den t-Test haben Edwards und Welch entwickelt.
   Die Methode von Edwards wollen wir in 5.3.2.3 darstellen.

(5) *Übungsbeispiele*

Nr. 5/11

Bei einer Gruppe von 10 Studenten wird die Wahl-Reaktions-Zeit unter den Bedingungen „Streß" (schwacher Elektroschock) und „kein Streß" gemessen. Die Frage ist, ob sich die mittleren Reaktionszeiten der beiden Bedingungen signifikant auf dem 5%-Niveau unterscheiden, d.h., ob die Studenten unter Streß signifikant schneller reagieren.
Unter Streß ergeben sich folgende Werte: 7, 9, 4, 15, 6, 3, 9, 5, 6, 12. Ohne Streß-Bedingung erreichen die Vpn die folgenden Zeiten: 5, 15, 7, 11, 4, 7, 8, 10, 6, 16.

Nr. 5/12

Die Effektivität einer Verhaltenstherapie zum Abbau von Tic-Bewegungen (z.B. Kopfnicken, Nasehochziehen, ...) soll untersucht werden. Die Anzahl der Bewegungen pro 5 Minuten wird jeweils vor und nach der Therapie gezählt. Die 10 Probanden erhalten die folgenden Werte vor der Therapie: 20, 18, 27, 28, 23, 30, 27, 27, 25, 26 und nach der Therapie: 10, 12, 10, 11, 15, 16, 16, 12, 15, 13.

5.3.2.3. Der t-Test für Stichproben mit heterogenen Varianzen

Der empirische t-Wert berechnet sich wie üblich, nur wird für $s_{(M_1-M_2)}$ eine ungewichtete Schätzung vorgenommen, so daß:

$$t = \frac{M_1 - M_2}{\sqrt{\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}}}$$

Der theoretische t-Wert wird in diesem Falle nicht Tab. III entnommen, sondern bestimmt nach der Formel:

$$t = \frac{t_1 \frac{s_1^2}{N_1 - 1} + t_2 \frac{s_2^2}{N_2 - 1}}{\sqrt{\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}}}$$

t_1 und t_2 sind dabei die beiden theoretischen t-Werte aus Tab. III, die den Freiheitsgraden der jeweiligen Stichproben und dem Signifikanzniveau entsprechen.
Anwendungsregeln, Voraussetzungen und Einschränkungen gelten wie bei Kap. 5.3.2.1 u. 5.3.2.2.

5.3.3. Der F-Test nach R. A. Fisher zur Überprüfung der Varianzhomogenität bei zwei oder mehr Stichproben

(1) *Untersuchungsproblem*

Unterscheiden sich die Varianzen zweier oder mehrerer Stichproben signifikant voneinander, oder weisen sie Homogenität (Gleichheit) auf? Wir wollen jetzt also nicht mehr die Mittelwerte von Stichproben oder Populationen miteinander vergleichen, sondern die Streuungen um den jeweiligen Stichprobenmittelwert.

(2) *Beispiel*

Zur Durchführung des t-Tests wird Varianzhomogenität gefordert. In unserem t-Test-Beispiel in Kap. 5.3.2 müßten wir also zunächst mit Hilfe der Prüfstatistik F testen, ob sich die Stichprobenvarianzen signifikant voneinander unterscheiden.

a) Versuchsplan
Siehe t-Test-Beispiel in Kap. 5.3.2, S. 192ff.

b) Hypothesen
H_0: Der empirische Unterschied zwischen den Stichprobenvarianzen ist rein zufällig.
H_1: Der Varianzunterschied ist bei einem Signifikanzniveau von .05 überzufällig (zweiseitige Fragestellung).

c) Ergebnisse
Differenz zwischen s_1 und s_2 (in unserem t-Test-Beispiel = 3)

d) Berechnung
Zur Überprüfung von Varianzhomogenitäten wird die Prüfstatistik F herangezogen, die sich nach R. A. Fisher immer linksschief verteilt. Der empirische F-Wert berechnet sich nach der Formel:

$$F = \frac{s_1^2}{s_2^2}$$

Dabei werden die beiden Stichprobenvarianzen immer so in den Bruch eingesetzt, daß die größere durch die kleinere geteilt wird, so daß der F-Wert nie weniger als 1 beträgt.
Bei mehr als zwei zu vergleichenden Stichproben wird in den Zähler die größte und in den Nenner die kleinste der Varianzen eingesetzt.
Für unser Beispiel ergibt sich ein F-Wert von F = 1.33.

e) Signifikanzprüfung
Der theoretische F-Wert zum Vergleich wird in Tab. IV unter Berücksichtigung des vorher festgelegten Signifikanzniveaus und der beiden Freiheitsgrade für s_1 und s_2 aufgesucht.
Die Freiheitsgrade berechnen sich nach der Formel:

$$df_1 = N_1 - 1; \quad df_2 = N_2 - 1$$

Für unser Beispiel enthält die Tab. IV bei einem Signifikanzniveau von .05 sowie zweiseitiger Fragestellung unter $df_1 = 39$ und $df_2 = 44$ einen F-Wert von F = 1.66.
Wenn nun der empirische F-Wert größer ist als der theoretische, muß H_0 abgelehnt werden, d.h., es besteht keine Varianzhomogenität. Wenn der empirische F-Wert kleiner ist als der theoretische, muß H_1 abgelehnt werden, d.h., es besteht

Varianzhomogenität, und der t-Test (für homogene Varianzen) kann ohne Schwierigkeiten durchgeführt werden.
In unserem Beispiel ist der empirische F-Wert kleiner als der theoretische Tabellenwert, so daß die t-Test-Voraussetzung der Varianzhomogenität erfüllt ist.

(3) *Anwendungsregeln*

a) M_1, s_1, N_1 und M_2, s_2, N_2 sind bekannt.
b) Das Signifikanzniveau und die Freiheitsgrade werden festgelegt.
c) Der F-Wert, der dem zugelassenen Fehler bei ein- oder zweiseitiger Fragestellung entspricht, wird in Tab. IV aufgesucht.
d) Der empirische F-Wert wird nach der o.a. Formel berechnet.
e) Vergleich der beiden F-Werte:
   wenn $F_{emp} < F_{theo}$: H_0 trifft zu
   wenn $F_{emp} \geqq F_{theo}$: H_1 trifft zu

(4) *Voraussetzungen und Einschränkungen*

Für die Anwendung des F-Tests gilt eine wesentliche Voraussetzung, nämlich die, daß die zu vergleichenden Stichproben *unabhängig* voneinander erhoben wurden. Bei *abhängigen* Stichproben ist die Prüfgröße t nach Ferguson zu verwenden:

$$t = \frac{(s_1^2 - s_2^2)\sqrt{N-2}}{\sqrt{4 s_1^2 s_2^2 (1 - r_{12}^2)}}$$

Dabei sind s_1^2 und s_2^2 die Varianzen der beiden Stichproben und r_{12} der Korrelationskoeffizient zwischen den beiden Datenreihen (siehe Korrelation, Kap. 4.1). Der theoretische t-Wert zum Vergleich ist wiederum in Tab. IV nachzuschlagen.

(5) *Übungsbeispiele*

Nr. 5/13
Testen Sie die Gruppe der neurotischen und die Gruppe der nicht-neurotischen Vpn aus Übungsbeispiel Nr. 5/10 hinsichtlich der Lernfähigkeits-Werte auf Varianzhomogenität hin (5%-Niveau).
Nr. 5/14
Die Daten von 3 experimentellen und einer Kontrollgruppe haben die folgenden Varianzen: $s_1^2 = 12$, $s_2^2 = 10$, $s_3^2 = 15$, $s_4^2 = 16$; $N_1 = 10$, $N_2 = 8$, $N_3 = 12$, $N_4 = 20$.
Besteht Varianzhomogenität für diese 4 Gruppen auf dem 1%-Niveau?

5.3.4. Die einfache Varianzanalyse zur Überprüfung von Mittelwertsdifferenzen zwischen mehreren Stichproben

(1) *Untersuchungsproblem*

Unterscheiden sich die Mittelwerte mehrerer Stichproben hinsichtlich eines quantitativen Merkmals signifikant voneinander? Es handelt sich bei den Stichproben um Zufallsstichproben, die der Wirkung verschiedener qualitativer Kategorien einer sog. unabhängigen Variablen unterworfen sind. Mit Hilfe der Varianzanalyse soll nun untersucht werden, ob die unabhängige Variable (z. B. alkoholische Getränke) einen Effekt ausübt auf die abhängige Variable bzw. das zu messende

Merkmal (z. B. die Reaktionsgeschwindigkeit), d. h. ob sich die abhängige Variable gesetzmäßig mit der Variation der unabhängigen ändert. Dieselbe Fragestellung ist uns schon einmal beim Untersuchungsproblem der Regression begegnet, jedoch mit dem Unterschied, daß dort nicht nur die abhängige Variable quantitativ erfaßbar war, sondern auch die unabhängige.
Formal stellt sich das Untersuchungsproblem wie folgt dar. Der Effekt, den die unabhängige Variable auf die abhängige ausübt, wird deutlich in der Differenz zwischen dem Stichprobenmittelwert und dem Gesamtmittelwert, d. h. dem Mittelwert über alle Stichproben:

$E_i = M_i - M_M =$ Effekt der Kategorie i der unabhängigen Variablen auf die abhängige Variable

In der Varianzanalyse wird von einem einfachen, linearen Modell ausgegangen, nach dem sich jeder Meßwert aus drei Komponenten zusammensetzt, aus dem *Gesamtmittelwert* über alle Messungen, dem *Effekt der speziellen Bedingung* der unabhängigen Variablen, dem gerade diese Vp unterworfen ist, und dem *Zufallsfehler,* der mit der Messung verbunden ist:

$x_{ij} = M_N + E_i + e_{ij}$
$M_N =$ Gesamtmittelwert
$E_i =$ Effekt der Bedingung i
$e_{ij} =$ Zufallsfehler, der mit der Messung x_{ij} verbunden ist

Wenn nun die unabhängige Variable keinen Effekt auf die abhängige ausübt, d. h. wenn alle $E = 0$, dann unterscheiden sich die Meßwerte lediglich hinsichtlich des Zufallsfehlers.
Ob dem so ist, kann mit Hilfe der Varianzanalyse untersucht werden.

(2) *Beispiel*

a) Versuchsplan
Im Verkehrsunterricht soll anhand des folgenden Untersuchungsbeispiels die Auswirkung von verschiedenen alkoholischen Getränken auf die Reaktionsgeschwindigkeit von Kraftfahrern demonstriert werden.

Unabhängige Variable:	Alkoholische Getränke	
	Kategorie 1: Bier	(der prozentuale Alkoholgehalt ist bei
	Kategorie 2: Wein	allen Kategorien gleich)
	Kategorie 3: Sekt	
Abhängige Variable:	Reaktionsgeschwindigkeit gemessen in Sekunden (sec.).	

Aus der Menge aller Kraftfahrer wird nun eine genügend große Zufallsstichprobe ermittelt, die wiederum per Zufall für die drei Kategorien der unabhängigen Variablen (Bier, Wein, Sekt) aufgeteilt wird, so daß drei unabhängige Zufallsstichproben entstehen.

b) Hypothesen

H_0: Auftretende Unterschiede hinsichtlich der Reaktionsgeschwindigkeit zwischen den drei Alkoholgruppen sind rein zufällig. Die Art des alkoholischen Getränks spielt für das Ausmaß der Beeinträchtigung der Reaktionsgeschwindigkeit keine Rolle.

H_1: Unterschiedliche alkoholische Getränke beeinträchtigen bei gleichem Alkoholgehalt die Reaktionsgeschwindigkeit in unterschiedlichem Ausmaß.

c) Ergebnisse

Die Daten für die Varianzanalyse lassen sich am sinnvollsten nach folgendem Schema anordnen:

Gruppe 1	...	Gruppe i	...	Gruppe k	
x_{11}	...	x_{i1}	...	x_{k1}	
x_{12}	...	x_{i2}	...	x_{k2}	
...	...	...	...	...	
...	...	...	...	...	
x_{1j}	...	x_{ij}	...	x_{kj}	
...	...	...	...	...	
...	...	...	...	...	
x_{1n}	...	x_{in}	...	x_{kn}	
$\sum x_1$	...	$\sum x_i$	...	$\sum x_k$	$\sum\sum x_{ij}$
M_1	...	M_i	...	M_k	M_N

Dabei bedeuten:

k = Anzahl der Gruppen (Kategorien der unabhängigen Variablen)
i = beliebige Gruppe aus der Anzahl aller Gruppen
N = Anzahl aller Versuchspersonen
x_{ij} = beliebige Messung einer Vp_j in einer beliebigen Gruppe i
$\sum x_i$ = Summe aller Meßwerte in der Gruppe i
$\sum\sum x_{ij}$ = Summe aller Messungen
M_i = Mittelwert der Gruppe i
M_N = Mittelwert über alle N Messungen

Datenschema für unser Beispiel:

	Bier	Wein	Sekt	
	8	5	11	
	3	10	11	
	5	8	6	
	6	8	8	
	3	9	10	
	9	7	9	
	5	4	5	
	6	9	9	
	7	7	10	
	8	9	8	
$\sum x$	60	76	87	740
M	6.0	7.6	8.7	7.4

d) Berechnung

Ziel der Varianzanalyse ist es, zu testen, ob die Varianz der Meßwerte zwischen den Kategorien (Bedingungen) der unabhängigen Variablen (s^2_{zw}) größer ist als die Varianz innerhalb der Kategorien (s^2_{in}).

In unserem Beispiel wäre also zu fragen, ob die Mittelwerte der verschiedenen Alkohol-Kategorien weiter um den Gesamtmittelwert streuen als die Meßwerte innerhalb der jeweiligen Gruppen um den Gruppenmittelwert. Ist die Varianz zwischen den Kategorien größer, so bedeutet das, wie man sich anhand des Datenschemas leicht veranschaulichen kann, daß sich die Gruppen signifikant vom Gesamtmittelwert bzw. auch untereinander unterscheiden, d.h., daß die unabhängige Variable einen Effekt auf die abhängige ausübt.

Formal läßt sich die Abweichung aller Meßwerte vom Gesamtmittelwert gemäß dem linearen Modell in der Varianzanalyse aufgliedern in die Abweichung aller Meßwerte vom jeweiligen Gruppenmittelwert und die Abweichung der Gruppenmittelwerte vom Gesamtmittelwert.

$$\overset{①}{\sum\sum(x_{ij}-M_N)^2} = \overset{②}{\sum\sum(x_{ij}-M_i)^2} + \overset{③}{N\sum(M_i-M_N)^2}$$

Ausdruck 1 stellt den Zähler der Totalvarianz (s^2_{tot}) dar, in Worten: Summe der Abweichungsquadrate aller Einzelwerte vom Gesamtmittelwert summiert über alle Gruppen.

Ausdruck 2 stellt den Zähler der Binnenvarianz (s^2_{in}) dar, in Worten: Summe der Abweichungsquadrate aller Einzelwerte vom Gruppenmittelwert summiert über alle Gruppen.

Ausdruck 3 stellt den Zähler der Zwischenvarianz (s^2_{zw}) dar, in Worten: Summe der Abweichungsquadrate aller Gruppenmittelwerte vom Gesamtmittelwert multipliziert mit N.

Für alle drei Ausdrücke gibt es folgende Formelvereinfachungen:

ad 1: $$\sum\sum(x_{ij}-M_N)^2 = \sum\sum x_{ij}^2 - \frac{(\sum\sum x_{ij})^2}{Nk}$$

ad 2: $$\sum\sum(x_{ij}-M_i)^2 = \sum\sum x_{ij}^2 - \frac{\sum(\sum x_{ij})^2}{N}$$

ad 3: $$N\sum(M_i-M_N)^2 = \frac{\sum(\sum x_{ij})^2}{N} - \frac{(\sum\sum x_{ij})^2}{Nk}$$

Dividiert man nun die Summen der Abweichungsquadrate (SAQ) durch die entsprechenden Freiheitsgrade, so erhält man die mittleren Abweichungsquadratsummen (MAQ), die gesuchten Varianzen.

Ausdruck 1 wird dividiert durch $Nk - 1$
Ausdruck 2 wird dividiert durch $Nk - k$
Ausdruck 3 wird dividiert durch $k - 1$

Das allgemein gebräuchliche Schema zur Darstellung der Berechnungen ist die folgende *Übersichtstabelle:*

Varianzquelle	Summe der Abweichungsquadrate SAQ	df Freiheitsgrade	Mittlere Abweichungsquadratsumme MAQ
Zwischenvarianz	$\frac{\Sigma(\Sigma x_{ij})^2}{N} - \frac{(\Sigma\Sigma x_{ij})^2}{Nk}$	$k-1$	$s^2_{zw} = \frac{SAQ_{zw}}{k-1}$
Binnenvarianz	$\Sigma\Sigma x^2_{ij} - \frac{\Sigma(\Sigma x_{ij})^2}{N}$	$Nk-k$	$s^2_{in} = \frac{SAQ_{in}}{Nk-k}$
Totalvarianz	$\Sigma\Sigma x^2_{ij} - \frac{(\Sigma\Sigma x_{ij})^2}{Nk}$	$Nk-1$	$s^2_{tot} = \frac{SAQ_{tot}}{Nk-1}$

Für unser Beispiel berechnen sich folgende Werte:

Varianzquelle	SAQ	df	MAQ
ZW	$\frac{16945}{10} - \frac{49730}{30} = 36.5$	2	$\frac{36.5}{2} = 18.3$
IN	$1801 - \frac{16945}{10} = 106.5$	27	$\frac{106.5}{27} = 3.9$
TOT	$1801 - \frac{49730}{30} = 143$	29	$\frac{143}{29} = 4.9$

e) Signifikanzprüfung

Obwohl die Varianzanalyse ein Verfahren zur Aufdeckung von Mittelwertsdifferenzen ist, wird der Signifikanztest über die Prüfgröße F vorgenommen, das darf bei der Interpretation der Ergebnisse nicht vergessen werden.

Geprüft wird, ob sich Binnen- und Zwischenvarianz signifikant voneinander unterscheiden. Wenn sie identisch sind, d.h. wenn H_0 zutrifft, verteilt sich ihr Bruch nach der F-Verteilung. Je größer die Differenz zwischen den beiden Varianzen ist, um so weiter weicht der Varianzquotient von der F-Verteilung ab, d.h. um so eher trifft H_1 zu.

Die Fragestellung beim F-Test ist in der Varianzanalyse immer einseitig. Es interessiert lediglich, ob die Zwischenvarianz größer ist als die Binnenvarianz. Bei exakter Versuchsplanung und -durchführung kann die Binnenvarianz die Zwischenvarianz nicht überschreiten. Formal lautet der F-Test in der Varianzanalyse:

$$F = \frac{s^2_{zw}}{s^2_{in}} \quad \text{mit} \quad df_1 = k-1 \quad \text{und} \quad df_2 = Nk-k$$

Der theoretische F-Wert wird in Tab. IV unter Berücksichtigung des vorher festgelegten Signifikanzniveaus bei einseitiger Fragestellung und der entsprechenden Freiheitsgrade aufgesucht.

Wenn der empirische F-Wert kleiner ist als der theoretische, trifft H_0 zu, d.h., die unabhängige Variable übt keinerlei Effekt aus auf die abhängige Variable.

Wenn der empirische F-Wert größer ist als der theoretische, trifft H_1 zu, d.h., die unabhängige Variable beeinflußt die abhängige Variable in überzufälligem Ausmaß.
Dabei ist dann jedoch noch ungewiß, welche Kategorie der unabhängigen Variablen einen Effekt bewirkt, bzw. ob alle Kategorien die abhängige Variable beeinflussen.

In unserem Beispiel ergibt sich ein F-Wert von $F = \frac{18.3}{3.9} = 4.69$

Bei einem Signifikanzniveau von .05, einseitiger Fragestellung und $df_1 = 2$, $df_2 = 27$ finden wir in Tab. IV einen theoretischen F-Wert von $F = 3.35$. Da F_{emp} größer ist als F_{theo}, können wir mit 95% Sicherheit sagen, daß unterschiedliche alkoholische Getränke bei gleichem prozentualem Alkoholgehalt einen unterschiedlichen Einfluß ausüben auf die Reaktionsgeschwindigkeit von Kraftfahrern. *Welche* der drei alkoholischen Getränke sich in ihrer Wirkungsweise auf die Reaktionsgeschwindigkeit voneinander signifikant unterscheiden, kann aufgrund der Varianzanalyse *nicht* gesagt werden, nur *daß* sie sich unterscheiden!

(3) *Anwendungsregeln*

a) Datensammlung nach dem o.a. Schema (S. 200).
b) Das Signifikanzniveau und die Freiheitsgrade werden festgelegt.
c) Die Varianzen werden berechnet und in der o.a. Übersichtstabelle zusammengefaßt (S. 202).
d) Der F-Wert, der dem zugelassenen Fehler bei einseitiger Fragestellung entspricht, wird nach Tab. IV bestimmt.
e) Der empirische F-Wert wird nach der o.a. Formel bestimmt (S. 202).
f) Vergleich der beiden F-Werte:
wenn $F_{emp} < F_{theo}$: H_0 trifft zu
wenn $F_{emp} \geqq F_{theo}$: H_1 trifft zu

(4) *Voraussetzungen und Einschränkungen*

a) Die abhängige Variable muß kontinuierlich auf einer Intervallskala meßbar sein.
b) Die Fehlerverteilung innerhalb jeder Kategorie der unabhängigen Variablen muß normal sein. Wenn das nicht der Fall ist, müssen die Stichproben relativ groß sein.
c) Es soll Varianzhomogenität für alle Kategorien bestehen, d.h., vor der eigentlichen Varianzanalyse sollte ein F-Test durchgeführt werden, der die Varianzen der Stichprobe auf Homogenität hin prüft. Für den F-Bruch wird die größte der Varianzen durch die kleinste geteilt. Diese Voraussetzung kann jedoch ohne großes Risiko verletzt werden, wenn alle Stichproben gleich groß sind.
d) Die Kategorienstichproben müssen völlig unabhängig voneinander sein. Diese Annahme darf keinesfalls verletzt werden. Abhängigkeit der Stichproben kann jedoch nur experimentell verhindert werden (s. Kap. 2.2.4).

(5) *Übungsbeispiele*

Nr. 5/15
Im Rahmen einer größeren Studie, die den empirischen Gehalt von Volksweisheiten überprüft, soll untersucht werden, wie es in dieser Hinsicht mit dem Spruch bestellt ist: „Die

dümmsten Bauern haben die dicksten Kartoffeln." Als Maß für die Dicke der Kartoffeln soll das durchschnittliche Gewicht von 100 nach Zufall ausgewählten Kartoffeln pro Bauer dienen. Für die unabhängige Variable „Dummheit" gibt es drei Kategorien: „starke D.", „mittlere D." und „geringe D.". Pro Kategorie gibt es 10 Vpn. Unterscheiden sich die 3 Gruppen hinsichtlich des Kartoffelgewichts, d.h., trifft die Volksweisheit für die untersuchte Stichprobe in Abbesbüttel zu?

Datenmatrix:

Vp	große D.	mittlere D.	geringe D.	Σ
1	20	25	21	
2	20	21	23	
3	19	21	20	
4	25	23	18	
5	29	21	17	
6	19	22	20	
7	20	26	25	
8	24	18	26	
9	27	19	23	
10	25	22	20	
Σx	228	218	213	659
Σx^2	5118	4806	5593	15517
M	22.8	21.8	21.3	

Nr. 5/16
Die Effektivität von 3 Förderkursen ($k = 3$) soll getestet werden. 24 Schüler stehen zur Verfügung, die per Zufall den 3 verschiedenen Kursen zugeordnet werden ($n = 8$). Nach der Kurs-Durchführung werden alle 24 Schüler mit einem Kenntnistest geprüft, wobei sich für die 3 Gruppen folgende Werte ergeben:

Vp	Kurs 1	Kurs 2	Kurs 3	Summe
1	3	4	6	
2	5	4	7	
3	2	3	8	
4	4	8	6	
5	8	7	7	
6	4	4	9	
7	3	2	10	
8	9	5	9	
Σx	38	37	62	137
Σx^2	224	199	496	919
M	4.75	4.62	7.75	

5.3.5. Der Newman-Keuls-Test

(1) *Untersuchungsproblem*

Das Untersuchungsproblem für die Durchführung dieses Tests ergibt sich nach vollzogener Varianzanalyse, und zwar, wenn sich die unabhängige Variable tatsächlich als mitbedingend für die abhängige Variable erwiesen hat, wenn also der F-Test zeigt, daß sich irgendwelche oder alle Stichprobenmittelwerte voneinander unterscheiden.
Die Frage lautet jetzt: Welche Kategorien der unabhängigen Variablen unterscheiden sich signifikant voneinander?
Welche Kategorie oder welche Kombination von Kategorien hat die größte Bedeutung für den in der vorangegangenen Varianzanalyse festgestellten Effekt der unabhängigen Variablen auf die abhängige?

(2) *Beispiel*

a) Versuchsplan
Siehe unser Beispiel zur Varianzanalyse.
Frage: Welche der drei möglichen einfachen Mittelwertsdifferenzen der 3 Versuchsbedingungen erweisen sich als signifikant?
b) Hypothesen
H_0 und H_1 werden getrennt für jeden Mittelwertsvergleich aufgestellt.
c) Ergebnisse
Die empirischen Mittelwertsdifferenzen lassen sich am zweckmäßigsten entsprechend folgender Matrix darstellen:

	M_1	M_2	...	M_i	...	M_k
M_1	—	$(M_1—M_2)$	...	$(M_1—M_i)$	...	$(M_1—M_k)$
M_2		—	...	$(M_2—M_i)$	...	$(M_2—M_k)$
...			—	...	...	...
M_i				—	...	$(M_i—M_k)$
...					—	...
M_k						—

Für unser Beispiel ergibt sich folgende Matrix:

	$M_{(Bier)}$	$M_{(Wein)}$	$M_{(Sekt)}$
$M_{(Bier)}$	—	1.6	2.7
$M_{(Wein)}$		—	1.1
$M_{(Sekt)}$			—

Für den Newman-Keuls-Test wesentlich ist bei der Anordnung der Datenmatrix, daß die Mittelwerte der Größe nach geordnet sind, da die Verteilung von Mittelwertsdifferenzen davon abhängt, ob die nacheinander verbleibenden jeweils größten Differenzen gebildet wurden. Nur in diesem Fall kann die Prüfstatistik des Newman-Keuls-Tests angewandt werden.

d) Berechnung

Die Prüfstatistik im Newman-Keuls-Test ist:

$$q = \frac{M_j - M_h}{\sqrt{\frac{s_{in}^2}{N}}}, \quad \text{wobei: } \begin{aligned} M_j &= \text{Mittelwert der Stichprobe j} \\ M_h &= \text{Mittelwert der Stichprobe h} \\ N &= \text{Größe der Stichproben} \\ s_{in}^2 &= \text{Binnenvarianz} \end{aligned}$$

Wenn die Stichproben mit unterschiedlich vielen Vpn besetzt sind, so bestimmt sich $\bar{n}_j$ aus dem harmonischen Mittel der Stichprobengrößen:

$$\bar{n}_j = \frac{k}{1/n_1 + 1/n_2 + \ldots + 1/n_j + \ldots + 1/n_k}$$

Für unser Beispiel ergibt sich:

$$q_{Sekt/Bier} = 3.7$$
$$q_{Wein/Bier} = 2.2$$
$$q_{Wein/Sekt} = 1.5$$

e) Signifikanzprüfung

Der theoretische q-Wert zum Vergleich wird in Tab. V unter Berücksichtigung des vorher festgelegten Signifikanzniveaus sowie der Freiheitsgrade der Binnenvarianz und der sog. Spannweite aufgesucht. Die Spannweite ist die um 1 vergrößerte Differenz zwischen den Indizes der beiden verglichenen Stichproben (die kleinste Stichprobe hat den Index 1, die größte den Index k):

$$S = j - i + 1$$

Wenn nun der empirische q-Wert kleiner ist als der Tabellenwert, ist der Unterschied zwischen den Mittelwerten der verglichenen Stichproben nicht signifikant, wenn der empirische Wert größer ist als der theoretische, so besteht ein überzufälliger Unterschied zwischen den beiden Mittelwerten. Die Reihenfolge der Mittelwertsvergleiche erfolgt der Größe der Spannweite nach, d.h., von unserer Matrix ausgehend, von oben rechts nach unten links.

Für unser Beispiel finden wir in Tab. V bei einem Signifikanzniveau von .05 für $df = 27$ und $S = 3$ ein q von 3.55 und für $S = 2$ ein q von 2.92.

Das heißt:

$$q_{Sekt/Bier} > 3.53$$
$$q_{Wein/Bier} < 2.92$$
$$q_{Wein/Sekt} < 2.92$$

Der einzige, signifikante Mittelwertsunterschied hinsichtlich der Wirkung auf die Reaktionsfähigkeit besteht demnach zwischen Bier und Sekt. Wein und Bier und auch Wein und Sekt unterscheiden sich nicht in ihrer diesbezüglichen Wirkung. (Achtung, das Beispiel ist konstruiert, die Daten wurden nicht empirisch gewonnen!)

(3) *Anwendungsregeln*

a) Der F-Test der Varianzanalyse war signifikant.
b) Die Mittelwertsdifferenzen werden der Größe nach geordnet.
c) Das Signifikanzniveau für die Mittelwertsvergleiche wird festgelegt.
d) Der q-Wert, der dem zugelassenen Fehler entspricht, wird für die Freiheitsgrade der Binnenvarianz der Varianzanalyse und für die Spannweite in Tab. V aufgesucht.

e) Der empirische q-Wert für die größte vorhandene Spannweite wird berechnet nach der o.a. Formel.
f) Vergleich der beiden q-Werte:
wenn $q_{emp} < q_{theo}$: H_0 trifft zu
wenn $q_{emp} \geqq q_{theo}$: H_1 trifft zu
g) Die Signifikanzprüfungen werden entsprechend der Größe der Spannweiten so lange von oben nach unten fortgesetzt, bis sich die Mittelwertsdifferenzen nicht mehr als signifikant erweisen.

(4) *Voraussetzungen und Einschränkungen*

a) Der F-Test in der Varianzanalyse muß sich als signifikant erwiesen haben.
b) Alle k Mittelwerte der Stichproben müssen der Größe nach geordnet sein, so daß die jeweils größten Differenzen gebildet werden können.
c) Wenn *Gruppen* von Stichprobenmittelwerten der in die Varianzanalyse eingegangenen Stichproben miteinander oder mit einem Mittelwert verglichen werden sollen, muß der etwas gröbere Scheffé-Test angewendet werden (siehe z.B. Fröhlich & Becker 1971).
d) Es sollten nur inhaltlich sinnvolle Mittelwertsvergleiche im Anschluß an die Varianzanalyse vorgenommen werden.

(5) *Übungsbeispiel*

Nr. 5/17
Fragestellung s. Übungsbeispiel Nr. 5/16:
Welche der 3 Kurse unterscheiden sich signifikant voneinander?

5.3.6. Der Omega-Wert zur Schätzung der Stärke einer statistischen Beziehung

Untersuchungsproblem

Ebenfalls im Anschluß an eine Varianzanalyse erhebt sich die Frage nach dem Ausmaß, in dem die unabhängige Variable die abhängige bedingt, d.h., wie stark der Zusammenhang zwischen beiden ist. Die Varianzanalyse gibt Antwort darauf, *ob* überhaupt ein Zusammenhang besteht, d.h. ob Differenzen auftreten. Der Newman-Keuls-Test beantwortet die Frage, *welche* Mittelwertsunterschiede bestehen, und der Omega-Wert nun bestimmt, *wie bedeutsam* die Mittelwertsdifferenzen, d.h. die Assoziationsstärke zwischen abhängiger und unabhängiger Variable, sind.

Beispiel

Analog zu unserem Varianzanalyse-Beispiel wollen wir fragen, in welchem Ausmaß verschiedene alkoholische Getränke die Reaktionsfähigkeit von Berufskraftfahrern beeinflussen, inwieweit die Werte in einem Reaktionsgeschwindigkeits-Test dadurch vorhergesagt werden können, daß bekannt ist, welches alkoholische Getränk eine Vp zu sich genommen hat.

Berechnung

Die Assoziationsstärke zwischen den beiden Variablen wird mit dem griechischen Buchstaben Omega (ω) bezeichnet. Dieses Maß repräsentiert die relative Reduktion

der Varianz der abhängigen Variablen (hier: Reaktionsgeschwindigkeit), die erreicht wird durch die Kenntnis der Kategorien der unabhängigen Variablen (Bier, Wein oder Sekt), unter deren Bedingung die jeweiligen Messungen gemacht werden. Omega^2 ist ein Wahrscheinlichkeitsmaß, das alle Werte zwischen 0 und 1 annehmen kann. 0 bedeutet, daß die Werte der abhängigen Variablen keineswegs durch die Kenntnis der unabhängigen Variablen vorhergesagt werden können, 1 bedeutet, daß die Werte der abhängigen Variablen vollständig durch die unabhängige Variable determiniert sind.
Die Schätzung von Omega^2 wird nach folgender Formel vorgenommen:

$$\text{Omega}^2 = \frac{SAQ_{zw} - (k-1)\,MAQ_{in}}{SAQ_{tot} + MAQ_{in}}$$

Für unser Beispiel ergibt sich ein Omega^2 von .19. Das heißt, daß der prozentuale Anteil der Varianz der Reaktionsgeschwindigkeit von Berufskraftfahrern, der durch den Genuß der 3 verschiedenen alkoholischen Getränke hervorgerufen wurde, also nicht auf Fehlern beruht, sondern systematischer Art ist, 19 Prozent beträgt. Immerhin: 81 Prozent der Varianz können nicht auf die Verschiedenheit der Getränke zurückgeführt werden.

Übungsbeispiel

Nr. 5/18
Fragestellung siehe Übungsbeispiel Nr. 5/16
Wie stark ist der Zusammenhang zwischen abhängiger und unabhängiger Variable?

5.4. Nonparametrische Verfahren zur Hypothesenprüfung

Im Gegensatz zu den in Kap. 5.3 erörterten Verfahren machen die sogenannten nonparametrischen (parameterfreien, verteilungsfreien) Verfahren keine Voraussetzungen über die Verteilung der Parameter in der Population. Es wird keine Normalverteilung der Daten in der Stichprobe oder der Population gefordert, die Daten brauchen kein Intervallskalenniveau aufzuweisen (es genügt Ordinal- oder auch Nominalskalenniveau). Natürlich müssen auch bei den nichtparametrischen Verfahren bestimmte Bedingungen erfüllt werden, jedoch sind diese in der Regel nicht so weitgehend wie bei den parametrischen Methoden. Im wesentlichen wird gefordert, daß Zufallsstichproben vorliegen und daß das Merkmal stetig verteilt ist. Andererseits sind die verteilungsfreien Verfahren auch weniger effizient als parametrische Prüfmethoden, d.h., für ein bestimmtes Signifikanzniveau benötigt man bei einem nichtparametrischen Test eine größere Stichprobe als bei einem parametrischen. Trotz der relativ einfachen Berechnungsweisen der nichtparametrischen Verfahren sollte man sie nur dann anwenden, wenn die Voraussetzungen der parametrischen Methoden nicht erfüllt werden können. In der nachstehenden Abb. 31 geben wir einen Überblick über die hier behandelten nichtparametrischen Tests.

	unabhängige Stichproben		abhängige Stichproben	
	Vergleich *zweier* Stichpr.	Vergleich *mehrerer* Stichpr.	Vergleich *zweier* Stichpr.	Vergleich *mehrerer* Stichpr.
Häufigkeiten (Nominalskala)	Chi^2-Test (s. S. 213ff.)	Chi^2-Test (s. S. 217ff.)	McNemar-Test (s. S. 233ff.)	Cochran-Q-Test (s. S. 235ff.)
Rangdaten (Ordinalskala)	U-Test von Mann-Whitney (s. S. 226ff.)	H-Test von Kruskal-Wallis – Rangvarianzanalyse – (s. S. 230ff.)	Wilcoxon-Test (s. S. 238ff.)	Friedman-Rangvarianzanalyse (s. S. 241ff.)

Abb. 31: Parameterfreie Prüfmethoden (Übersichtsschema)

5.4.1. Der Chi^2-Test

5.4.1.1. Der einfache Chi^2-Test (Ein-Stichprobenfall)

(1) *Untersuchungsproblem*

Unterscheiden sich die bei einer Stichprobe *empirisch ermittelten* Häufigkeiten bezüglich eines Merkmals signifikant von den für die Kategorien dieses Merkmals *erwarteten* Häufigkeiten?

(2) *Beispiel*

a) Versuchsplan
Ein Lehrer verteilt einen Fragebogen an seine Schüler mit der Frage: Lest ihr die politischen Nachrichten in eurer Tageszeitung?
Antwortkategorien: „immer“ „meist“ „selten“ „nie“
Stichprobe: 40 Schüler der 5. Klasse eines Gymnasiums in X
Merkmal: Interesse an politischen Tagesfragen

b) Hypothesen
Nullhypothese (H_0):
Es besteht kein Unterschied in den Häufigkeitsbesetzungen der vier Kategorien.
Alternativhypothese (H_1):
Die Häufigkeiten in den vier Kategorien unterscheiden sich überzufällig.
Signifikanzniveau: .05

c) Ergebnisse

immer	meist	selten	nie
13	10	8	9

$k = 4$
$N = 40$

d) Berechnung
Für die Analyse solcher oder ähnlicher Daten kann der Chi^2-Test verwendet werden.
Chi^2 wird nach folgender Formel berechnet:

$$Chi^2 = \sum_{i=1}^{k} \frac{(0_i - E_i)^2}{E_i}$$

0_i = beobachtete („observed") Zahl der Fälle in der i-ten Kategorie;
bei $i = 1$ also $0_1 = 13$
bei $i = 2$ also $0_2 = 10$ usw.

E_i = erwartete („expected") Zahl der Fälle in der i-ten Kategorie

$\sum_{i=1}^{k}$ = Summation des Ausdrucks $\frac{(0_i - E_i)}{E_i}$ über alle k Kategorien

Wie die Formel zeigt, werden den beobachteten Häufigkeiten „erwartete" Häufigkeiten gegenübergestellt. Woher erhält man nun diese erwarteten Häufigkeiten? Dafür gibt es verschiedene Möglichkeiten. Es kann sein, daß man den Fragebogen schon im vorigen Jahr verteilt hatte. Dann kann man die damals erhaltenen Antworthäufigkeiten heranziehen, um damit etwa zu prüfen, ob sich das politische Interesse der Schüler von einem Jahr zum anderen verändert hat. Erwartete Häufigkeiten sind dann die im Vorjahr erhobenen Zahlen.
In der Regel wird man aber – wenn keine anderen Daten vorliegen und auch theoretisch begründete Erwartungen hinsichtlich der Verteilung der Antworthäufigkeiten fehlen – die empirisch ermittelten Häufigkeiten der sogenannten *Gleichverteilung* gegenüberstellen. D.h. man erwartet, daß sich die Antworthäufigkeiten auf die Kategorien gleichmäßig verteilen (s. unser Beispiel).
Die erwarteten Häufigkeiten werden berechnet:

$$E_i = \frac{N}{k}$$

N = Gesamtzahl aller Versuchspersonen
k = Anzahl der Kategorien
Dies ergibt für unser Beispiel:

$$E = \frac{40}{4} = 10$$

Wir tragen die erwarteten Häufigkeiten in unsere Tabelle ein:

	immer	meist	selten	nie
beobachtete Häufigkeiten	13	10	8	9
erwartete Häufigkeiten	10	10	10	10

Nun setzen wir die entsprechenden Werte in die Formel ein:

$$\mathrm{Chi}^2 = \frac{(13-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(9-10)^2}{10}$$

$$= \frac{9}{10} + 0 + \frac{4}{10} + \frac{1}{10}$$

$$= .9 + 0 + .4 + .1 = \underline{\underline{1.4}}$$

e) Signifikanzprüfung
Als Prüfverteilung benutzt man die Chi-Quadrat-Verteilung (Tab. XI). Die Zahl der Freiheitsgrade beträgt df[1] = $k - 1$, im Beispiel also $df = 4 - 1 = 3$.
Man stellt nun fest, ob der errechnete Wert den Tabellenwert bei $df = 3$ und Signifikanzniveau .05 über- oder unterschreitet. Für unser Beispiel finden wir einen Tabellenwert von 7.82. Der errechnete Wert ist kleiner als der Tabellenwert, d.h. also, unsere beobachteten Werte weichen von der angenommenen Gleichverteilung nicht signifikant, sondern nur zufällig ab. Die Nullhypothese kann nicht zurückgewiesen werden.

(3) *Anwendungsregeln*

a) Die beobachteten Häufigkeiten („0") werden für jede Kategorie ausgezählt.
b) Die erwarteten Häufigkeiten („E") werden für jede Kategorie berechnet $\left(\text{bei Annahme der Gleichverteilung } E_i = \frac{N}{k}\right)$, wenn nicht schon Erwartungswerte vorliegen.
c) Der Wert für Chi^2 wird nach der Formel $\mathrm{Chi}^2 = \sum_{i=1}^{k} \frac{(0_i - E_i)^2}{E_i}$ berechnet.
d) Die Zahl der Freiheitsgrade wird errechnet: $df = k - 1$
e) In Tab. XI wird unter Berücksichtigung des gewählten Signifikanzniveaus und der Zahl der Freiheitsgrade der kritische Chi^2-Wert abgelesen. Ist der errechnete Chi^2-Wert kleiner als der in der Tabelle abgelesene, können wir H_0 nicht zurückweisen; ist er gleich oder größer, weisen wir H_0 zurück.

(4) *Voraussetzungen und Einschränkungen*

a) Die Kategorien des Merkmals müssen einander ausschließen, d.h., jede Beobachtung darf immer nur einer Kategorie zugeordnet werden können.
b) Die beobachteten Werte müssen unrelativierte Häufigkeiten sein, also z.B. keine Prozentwerte.
c) Die Summe der Häufigkeiten muß gleich sein der Zahl der Vpn, von jeder Vp darf also nur ein Meßwert stammen.
d) Bei $k = 2$ Kategorien müssen die erwarteten Häufigkeiten $\geqq 5$ sein (sonst Binomialtest, s. Siegel 1956).
e) Bei $k > 2$ Kategorien darf höchstens 1/5 der Erwartungswerte < 5 sein, wobei aber kein Erwartungswert < 1 sein darf.
Wenn inhaltlich möglich, können Kategorien zusammengefaßt werden, um so die Erwartungswerte zu vergrößern.

---

1 *df = degrees of freedom*

f) Bei k = 2 Kategorien (df = 1) ist zur Berechnung von Chi² die Kontinuitätskorrektur nach Yates zu berücksichtigen:

$$\text{Chi}^2 = \sum_{i=1}^{k} \frac{(|0_i - E_i| - .5)^2}{E_i}$$

(5) *Übungsbeispiel*

Nr. 5/19

Folgende Verteilung der Mathematikzensuren in einer Klasse ist gegeben:

Note	1	2	3	4	5
Häufigkeit	5	15	30	30	20

Es ist die Nullhypothese zu prüfen: Alle Zensuren werden gleich häufig gegeben.

5.4.1.2. Der Chi²-Test zum Vergleich von zwei oder mehreren (k) unabhängigen Stichproben (Komplexer Chi²-Test)

Abb. 32 gibt einen Überblick über die Anwendungsmöglichkeiten des Chi²-Tests. Welches Chi²-Verfahren im konkreten Fall zu verwenden ist, hängt ab von der Zahl der zu vergleichenden Stichproben und der Art des vorliegenden Merkmals (Alternativmerkmal oder mehrklassiges Merkmal). Die einzelnen Verfahren werden im folgenden ausführlich beschrieben.

	Vergleich *zweier* Stichproben (k = 2)	Vergleich *mehrerer* Stichproben (k > 2)
Alternativ verteiltes *Merkmal* (r = 2)	2 × 2 A B C D Vierfelder-Chi² $\text{Chi}^2 = \frac{N(\lvert AD - BC\rvert - \frac{N}{2})^2}{(A+B)(C+D)(A+C)(B+D)}$ df = 1	1 2 3 ... k Stichproben 2 × k *Merkmal* 1 2 $\text{Chi}^2 = \sum_{i=1}^{2}\sum_{j=1}^{k} \frac{(0_{ij} - E_{ij})^2}{E_{ij}}$ df = (k − 1)
Mehrklassiges Merkmal (r > 2)	1 2 Stichproben r × 2 *Merkmal* 1 2 3 ... r $\text{Chi}^2 = \sum_{i=1}^{r}\sum_{j=1}^{2} \frac{(0_{ij} - E_{ij})^2}{E_{ij}}$ df = (r − 1)	1 2 3 ... k Stichproben r × k *Merkmal* 1 2 3 ... r $\text{Chi}^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(0_{ij} - E_{ij})^2}{E_{ij}}$ df = (r − 1)(k − 1)

Abb. 32: Übersicht über die komplexen Chi²-Tests

A) Vierfelder-Chi²-Test zum Vergleich von 2 unabhängigen Stichproben (Alternativmerkmal) – Fall: 2×2

(1) *Untersuchungsproblem*

Unterscheiden sich *zwei* unabhängige Gruppen hinsichtlich ihrer Häufigkeitsverteilung bei einem *Alternativmerkmal?*
Es existieren also 2 unabhängige Gruppen (z.B. Jungen und Mädchen) und ein zweiklassiges Merkmal (z.B. ja/nein; Aufgabe gelöst/Aufgabe nicht gelöst).

(2) *Beispiel*

a) Versuchsplan
Der Lehrer will wissen, ob Mädchen und Jungen eine bestimmte Rechenaufgabe unterschiedlich häufig lösen können. Dazu legt er einer Gruppe von 50 Jungen und 40 Mädchen diese Aufgabe vor.
Stichprobe 1: 40 Mädchen; Stichprobe 2: 50 Jungen;
Merkmal: Aufgabe gelöst (+), Aufgabe nicht gelöst (−).

b) Hypothesen
H_0: Auftretende Unterschiede in der Lösungshäufigkeit bei Mädchen und Jungen sind rein zufällig.
H_1: Mädchen und Jungen lösen die Aufgabe unterschiedlich häufig.
Signifikanzniveau: .05

c) Ergebnisse
Die Ergebnisse lassen sich in einer sogenannten *Vierfeldertafel* zusammenstellen:

		Jungen	Mädchen	
Aufgabe	+	A 32	B 14	A + B 46
	−	C 18	D 26	C + D 44
		A + C 50	B + D 40	N = 90

d) Berechnung
Bei Daten, die sich in der Form einer Vierfeldertafel darstellen lassen, kann Chi² über eine vereinfachte Methode berechnet werden.

$$\mathrm{Chi}^2 = \frac{N\left(|AD - BC| - \frac{N}{2}\right)^2}{(A+B)(C+D)(A+C)(B+D)}; \quad df = 1$$

Im Beispiel ergibt sich: $\mathrm{Chi}^2 = 6.36$

e) Signifikanzprüfung
In Tab. XI lesen wir bei df = 1 Signifikanzniveau .05 den Wert 3.84 ab. Da der errechnete Wert größer ist als der abgelesene, ist die Nullhypothese zurückzuweisen. Mädchen und Jungen lösen die Aufgabe also unterschiedlich häufig.

(3) *Anwendungsregeln*

a) Die beobachteten Häufigkeiten werden in eine Vierfeldertafel eingetragen.
b) Chi^2 wird nach der Formel

$$\text{Chi}^2 = \frac{N\left(|AD - BC| - \frac{N}{2}\right)^2}{(A+B)(C+D)(A+C)(B+D)}$$

berechnet.

c) In Tab. XI wird unter Berücksichtigung der Freiheitsgrade ($df = 1$) und des gewählten Signifikanzniveaus der kritische Chi^2-Wert abgelesen. Ist der errechnete Wert kleiner als der abgelesene, behalten wir H_0 bei, ist er gleich oder größer, dann können wir H_0 zurückweisen.

(4) *Voraussetzungen und Einschränkungen*

a) Es gelten die beim einfachen Chi^2-Test unter 4a bis 4c gemachten Angaben (S. 211).
b) Wenn N zwischen 20 und 40 liegt, kann die angegebene Formel nur verwendet werden, wenn alle Erwartungswerte > 5 sind (zur Berechnung der Erwartungswerte s. u. Abschnitt B, 2d).
Ist eine erwartete Häufigkeit < 5, muß auf dieses Verfahren verzichtet werden und statt dessen der Fisher-Test (s. Siegel 1956) oder Kullbacks $2\,\hat{I}$-Test (s. Kap. 5.4.2) herangezogen werden.
c) Ist $N < 20$, kann das Verfahren nicht verwendet werden (dann Fisher-Test, s. Siegel 1956).
d) Ist die Stichprobe $N > 100$, darf eine unkorrigierte Formel verwendet werden, wenn die Voraussetzungen hinsichtlich der Erwartungswerte erfüllt sind:

$$\text{Chi}^2 = \frac{N(AD - BC)^2}{(A+B)(C+D)(A+C)(B+D)}; \quad df = 1$$

(5) *Übungsbeispiel*

Nr. 5/20

24 Mädchen und 19 Jungen wurden befragt, ob sie ihre Ferien gemeinsam mit den Eltern verbringen oder nicht. Es ergab sich:

	Mädchen	Jungen
mit Eltern	20	10
ohne Eltern	4	9

Es ist zu prüfen, ob zwischen Mädchen und Jungen ein Unterschied besteht hinsichtlich ihrer Feriengestaltung.

B) Chi^2-Test zum Vergleich von 2 unabhängigen Stichproben (mehrklassiges Merkmal) – Fall: $r \times 2$

Die im folgenden dargestellte Methode zur Chi^2-Berechnung kann auch auf Vierfeldertafeln angewendet werden, obwohl man aus rechenökonomischen Gründen die oben angeführten Formeln bevorzugen wird.

(1) *Untersuchungsproblem*

Unterscheiden sich *zwei* unabhängige Gruppen hinsichtlich ihrer Häufigkeitsverteilung bei einem *mehrklassigen* Merkmal?
Es existieren hier also ein in $k > 2$ Kategorien aufgeteiltes Merkmal (z.B. überdurchschnittlich intelligent) und die Häufigkeitsverteilungen zweier Gruppen hinsichtlich dieses Merkmals.

(2) *Beispiel* (fiktiv)

a) Versuchsplan
Es soll geprüft werden, ob Kinder aus ländlichen und städtischen Wohnregionen sich in bezug auf ihren Schulerfolg unterscheiden.
98 Schüler wurden von ihren Lehrern bezüglich ihres Schulerfolgs eingestuft in überdurchschnittlich erfolgreiche, durchschnittlich erfolgreiche und unterdurchschnittlich erfolgreiche.
Stichprobe 1: 46 Landkinder; Stichprobe 2: 52 Stadtkinder (Zufallsstichproben).
Merkmal: überdurchschnittlich erfolgreich ($>\varnothing$); durchschnittlich erfolgreich ($\varnothing$); unterdurchschnittlich erfolgreich ($<\varnothing$).

b) Hypothesen
H_0: Stadt- und Landkinder unterscheiden sich nicht hinsichtlich ihres Schulerfolgs.
H_1: Stadt- und Landkinder haben in der Schule einen unterschiedlichen Erfolg.
Signifikanzniveau: .05

c) Ergebnisse
Die Häufigkeiten werden in eine Tafel mit 2 (k) Spalten und 3 (r) Reihen eingetragen:

		Stadt	Land	
	$>\varnothing$	15	29	44
Schulerfolg	$\varnothing$	25	11	36
	$<\varnothing$	12	6	18
		52	46	98

d) Berechnung
Chi^2 wird hier nach folgender Formel berechnet:

$$Chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}; \quad df = (r-1)$$

Es bedeuten:

O_{ij} = die beobachtete Häufigkeit in der i-ten Reihe der j-ten Spalte ($O_{11} = 15$; $O_{31} = 12$; $O_{22} = 11$; usw.)

E_{ij} = die erwartete Häufigkeit in der i-ten Reihe der j-ten Spalte (analog zu „O")

$\sum_{i=1}^{r} \sum_{j=1}^{2}$ = Für jede Zelle ist der Ausdruck $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ zu berechnen und über alle Zellen die Summe zu bilden.

Die *Freiheitsgrade* betragen $df = (r-1)(2-1)$ oder allgemein $df = (r-1)(k-1)$, also *Zahl der Reihen minus 1 mal Zahl der Spalten minus 1;* im Beispiel: $(3-1)(2-1) = 2$.

Die Berechnung der *erwarteten Häufigkeiten* erfolgt über die Randsummen nach folgender Beziehung:

$$E_{ij} = \frac{\text{Reihensumme} \times \text{Spaltensumme}}{N}$$

Die erwartete Häufigkeit für die Zelle in der 1. Reihe und der 1. Spalte (mit der beobachteten Häufigkeit 15) lautet also:

$$E_{11} = \frac{52 \cdot 44}{98} = 23.4$$

E_{12} ergibt sich aus: $\frac{46 \cdot 44}{98} = 20.7$

Ob bei einer Vierfeldertafel die erwarteten Häufigkeiten groß genug sind (> 5), läßt sich rasch auf folgende Weise ermitteln. Man multipliziert die kleinste Reihensumme mit der kleinsten Spaltensumme und dividiert das Produkt durch N (Gesamt-N). Das Ergebnis ist die kleinste erwartete Häufigkeit.

Die beschriebene Berechnung der erwarteten Häufigkeiten wird für jede Zelle nacheinander durchgeführt, so daß nach Abschluß der Rechnung in jeder Zelle zwei Werte stehen – die beobachtete und die erwartete Häufigkeit. Wir erhalten also:

		Stadt	Land	
	> Ø	*23.4* 15	*20.7* 29	44
Schulerfolg	Ø	*19.1* 25	*16.9* 11	36
	< Ø	*9.6* 12	*8.5* 6	18
		52	46	98

Zahlen im Kursivdruck = erwartete Häufigkeiten

Ob die Erwartungswerte richtig berechnet worden sind, kann man überprüfen, indem man ihre Randsummen bildet. Die Randsummen der Erwartungswerte und die Randsummen der beobachteten Werte müssen übereinstimmen, z.B. $25 + 11 = 36$ (beobachtete Werte); $19.1 + 16.9 = 36$ (erwartete Werte).

Chi^2 wird nun ermittelt, indem man in die Chi^2-Formel einsetzt:

$$Chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} =$$

$$= \frac{(15-23.4)^2}{23.4} + \frac{(25-19.1)^2}{19.1} + \frac{(12-9.6)^2}{9.6} + \frac{(29-20.7)^2}{20.7} + \frac{(11-16.9)^2}{16.9} + \frac{(6-8.5)^2}{8.5}$$

$$= \frac{8.4^2}{23.4} + \frac{5.9^2}{19.1} + \frac{2.4^2}{9.6} + \frac{8.3^2}{20.7} + \frac{5.9^2}{16.9} + \frac{2.5^2}{8.5}$$

$$= 3.02 + 1.82 + .60 + 3.33 + 2.06 + .74$$

$$= \underline{\underline{11.57}}$$

e) Signifikanzprüfung
Wie üblich vergleicht man den errechneten Wert mit dem in Tab. XI abgelesenen Wert. Für das Signifikanzniveau .05 lautet dieser Tabellenwert 5.991 bei $df = 2$. Der errechnete Wert übersteigt den abgelesenen Wert, die Nullhypothese kann zurückgewiesen werden. Stadt- und Landkinder unterscheiden sich also hinsichtlich des Schulerfolges.

(3) *Anwendungsregeln*

a) Die beobachteten Häufigkeiten werden in eine $r \times 2$-Tafel eingetragen, wobei r die Klassen des Merkmals bedeuten (in den Reihen). In den Spalten sind die 2 Stichproben einzutragen.
b) Die erwarteten Häufigkeiten werden über die Randsummen berechnet, indem man für jede Zelle bildet:

$$\frac{\text{Reihensumme} \times \text{Spaltensumme}}{\text{Gesamt-N}}$$

c) Chi^2 wird nach der Formel

$$Chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

berechnet.
d) Die Freiheitsgrade werden ermittelt: $df = (r - 1)$, wobei r = Zahl der Reihen.
e) Der kritische Chi^2-Wert wird unter Berücksichtigung des Signifikanzniveaus und der Freiheitsgrade in Tab. XI abgelesen. Ist der errechnete Wert kleiner, muß die Nullhypothese beibehalten werden, ist der Wert gleich oder größer als der Tabellenwert, kann H_0 zurückgewiesen werden.

(4) *Voraussetzungen und Einschränkungen*

a) Es gelten die unter 4a bis 4c beim einfachen Chi^2-Test gemachten Angaben (S. 211).
b) Nicht mehr als ein Fünftel der erwarteten Häufigkeiten darf < 5 sein, kein erwarteter Wert < 1 sein. Andernfalls muß durch Zusammenfassung von Kategorien (wenn inhaltlich vertretbar) eine Erhöhung der Erwartungswerte angestrebt werden.

C) Chi^2-Test zum Vergleich von mehreren ($k > 2$) Stichproben (Alternativmerkmal) – Fall: $2 \times k$

(1) *Untersuchungsproblem*

Unterscheiden sich *mehrere* Stichproben hinsichtlich ihrer Häufigkeitsverteilung bei einem *Alternativmerkmal?*
Es existieren also ein Alternativmerkmal (zweiklassiges Merkmal) und die Häufigkeitsverteilungen mehrerer Gruppen bezüglich dieses Merkmals.

(2) *Beispiel*

a) Versuchsplan
Es soll untersucht werden, ob sich unterschiedlich intelligente Kinder hinsichtlich der Lösung einer Testaufgabe unterscheiden. Es werden drei Zufallsstichproben von Kindern gebildet, die bei einem Intelligenztest überdurchschnittliche, unter-

durchschnittliche und durchschnittliche Ergebnisse erreicht hatten. Diesen drei Gruppen wird eine Testaufgabe zur Lösung vorgelegt.
Stichprobe $k = 1$: 52 überdurchschnittlich intelligente Kinder ($> \varnothing$)
Stichprobe $k = 2$: 28 durchschnittlich intelligente Kinder ($\varnothing$)
Stichprobe $k = 3$: 35 unterdurchschnittlich intelligente Kinder ($< \varnothing$)
Alternativmerkmal: Aufgabe gelöst (+); Aufgabe nicht gelöst (−).

b) Hypothesen

H_0: Unterschiedlich intelligente Kinder unterscheiden sich nicht hinsichtlich der Lösungshäufigkeit bei dieser Aufgabe.

H_1: Die Lösungshäufigkeit bei unterschiedlich intelligenten Kindern ist verschieden.

Signifikanzniveau: .05

c) Ergebnisse

		Intelligenz			
		$>\varnothing$	$\varnothing$	$<\varnothing$	
Aufgabe	+	35	12	7	54
	−	17	16	28	61
		52	28	35	115

$Chi^2 = 18.99$

d) Berechnung

Die Berechnung von Chi^2 für diese $2 \times k$-Tafel erfolgt in der gleichen Weise wie für die $r \times 2$-Tafel, und zwar nach der Formel:

$$Chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}; \quad df = (k-1)$$

(3) *Anwendungsregeln* und

(4) *Voraussetzungen und Einschränkungen* gelten analog Abschnitt B.

D) Chi²-Test für mehrere (k > 2) unabhängige Stichproben (mehrklassiges Merkmal) – Fall: r × k

(1) *Untersuchungsproblem*

Unterscheiden sich *mehrere* unabhängige Stichproben hinsichtlich ihrer Häufigkeitsverteilung bei einem *mehrklassigen* Merkmal?
Es ist dies die allgemeinste Form der Chi^2-Tafel.

(2) *Beispiel* (fiktiv)

a) Versuchsplan

Es soll untersucht werden, ob sich verschieden intelligente Kinder hinsichtlich ihres Schulerfolges unterscheiden.
Dazu werden wiederum drei Zufallsstichproben von Kindern unterschiedlicher Intelligenz gebildet und die Beurteilung ihres bisherigen Schulerfolges beim Lehrer eingeholt.

Stichprobe k = 1: 60 überdurchschnittlich intelligente Kinder (> ∅)
Stichprobe k = 2: 80 durchschnittlich intelligente Kinder (∅)
Stichprobe k = 3: 50 unterdurchschnittlich intelligente Kinder (< ∅)
Merkmalsklasse r = 1: überdurchschnittlicher Schulerfolg (> ∅)
Merkmalsklasse r = 2: durchschnittlicher Schulerfolg (∅)
Merkmalsklasse r = 3: unterdurchschnittlicher Schulerfolg (< ∅)

b) Hypothesen

H_0: Unterschiedlich intelligente Kinder unterscheiden sich nicht hinsichtlich ihres Schulerfolgs.

H_1: Unterschiedlich intelligente Kinder werden sich hinsichtlich ihres Schulerfolges unterscheiden.

Signifikanzniveau: .05

c) Ergebnisse

		Intelligenz			
		> ∅	∅	< ∅	
Schulerfolg	> ∅	31	17	11	59
	∅	20	53	22	95
	< ∅	9	10	17	36
		60	80	50	190

$Chi^2 = 30.53$

d) Berechnung

Die Berechnung von Chi^2 für solche $r \times k$-Tafeln (wobei $r \neq k$ sein kann) unterscheidet sich grundsätzlich nicht von der Berechnung des Chi^2 für die bisher besprochenen Tafeln.

Es gilt allgemein die Formel: $Chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$; $df = (r-1)(k-1)$

e) Signifikanzprüfung

Der errechnete Wert übersteigt den in Tab. XI bei df = 4 und Signifikanzniveau .05 abgelesenen. H_0 ist daher zurückzuweisen.

(3) *Anwendungsregeln* und

(4) *Voraussetzungen und Einschränkungen* gelten analog Abschnitt B.

(5) *Übungsbeispiel*

Nr. 5/21

30 Männern und 30 Frauen wurde die Frage vorgelegt: „Sollen Frauen studieren?" Ergebnis:

	Männer	Frauen
ja	9	15
nein	12	2
unentschieden	9	13

Prüfen Sie, ob zwischen Männern und Frauen ein Unterschied hinsichtlich der Beantwortung dieser Frage besteht.

5.4.1.3. Der Chi²-Test für die Güte der Anpassung (Prüfung auf Normalverteilung)

(1) *Untersuchungsproblem*

Mit einem Anpassungstest versucht man generell die Frage zu beantworten, ob die für die Klassen eines Merkmals empirisch ermittelten Häufigkeiten von den für die Klassen dieses Merkmals theoretisch erwarteten Häufigkeiten abweichen.
Für die Anwendung bestimmter Prüfmethoden (parametrische Tests) ist es u.a. wichtig, zu wissen, ob die empirisch ermittelten Werte normal verteilt sind oder nicht. Es muß also überprüft werden, ob die Häufigkeiten, die wir für die Klassen eines Merkmals empirisch gewonnen haben, von den Häufigkeiten signifikant abweichen, die wir erwarten, wenn die Daten normal verteilt sind. Es wird also eine empirische Verteilung der Normalverteilung gegenübergestellt.

(2) *Beispiel*

a) Versuchsplan (nach Clauß & Ebner 1971)
Im Rahmen einer empirischen Untersuchung ergab sich die unter Punkt c dargestellte Häufigkeitsverteilung für die verschiedenen Klassen des untersuchten Merkmals.
Stichprobe: 81 Schüler der 6. Klassen eines Gymnasiums
Merkmal: Leistung der Schüler in einem Rechentest (Punktwerte)

b) Hypothesen
H_0: Die ermittelten Punktwerte der Schüler sind normal verteilt.
H_1: Die empirische Verteilung der Punktwerte ist keine Normalverteilung.
Signifikanzniveau: .05

c) Ergebnisse (s. Spalte 7)

$M = 21.11$; $s = 3.84$; $N = 81$

(1) Klassen	(2) Exakte Klassengrenzen	(3) $z = \frac{X-M}{s}$	(4) f(z)	(5) Differenzen	(6) Differenzen $\times N = f_E$	(7) f_0	(8) $(f_0 - f_E)^2$	(9) $\frac{(f_0 - f_E)^2}{f_E}$
< 9					0*	0		
	8.5	−3.28	.0005					
9–11				.0057	.5 } 3.5	1 } 4	.25	.07
	11.5	−2.50	.0062					
12–14				.0365	3.0	3		
	14.5	−1.72	.0427					
15–17				.1309	10.6	10	.36	.03
	17.5	− .94	.1736					
18–20				.2628	21.3	18	10.89	.51
	20.5	− .16	.4364					
21–23				.2960	24.0	27	9.00	.38
	23.5	+ .62	.7324					
24–26				.1868	15.1	17	3.61	.24
	26.5	+1.40	.9192					
27–29				.0662	5.4	5	2.56	.39
	29.5	+2.18	.9854		} 6.6	} 5		
> 29				.0146**	1.2	0		
Σ					81.1	81		$Chi^2 = 1.62$

* $.0005 \cdot 81 = 0$
** .0146 folgt aus $1.0 - .9854$

d) Berechnung

Der Vergleich einer empirischen Verteilung mit der Normalverteilung kann mit der Chi^2-Prüfmethode durchgeführt werden. Der Berechnungsvorgang läßt sich in zwei Arbeitsabschnitte gliedern:

Die Berechnung der erwarteten Häufigkeiten.

Der Vergleich der erwarteten mit den beobachteten Häufigkeiten mit Hilfe der Chi^2-Methode.

Berechnung der erwarteten Häufigkeiten:

- Die Werte werden in Klassen von gleicher Intervallgröße zusammengefaßt (Spalte 1); Mittelwert (M) und Standardabweichung (s) werden berechnet.
- Die exakten Klassengrenzen (es genügen die Klassenenden) werden aufgeschrieben (Spalte 2). Beispielsweise schreibt sich
  27–29 exakt: 26.5–29.5,
  30–32 exakt: 29.5–32.5.
  Das Klassenende entspricht also immer dem Klassenanfang der nächsthöheren Klasse.
- Die erwarteten Häufigkeiten für jede Klasse werden wie folgt berechnet:

a) In die Formel $z = \frac{X - M}{s}$ werden für X nacheinander alle Klassenenden eingesetzt und z errechnet (Spalte 3).

b) Für jedes z wird in Tab. II der Häufigkeitswert in der Normalverteilung abgelesen (Spalte 4).
Z.B. für $z = -1.72$ ergibt sich .0427.

c) Es werden nun jeweils die Differenzen zweier aufeinanderfolgender Häufigkeitsangaben (aus Spalte 4) gebildet, also jeweils die Differenzen zwischen Klassenende und Klassenanfang,
z.B.: $.0062 - .0005 = .0057$.
Diese Differenzen werden so in Spalte 5 geschrieben, daß sie in der gleichen Reihe mit der Klassenbezeichnung stehen.

d) Die Differenzen werden mit N (N = Stichprobengröße) multipliziert und in Spalte 6 eingetragen.
Es sind dies die erwarteten Häufigkeiten für jede Klasse. Sind die erwarteten Häufigkeiten für eine Klasse zu klein (< 5), dann werden für die weitere Berechnung jeweils zwei Klassen zusammengefaßt.

Vergleich der erwarteten mit den beobachteten Häufigkeiten mit Hilfe der Chi^2-Methode:

- In Spalte 7 werden die beobachteten Häufigkeiten eingetragen.
- Das Quadrat aus der Differenz von beobachteten und erwarteten Häufigkeiten wird pro Reihe gebildet und in Spalte 8 eingetragen.
- Die Zahlen der Spalte 8 werden für jede Reihe durch die zugehörigen Erwartungswerte dividiert (Spalte 9).
- Die Zahlen der Spalte 9 werden addiert: Die Summe ist Chi^2.

e) Signifikanzprüfung

In Tab. XI wird bei $df = k - 3$ (k = Anzahl der Klassen) die Überschreitungswahrscheinlichkeit für den errechneten Chi^2-Wert abgelesen. Ist diese größer als .05, dann ist die Abweichung der empirisch ermittelten von der erwarteten Häufigkeits-

verteilung nicht signifikant. H_0 muß beibehalten werden, die Abweichungen der empirischen Verteilung von der Normalverteilung liegen noch im Zufallsbereich. Ist die für Chi^2 abgelesene Überschreitungswahrscheinlichkeit kleiner als .05, dann ist H_0 zurückzuweisen, die empirische Verteilung ist keine Normalverteilung. Als Maß für die Güte der Anpassung der theoretischen Verteilung an die empirische Verteilung gibt Lienert (1962) folgende Werte an.
Überschreitungswahrscheinlichkeit für Chi^2:

	$p \geqq .50$	$.50 \geqq p \geqq .20$	$.20 \geqq p \geqq .05$	$.05 \geqq p \geqq .001$
Anpassung:	gut	mäßig	schwach	fehlend

(4) *Voraussetzungen und Einschränkungen*

Für die Durchführung des Chi^2-Tests gelten die unter Punkt 3, Abschnitt A und B gemachten Angaben.

5.4.2. Kullbacks 2 Î-Test als Chi^2-Alternative

(1) *Untersuchungsproblem*

Der 2 Î-Test von Kullback (Informationsanalyse) kann zur Lösung der in den Kapiteln 5.4.1.1 bis 5.4.1.3, 4.1.7 und 4.1.8 behandelten Problemstellungen eingesetzt werden.
Es handelt sich dabei
a) um den Vergleich einer empirischen Häufigkeitsverteilung mit einer theoretischen (Gleichverteilung bzw. Normalverteilung),
b) um die Prüfung der Differenz von Häufigkeiten aus zwei oder mehreren Stichproben,
c) um die Ermittlung des Zusammenhangs zwischen zwei Merkmalen, die an einer Gruppe von Merkmalsträgern gemessen wurden (Kontingenz).
In den genannten Kapiteln wurde die Bearbeitung der angeführten Fragestellungen mit Hilfe der Berechnung von Chi^2 dargestellt. Bei Mehrfeldertafeln, in denen k und/oder $r > 2$ sind, wird die Berechnung von Chi^2 nach der Formel

$$Chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

recht aufwendig. Hinzu kommt, daß Chi^2 nicht berechnet werden kann, wenn eine größere Zahl von Erwartungswerten < 5 vorliegt. Beide Schwierigkeiten lassen sich mit dem 2 Î-Test von Kullback umgehen, bei dem die Berechnung der entsprechenden Prüfgröße ökonomischer ist als die Bestimmung von Chi^2 und der auch für Tafeln mit niedrigen Erwartungswerten verwendet werden kann.
Der 2 Î-Test soll hier nur an dem Beispiel einer $2 \times k$-Feldertafel erläutert werden. Ausführliche Darstellungen finden sich bei Kullback (1959) und Blöschl (1966).

(2) *Beispiel* (nach Blöschl 1966)

a) Versuchsplan
146 Studenten und 54 Studentinnen sollten den Grad ihres Interesses an weltanschaulichen Fragen auf einer sechsstufigen Skala einstufen (1 = sehr interessiert, 6 = gleichgültig).
Stichprobe: 146 Studenten und 54 Studentinnen; N = 200
Merkmal: Interesse an weltanschaulichen Fragen

b) Hypothesen

H_0: Zwischen Studenten und Studentinnen besteht kein Unterschied hinsichtlich ihres Interesses an weltanschaulichen Fragen.

H_1: Studenten und Studentinnen unterscheiden sich hinsichtlich ihres Interesses an weltanschaulichen Fragen.

Signifikanzniveau: .05

c) Ergebnisse

Grad des Interesses

		1	2	3	4	5	6	N
Geschlecht	männl.	12	66	50	17	1	0	146
	weibl.	4	22	22	5	0	1	54
		16	88	72	22	1	1	200

d) Berechnung

Es ist folgende Prüfgröße zu berechnen:

$$2\hat{I} = \left(\sum_{i=1}^{r}\sum_{j=1}^{k} 2f_{ij}\ln f_{ij} + 2N\ln N\right) - \left(\sum_{i=1}^{r} 2f_i \ln f_i + \sum_{j=1}^{k} 2f_j \ln f_j\right)$$

$$2\hat{I} = \qquad K_1 \qquad - \qquad K_2$$

Es bedeuten: f_{ij} = die Häufigkeit in der i-ten Reihe und der j-ten Spalte (z.B. $f_{11} = 12$; $f_{24} = 5$)

ln = natürlicher Logarithmus

N = Gesamt-N

K = Klammerausdruck

Diese Prüfgröße gilt für 2×2-, $2 \times k$-, $r \times 2$- und $r \times k$-Tafeln.

Für die Funktion 2flnf liegen Tabellen vor für f = 1 bis 10000 (Kullback, Kupperman & Ku 1962) bzw. für f = 1 bis 5000 bei Blöschl (1966).

Es wird nun der erste Klammerausdruck (K_1) berechnet.

Für jeden Häufigkeitswert jeder Zelle der Mehrfeldertafel wird der entsprechende 2flnf-Wert in der Tabelle abgelesen. Diese Werte werden sodann addiert.

Häufigkeit (f_{ij})	2flnf
12	59.638
66	553.034
50	391.202
17	96.329
1	0.000
0	0.000
4	11.090
22	136.006
22	136.006
5	16.094
0	0.000
1	0.000

$\sum_{i=1}^{r} \sum_{j=1}^{k} 2 f_{ij} \ln f_{ij}$	$= 1409.399$
Für $2 N \ln N = 200$ wird abgelesen	$\underline{2119.327}$
$\left(\sum_{i=1}^{r} \sum_{j=1}^{k} 2 f_{ij} \ln f_{ij} + 2 N \ln N \right)$	$= 3518.726 = K_1$

Der zweite Klammerausdruck wird errechnet, indem man für alle r- und alle k-Randsummen die Tabellenwerte abliest und addiert.

Randsummen	
146	1455.213
54	430.810
$\sum_{i=1}^{r} 2 f_i \ln f_i$	$= 1886.023$
16	88.723
88	788.011
72	615.840
22	136.006
1	0.000
1	0.000
$\sum_{j=1}^{k} 2 f_j \ln f_j$	$= \underline{1628.580}$
$\left(\sum_{i=1}^{r} 2 f_i \ln f_i + \sum_{j=1}^{k} 2 f_j \ln f_j \right)$	$= 3514.603 = K_2$

Es müssen für die Auswertung einer Mehrfeldertafel jeweils insgesamt $(r + 1)(k + 1)$ Tabellenwerte abgelesen werden.
Der Wert für $2\hat{I}$ ergibt sich nun durch Subtraktion des zweiten vom ersten Klammerausdruck:

$$2\hat{I} = K_1 - K_2$$
$$2\hat{I} = 3518.726 - 3514.603$$
$$2\hat{I} = 4.123$$

e) Signifikanzprüfung
Die Signifikanzprüfung erfolgt über die Chi-Quadrat-Verteilung (Tab. XI). Die Zahl der Freiheitsgrade beträgt $(r - 1)(k - 1)$, im Beispiel $(2 - 1)(6 - 1) = 5$. Der errechnete $2\hat{I}$-Wert ist kleiner als der bei Signifikanzniveau .05 und $df = 5$ in der Tabelle abgelesene. Die Nullhypothese ist daher beizubehalten.

(3) *Anwendungsregeln*

a) Die Ergebnisse sind wie üblich in eine Mehrfeldertafel einzutragen.
b) Für jeden Häufigkeitswert jeder Zelle der Mehrfeldertafel und für das Gesamt-N werden die Tabellenwerte (Tab. XII) abgelesen und addiert. Man erhält den Betrag für den ersten Klammerausdruck K_1.
c) Für jede Randsumme wird der Tabellenwert abgelesen, die Werte werden addiert, und es ergibt sich der Betrag für den zweiten Klammerausdruck K_2.

d) Den Wert für $2\hat{I}$ erhält man durch Subtraktion des zweiten vom ersten Klammerausdruck:

$$2\hat{I} = K_1 - K_2$$

wobei $K_1 = \left(\sum_{i=1}^{r} \sum_{j=1}^{k} 2 f_{ij} \ln f_{ij} + 2 N \ln N \right)$

$$K_2 = \left(\sum_{i=1}^{r} 2 f_i \ln f_i + \sum_{j=1}^{k} 2 f_j \ln f_j \right)$$

e) Die Signifikanzprüfung erfolgt über die Chi-Quadrat-Verteilung (Tab. XI). Die Zahl der Freiheitsgrade beträgt $(r-1)(k-1)$. Ist der errechnete $2\hat{I}$-Wert gleich oder größer als der unter Berücksichtigung des Signifikanzniveaus und der Freiheitsgrade abgelesene, kann H_0 zurückgewiesen werden; ist er kleiner, muß H_0 beibehalten werden.

(4) *Voraussetzungen und Einschränkungen*

a) Der $2\hat{I}$-Test kann auch bei Erwartungswerten < 5 verwendet werden.

b) Bei Verwendung des Verfahrens von Kullback für die Berechnung des Zusammenhanges zwischen zwei Merkmalen (Kontingenz) gelten folgende Beziehungen:

$$\text{Phi} = \sqrt{\frac{2\hat{I}}{N}}$$

$$C = \sqrt{\frac{2\hat{I}}{N + 2\hat{I}}}$$

(5) *Übungsbeispiel* (nach Blöschl 1966)

Nr. 5/22

Es soll geprüft werden, ob Werkstudenten sich in ihren Examensergebnissen von Studenten unterscheiden, die während des Studiums nicht arbeiten.

Folgende Ergebnisse liegen vor:

	Examenserfolg gut	Examenserfolg schlecht	
Werkstudenten	30	43	73
Keine Werkstudenten	70	41	111
	100	84	184

Verfahren für Rangdaten

Im Gegensatz zu den bisher geschilderten Verfahren setzen die folgenden Methoden Daten mindestens auf Rangskalenniveau voraus.

5.4.3. Der U-Test von Mann-Whitney zum Vergleich zweier unabhängiger Stichproben

(1) *Untersuchungsproblem*

Unterscheiden sich zwei unabhängige Stichproben hinsichtlich eines auf Rangskalenniveau gemessenen Merkmals?

5.4.3.1. U-Test für Stichproben $N_1 < 9$ und $N_2 < 9$

(Die größere Stichprobe wird immer mit N_2 bezeichnet.)

(2) *Beispiel*

a) Versuchsplan

Beeinflußt Lärm die Leistung von Schülern bei einem Rechentest? Es werden zwei Zufallsstichproben von 5 bzw. 4 Schülern gebildet, die einem Rechentest unterzogen werden, wobei die erste Gruppe unter Lärmbedingungen, die zweite unter Normalbedingungen arbeitet.
Stichprobe 1: 5 Schüler unter Lärmbedingungen (Experimentalgruppe = E)
Stichprobe 2: 4 Schüler unter Normalbedingungen (Kontrollgruppe = K)
Merkmal: Leistungen im Rechentest

b) Hypothesen

H_0: Die Schüler in der Experimentalgruppe werden nur zufällig verschieden hohe Punktzahlen im Rechentest erzielen im Vergleich zu den Schülern der Kontrollgruppe.

H_1: Die Schüler unter Normalbedingungen werden bessere Leistungen erzielen als die Schüler in der Experimentalgruppe.

Signifikanzniveau: .05 (einseitiger Test)

c) Ergebnisse

Die im Rechentest erzielten Punkte werden in die folgende Tabelle für beide Gruppen eingetragen:

Experimentalgruppe (E)	48	31	51	49	43	$N_2 = 5$
Kontrollgruppe (K)	42	35	45	52		$N_1 = 4$

d) Berechnung

Zur Anwendung des U-Tests ordnen wir die Werte *beider* Stichproben ihrer algebraischen Größe nach:

31	35	42	43	45	48	49	51	52
E	K	K	E	K	E	E	E	K

Der U-Wert berechnet sich wie folgt:
Wir zählen aus, wie oft ein E-Wert vor einem K-Wert liegt.
Bei der Anordnung KKKKEEEEE wäre U = 0, denn kein E-Wert liegt vor einem K-Wert.
Im Falle der Anordnung EEEEEKKKK wäre U = 20, denn vor jedem K stehen

5 E. Bei der Anordnung KKKEKEEEE ist $U = 1$, denn nur ein E-Wert steht vor einem K-Wert.
In unserem Beispiel errechnet sich:

$$U = 1 + 1 + 2 + 5 = 9$$

Wir könnten auch auszählen, wie viele K vor einem E stehen und erhalten dann:

$$U' = 1 + 1 + 3 + 3 + 3 = 11$$

Zwischen U und U' besteht der Zusammenhang:

$$\begin{aligned} U + U' &= N_1 N_2 \\ U &= N_1 N_2 - U' \\ U &= 4 \cdot 5 - 11 = 9 \\ U' &= N_1 N_2 - U \\ U' &= 4 \cdot 5 - 9 = 11 \end{aligned}$$

e) Signifikanzprüfung
Der Zusammenhang zwischen U und U' ist für die Überprüfung der Signifikanz von Bedeutung.
Haben wir nämlich den U-Wert in der geschilderten Weise berechnet, dann schlagen wir in Tab. XIIIa nach. Wir finden dort im Kopf der kleinen Tabellen jeweils die Stichprobengrößen der größeren Stichprobe N_2 und seitlich die U-Werte.
In unserem Beispiel lesen wir in der Tabelle mit der Überschrift $N_2 = 5$, in der Spalte $N_1 = 4$ und in der Reihe $U = 9$ die Überschreitungswahrscheinlichkeit von $p = .452$ ab. Da .452 größer ist als unser Signifikanzniveau .05, können wir die Nullhypothese nicht zurückweisen. Die Leistungen von Schülern unter der Lärmbedingung sind also nicht signifikant von denen unter Normalbedingungen verschieden.
Es ist zu beachten, daß die Tafeln für einseitige Tests gelten; bei zweiseitigem Test sind die Werte in Tab. XIIIa zu verdoppeln. Hätten wir nun anstelle des U-Wertes den Wert $U' = 11$ errechnet, indem wir alle K vor einem E ausgezählt hätten, würden wir in Tab. XIIIa keinen solchen Wert vorfinden. Wir können in einem solchen Fall nach der oben angegebenen Beziehung leicht den zugehörigen U-Wert errechnen. Für die Signifikanzprüfung darf nämlich immer nur der *kleinere Wert* (U oder U') herangezogen werden.
Es ist übrigens immer ratsam, beide Werte zu ermitteln. Zum einen, weil immer, wie gesagt, der kleinere beider Werte zur Signifikanzprüfung herangezogen wird, zum anderen, weil man damit überprüfen kann, ob der U-Wert richtig berechnet wurde.

(3) *Anwendungsregeln*

a) Die Werte beider Stichproben werden ihrer algebraischen Größe nach in eine Rangreihe gebracht.
b) Der U-Wert wird ausgezählt. Ist der resultierende Wert $> N_1 N_2/2$, berechnet man $U' = N_1 N_2 - U$.
(Der U-Wert kann auch nach der unter B angegebenen Methode berechnet werden. Für kleine Stichproben ist die Auszählmethode jedoch ökonomischer.)
c) Die Überschreitungswahrscheinlichkeit für den *kleineren* U-Wert (oder U') wird in Tab. XIIIa nachgeschlagen; bei zweiseitigem Test muß dieser Wert verdoppelt werden.
d) Ist die gefundene Überschreitungswahrscheinlichkeit gleich oder kleiner als unser Signifikanzniveau, ist die Nullhypothese zurückzuweisen.

5.4.3.2. U-Test für Stichproben, bei denen $9 \leqq N_2 \leqq 20$ ist

Das oben beschriebene Verfahren ist bei größeren Stichproben recht aufwendig. Der U-Wert wird dann auf eine andere Weise ermittelt, die aber zu identischen Resultaten führt.

Dazu werden die Ergebnisse folgendermaßen geordnet (wir benutzen als Beispiel die gleichen Werte wie oben):

Experimental-gruppe	Rang	Kontroll-gruppe	Rang
48	6	42	3
31	1	35	2
51	8	45	5
49	7		9
43	4		

$R_2 = 26$ $\qquad\qquad\qquad$ $R_1 = 19$

Auch hier werden die Daten so behandelt, als ob sie aus einer Stichprobe $(N_1 + N_2)$ stammten. Jedem Wert wird eine Rangzahl zugeordnet. Der niedrigste Wert aus beiden Stichproben (31) erhält den Rang 1, der nächsthöhere (35) den Rang 2 und so fort. Sollten einmal mehrere gleich große Werte (sog. *ties*) auftreten, dann geht man wie folgt vor:

Beispiel: 15 20 22 22 22 24 25 25 27 (Meßwerte)
Ränge: 1 2 4 4 4 6 7.5 7.5 9

$$\frac{3+4+5}{3} = 4 \qquad \frac{7+8}{2} = 7.5$$

D.h. also, bei mehreren gleichen Werten erhalten alle diese Werte den mittleren Rang; der nächstfolgende Wert erhält den Rang, den man gegeben hätte, wenn man in der Rangreihe einfach weitergezählt hätte. Bei einer ungeraden Zahl von gleichen Werten kann man sich eine Division ersparen, alle Werte erhalten den in der Mitte liegenden Rang zugeordnet (im Beispiel Rang 4).

Zur Kontrolle, ob man die Rangplätze richtig zugeordnet hat, kann man folgende Gleichung benutzen:

$$R_1 + R_2 = \frac{N(N+1)}{2}; \quad (N = N_1 + N_2)$$

D.h., die Summe der Rangplätze ist gleich der Summe der Zahlen von 1 bis N.

$$26 + 19 = \frac{9(9+1)}{2}$$

$$45 = 45$$

Der U-Wert wird nun nach folgenden Formeln berechnet:

$$U = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1 \quad \text{oder}$$

$$U = N_1 N_2 + \frac{N_2(N_2+1)}{2} - R_2$$

R_1, R_2 = Summe der Ränge in der ersten bzw. zweiten Stichprobe.

Beide Formeln ergeben verschiedene U-Werte.

Nach der ersten Formel ist $U = 4 \cdot 5 + \frac{4(4+1)}{2} - 19 = 11$

Nach der zweiten Formel ist $U = 4 \cdot 5 + \frac{5(5+1)}{2} - 26 = 9$

Um festzustellen, ob der errechnete U-Wert der kleinere ist, kann man entweder den U-Wert nach der jeweils anderen Formel berechnen oder die Beziehung $U = N_1 N_2 - U'$ benutzen.

Für die *Signifikanzprüfung* wird auch hier der kleinere U-Wert benutzt. Dazu dient jetzt Tab. XIIIb. Hier finden sich die kritischen U-Werte für ein- und zweiseitige Tests.

Ist der errechnete U-Wert gleich oder *kleiner* als der in der Tabelle gefundene, dann ist die Nullhypothese auf dem entsprechenden Signifikanzniveau zurückzuweisen.

(3) *Anwendungsregeln*

a) Die Daten werden für jede Stichprobe getrennt in eine Tabelle eingetragen.
b) Jedem Meßwert wird ein Rang zugeordnet, wobei der niedrigste Meßwert den Rangplatz 1 erhält, der höchste den Rangplatz $(N_1 + N_2)$, da beide Stichproben kombiniert werden. Gleiche Meßwerte erhalten mittlere Rangplätze.
c) Der U-Wert wird nach den angegebenen Formeln berechnet.
d) In Tab. XIIIb wird geprüft, ob der kleinere U-Wert den kritischen Wert übersteigt. Ist der errechnete Wert gleich oder *kleiner* als der Tafelwert, kann H_0 zurückgewiesen werden.

5.4.3.3. U-Test für große Stichproben ($N_2 > 20$)

Bei zunehmender Stichprobengröße nähert sich die U-Verteilung der Normalverteilung. Die Signifikanz eines errechneten U-Wertes kann dann über z geprüft werden.

$$z = \frac{U - \frac{N_1 N_2}{2}}{\sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}}$$

Der U-Wert wird nach der unter B beschriebenen Methode berechnet, wobei es hier gleichgültig ist, ob man den kleineren oder größeren U-Wert errechnet und in die Formel für z einsetzt. Es ändert sich dadurch nur das Vorzeichen von z, nicht aber sein Absolutbetrag, den man zur Signifikanzprüfung heranzieht.

(3) *Anwendungsregeln*

a) Die Anordnung der Daten, die Zuteilung der Rangplätze und die Berechnung von U erfolgt in der gleichen Weise wie in Abschnitt B erläutert.
b) Der z-Wert wird nach der angegebenen Formel berechnet.
c) In Tab. II wird die dem errechneten z-Wert zugehörige Auftretenswahrscheinlichkeit abgelesen. Ist sie gleich oder kleiner als unser Signifikanzniveau, dann ist H_0 zurückzuweisen.

(4) *Voraussetzungen und Einschränkungen*

a) Die Werte müssen mindestens Rangskalenniveau haben oder in Rangdaten umgewandelt worden sein.
b) Treten sehr viele gleiche Werte auf *(ties)*, dann kann eine Korrekturformel benutzt werden (Siegel 1956, S. 123ff.). Sie führt allerdings, selbst bei einer großen Zahl von *ties*, nur zu einer geringfügigen Erhöhung des z-Wertes. Siegel empfiehlt, die Korrektur nur dann durchzuführen, wenn die Zahl der *ties* sehr groß ist und die für z gefundene Auftretenswahrscheinlichkeit nahe an dem von uns gewählten Signifikanzniveau liegt.
c) Es ist anzumerken, daß der U-Test einer der effizientesten nichtparametrischen Tests ist (Effizienz 95%).

(5) *Übungsbeispiele*

Nr. 5/23
In einer psychologischen Untersuchung ergaben sich folgende Punktwerte:

Experimentalgruppe:	8	3	4	6	
Kontrollgruppe:	1	7	9	10	12

Besteht ein Unterschied zwischen Experimental- und Kontrollgruppe?

Nr. 5/24
Zwei Gruppen von Schülern wurden Mathematikaufgaben vorgelegt, deren Lösung selbstständiges mathematisches Denken erforderte. Die Gruppe 1 (N = 13) bestand aus Schülern, die für den Besuch der höheren Schule vorgeschlagen waren, die Gruppe 2 (N = 24) aus Schülern, die nicht für den Besuch der höheren Schule geeignet erschienen. Es ist die Hypothese zu prüfen: Die Schüler aus Gruppe 1 werden bessere Leistungen erbringen als die Schüler der Gruppe 2.
Es wurden folgende Punktwerte erreicht:

Gruppe 1:	32	14	10	42	30	10	48	36	57	7	36	28	31					
Gruppe 2:	7	27	18	22	40	42	18	7	25	19	6	13	15	22	4	7	36	5
	19	7	31	25	0	8												

5.4.4. Der H-Test von Kruskal-Wallis zum Vergleich mehrerer (k) unabhängiger Stichproben (Rangvarianzanalyse)

(1) *Untersuchungsproblem*

Unterscheiden sich mehrere (k) unabhängige Stichproben hinsichtlich eines auf Rangskalenniveau gemessenen Merkmals?

(2) *Beispiel* (fiktiv)

a) Versuchsplan
Es soll untersucht werden, ob sich Hauptschul-, Realschul- und Gymnasiallehrer hinsichtlich ihrer autoritären Einstellung unterscheiden.
Dazu wurden aus diesen drei Lehrergruppen drei Zufallsstichproben gezogen. Den Lehrern wurde die F-Skala vorgelegt.

b) Hypothesen
H_0: Es besteht kein Unterschied hinsichtlich des Autoritarismus zwischen Hauptschul-, Realschul- und Gymnasiallehrern.

H_1: Hauptschul-, Realschul- und Gymnasiallehrer haben eine unterschiedlich ausgeprägte autoritäre Einstellung.

Signifikanzniveau: .05

c) Ergebnisse

Hauptschul-lehrer	Ränge	Realschul-lehrer	Ränge	Gymnasial-lehrer	Ränge
96	4	82	2	115	7
128	9	124	8	149	13
83	3	132	10	166	14
61	1	135	11	147	12
101	5	109	6		

$R_1 = 22$ $\qquad R_2 = 37$ $\qquad R_3 = 46$

$$\sum_{j=1}^{k} R_j = 105$$

d) Berechnung

Man geht zunächst vor wie beim U-Test und behandelt die Werte aller k Stichproben, als ob sie aus einer gemeinsamen großen Stichprobe ($k = 1 + k = 2 + \ldots + k = j$) stammten. Allen Werten werden ihrer algebraischen Größe nach Ränge zugeteilt, der niedrigste Wert aus allen Stichproben erhält den Rang 1, der höchste Wert den Rang N, wobei N die Gesamtzahl der unabhängigen Beobachtungen in den k (hier: $k = 3$) Stichproben ist. Gleiche Meßwerte erhalten mittlere Ränge, wie beim U-Test. Nach Zuordnung der Ränge wird die Summe aller Ränge pro Stichprobe gebildet (R_1, R_2 bis R_k).

Zur Kontrolle, ob die Rangzahlen richtig zugeordnet wurden, berechnen wir:

$$\sum_{j=1}^{k} R_j = \frac{N(N+1)}{2}$$

Wenn auf beiden Seiten der Gleichung der gleiche Betrag steht, sind die Ränge korrekt zugeordnet worden.

Für unser Beispiel:

$$22 + 37 + 46 = \frac{14(14+1)}{2}$$

$$105 = 105$$

Die Ränge wurden also richtig zugeordnet.

Es wird nun der Wert H berechnet:

$$H = \frac{12}{N(N+1)} \left(\sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1)$$

k = Zahl der Stichproben

N = Zahl der Vpn in allen k Stichproben (Gesamt-N)

n_j = Zahl der Vpn in der j-ten Stichprobe, wobei j von 1 bis k läuft

R_j = Rangsumme in der j-ten Stichprobe, wobei j von 1 bis k läuft

Wir setzen unsere Werte in die Formel ein und erhalten:

$$H = \frac{12}{14(14+1)} \left(\frac{22^2}{5} + \frac{37^2}{5} + \frac{46^2}{4} \right) - 3(14+1) = 6.4$$

e) Signifikanzprüfung

Zur Signifikanzprüfung gibt es nun *zwei Methoden:*

aa) Die erste Methode (die für unser Beispiel gilt) ist anzuwenden, wenn die Zahl der Stichproben k = 3 ist und die Zahl der Vpn in jeder Stichprobe 5 oder kleiner ist.
In diesem Fall bedienen wir uns der Tafel XIV.
In der linken Spalte sind die Stichprobengrößen in verschiedenen Kombinationen angegeben. In der zweiten Spalte finden sich die H-Werte. Die dritte Spalte enthält die Auftretenswahrscheinlichkeiten von H.
Für unser Beispiel lesen wir ab bei $n_1, n_2, n_3 = 5, 5, 4$, daß ein $H \geqq 6.4$ eine Auftretenswahrscheinlichkeit von $p < .049$ hat.
Da dieser Wert kleiner ist als .05, können wir die Nullhypothese zurückweisen. Die Lehrer aus den verschiedenen Schultypen unterscheiden sich also hinsichtlich ihrer autoritären Einstellung.

bb) Bei mehr als drei Stichproben und bei größeren Vpn-Zahlen in den Stichproben ist der H-Wert verteilt wie Chi-Quadrat mit $df = k - 1$. Die Signifikanz kann dann in der gewohnten Weise über die Tab. XI der kritischen Werte für Chi^2 geprüft werden, unter Beachtung der Freiheitsgrade $(k - 1)$.

(3) *Anwendungsregeln*

a) Die Werte sind für jede der k Gruppen in einer Tabelle zusammenzufassen.
b) Den Werten werden Ränge von 1 bis N zugeordnet, und zwar über alle Stichproben. Der kleinste Wert erhält Rang 1, der höchste Rang N. Die Rangsummen für jede Stichprobe werden gebildet.
c) Der H-Wert wird nach der genannten Formel berechnet.
d) Die Signifikanzprüfung hängt ab von der Zahl und der Größe der Stichproben: Bei $k = 3$ und $n_1, n_2, n_3 \leqq 5$ wird Tab. XIV verwendet.
Ansonsten ist Tab. XI heranzuziehen, wobei $df = k - 1$ ist.

(4) *Voraussetzungen und Einschränkungen*

a) Die Meßwerte müssen stetig verteilt sein und mindestens Rangskalenniveau haben.
b) Treten sehr viele gleich große Werte auf, dann sollte H korrigiert werden (Korrekturformel s. Siegel 1956, S. 188). Allerdings weist Siegel darauf hin, daß die Wirkung der Korrektur nur sehr gering ist. Selbst bei sehr vielen *ties* ergibt sich nur eine geringfügige Verkleinerung des H-Wertes.
c) Anmerkung zur Effizienz: Im Vergleich zum F-Test hat der H-Test eine Effizienz von 95,5%.

(5) *Übungsbeispiel*

Nr. 5/25

Drei Gruppen von je 12 zufällig ausgewählten Schülern gleichen Geschlechts und Alters mußten eine Probearbeit absolvieren. Besteht ein Unterschied zwischen den Leistungen der drei Gruppen?

Punktwerte:

Gruppe 1:	6	11	12	20	24	21	18	15	14	10	8	14
Gruppe 2:	31	7	9	11	16	19	17	11	22	23	27	26
Gruppe 3:	13	32	31	30	28	29	25	26	26	27	26	19

5.4.5. Der McNemar-Test zum Vergleich zweier abhängiger Stichproben

(1) *Untersuchungsproblem*

Unterscheidet sich die Häufigkeitsverteilung eines alternativen Merkmals bei einer Gruppe von Versuchspersonen unter der Bedingung a von der Häufigkeitsverteilung unter der Bedingung b?
In der Hauptsache kann man mit dieser Methode Untersuchungen analysieren, die das Problem der Veränderung von Leistungen, Einstellungen, Bevorzugungen usw. nach Setzung einer bestimmten Bedingung zum Gegenstand haben.
Da von den Vpn je zwei Meßwerte erhoben werden (unter Bedingung a und unter Bedingung b), liegt der Fall abhängiger Stichproben vor.

(2) *Beispiel* (fiktiv)

a) Versuchsplan
Es soll untersucht werden, ob der Unterricht über die Probleme ethnischer Minderheiten zu einer Änderung der Einstellung der Schüler solchen Minderheiten gegenüber führt.
Dazu wird mit Hilfe eines Fragebogens festgestellt, wieviele Schüler einer Klasse eine positive bzw. negative Einstellung gegenüber ethnischen Minderheiten haben (vor dem Unterricht). Nach dem Unterricht wird von denselben Schülern nochmals die Zahl der positiven und negativen Einstellungen erhoben.

Stichprobe (abhängig):	26 Schüler einer Klasse; 1. Messung vor, 2. Messung nach dem Unterricht
Alternativmerkmal:	Einstellung zu ethn. Minderh.: positiv/negativ

b) Hypothesen
H_0: Die Zahl der Schüler, die ihre Einstellung nach dem Unterricht in positive Richtung geändert haben, ist nur zufällig verschieden von der Zahl der Schüler, die ihre Einstellung in die negative Richtung geändert haben.
H_1: Nach dem Unterricht werden mehr Schüler ihre Einstellung in die positive Richtung geändert haben als in die negative.
Signifikanzniveau: .05 (einseitiger Test)

c) Ergebnisse
Die Ergebnisse werden in eine Vierfeldertafel eingetragen:

		Einstellung nach dem Unterricht −	+
Einstellung vor dem Unterricht	+	A 5	B 3
	−	C 2	D 16

Wie der Tabelle zu entnehmen ist, hatten 5 Schüler vor dem Unterricht eine positive, nach dem Unterricht eine negative Einstellung (A). 16 Schüler (D) hatten vorher eine negative, nachher eine positive Einstellung. Bei 5 Kindern (3 in B und

2 in C) blieb die Einstellung unverändert. Schüler, die ihre Einstellung verändert haben, sind also in den Feldern A und D zu finden, und zwar in A Veränderung vom Positiven zum Negativen, in D vom Negativen zum Positiven. Geprüft werden muß nun, ob signifikant mehr Schüler ihre Einstellung in die positive Richtung änderten als in die negative bzw. umgekehrt.

d) Berechnung

Für die Analyse derartiger Daten sind nur die Versuchspersonen von Bedeutung, die sich von der ersten zur zweiten Bedingung „verändert" haben, also die, welche in den Feldern A und D stehen. Entsprechend der Nullhypothese würde man erwarten, daß die eine Hälfte der sich „ändernden" Vpn, nämlich $[\frac{1}{2}(A + D)]$, ihre Einstellung in die eine (positive) Richtung, die andere Hälfte $[\frac{1}{2}(A + D)]$ in die andere (negative) Richtung ändert. Die erwarteten Häufigkeiten für die Zellen A und D sind also:

$$E = \frac{1}{2}(A + D)$$

Es darf nur dann weitergerechnet werden, wenn die erwarteten Häufigkeiten für die Zellen A und D > 5 sind.

Im Beispiel:

$$E = \frac{1}{2}(5 + 16) = 10.5$$

Ist die Bedingung erfüllt, wird Chi^2 wie folgt berechnet:

$$\text{Chi}^2 = \frac{(A - D)^2}{A + D}; \quad df = 1$$

$$\text{Chi}^2 = \frac{(5 - 16)^2}{5 + 16} = 5.75$$

e) Signifikanzprüfung

Zur Signifikanzprüfung benutzen wir Tab. XI. Wir lesen ab, daß ein Chi^2-Wert von 5.75 bei $df = 1$ eine Auftretenswahrscheinlichkeit von $p < 1/2$ (.02) hat, was .01 ergibt. (Die Wahrscheinlichkeitswerte in Tab. XI gelten für zweiseitige Tests, sie müssen bei einseitigen Tests halbiert werden.) Wir sind also berechtigt H_0 zurückzuweisen, und können H_1 akzeptieren. Nach dem Unterricht haben also sehr signifikant mehr Schüler eine positive Einstellung ethnischen Minderheiten gegenüber als vor dem Unterricht.

(3) *Anwendungsregeln*

a) Die erhobenen Häufigkeiten werden in eine Vierfeldertafel eingetragen.
b) Die Erwartungswerte für die Zellen A und D werden bestimmt nach der Beziehung: $E = 1/2\,(A + D)$.
c) Chi^2 wird nach der Formel $\text{Chi}^2 = \frac{(A - D)^2}{A + D}$ berechnet.
d) Vergleich des errechneten Chi^2-Wertes mit dem in der Tab. XI unter Berücksichtigung der Freiheitsgrade ($df = 1$) und des Signifikanzniveaus abgelesenen Wert. Bei einseitigem Test ist der Wahrscheinlichkeitswert zu halbieren, da die Tabelle für zweiseitige Tests angelegt ist.

(4) *Voraussetzungen und Einschränkungen*

a) Die erwarteten Häufigkeiten für die Zellen A und D $(= 1/2\,(A + D))$ müssen > 5 sein.

b) Wenn $1/2(A + D) < 5$ ist, kann das Verfahren nicht angewendet werden (dann Binomialtest; s. Siegel 1956).
c) Wenn $(A + D) < 20$, ist eine korrigierte Formel zu verwenden:

$$Chi^2 = \frac{(|A - D| - 1)^2}{A + D}$$

(5) *Übungsbeispiel*

Nr. 5/26
25 Kinder einer Klasse wurden befragt, ob sie vor dem Schulantritt regelmäßig ein Frühstück einnehmen. Danach wurde ihnen im Unterricht der Vorteil eines regelmäßigen Frühstücks auseinandergesetzt. Nach einem Monat wurden die Kinder erneut befragt. Prüfen Sie die Hypothese: Nach dem Unterricht werden mehr Kinder regelmäßig ihr Frühstück einnehmen als vorher.

		Regelmäßiges Frühstück vor dem Unterricht: nein	ja
Regelmäßiges Frühstück nach dem Unterricht	ja	14	4
	nein	3	4

5.4.6. Der Cochran-Q-Test zum Vergleich mehrerer (k) abhängiger Stichproben

Der Cochran-Q-Test stellt eine Extension des im vorigen Kapitel dargestellten McNemar-Tests von 2 auf k abhängige Stichproben dar.

(1) *Untersuchungsproblem*

a) Es werden k Gruppen von Vpn gebildet, wobei sich die Versuchspersonen in jeder Gruppe hinsichtlich bestimmter relevanter Kontrollmerkmale (z.B. Intelligenz, Alter usw.) gleichen (sog. *matched groups*). Jede Gruppe enthält so viele Individuen, wie Untersuchungsbedingungen vorliegen. Die Vpn einer Gruppe werden jeweils per Zufall auf die verschiedenen Untersuchungsbedingungen aufgeteilt.
b) Es existieren k Versuchspersonen. Jede Vp wird jeder Untersuchungsbedingung in zufälliger Reihenfolge unterworfen (sog. „wiederholte Messungen").

Problem: Unterscheiden sich die verschiedenen Untersuchungsbedingungen unterworfenen Individuen bzw. Gruppen von anderen Individuen hinsichtlich der Häufigkeitsverteilung bei einem alternativ verteilten Merkmal?
Das gemessene Merkmal kann nur zwei Werte annehmen:
1 (für „Erfolg" – z.B. Lösung einer Aufgabe) und 0 (für „Mißerfolg" – Aufgabe nicht gelöst).

(2) *Beispiel*

a) Versuchsplan
Es soll geprüft werden, ob die Lösung einer Denksportaufgabe, für die eine Minute Zeit zur Verfügung steht, von der Tageszeit abhängt (Tageszeiten: 8.00, 12.00 und 16.00 Uhr). Es werden 10 Gruppen mit je 3 vergleichbaren Schülern gebildet. Die 3 Schüler jeder Gruppe werden per Zufall auf die drei Tageszeiten verteilt.

Stichproben: 3 abhängige Stichproben ($k = 3$) mit je 10 Vpn ($N = 10$)
Merkmal: Alternativ-Aufgabe gelöst: „1"; Aufgabe nicht gelöst: „0"

b) Hypothesen

H_0: Die Tageszeit der Aufgabenstellung hat keinen Einfluß auf die Lösung der Denksportaufgabe.

H_1: Die Lösung der Denksportaufgabe ist abhängig von der Tageszeit, zu verschiedenen Tageszeiten ergeben sich verschieden viele richtige Lösungen.

Signifikanzniveau: .05.

c) Ergebnisse

Die Ergebnisse werden in eine Tabelle eingetragen, deren Spalten die Untersuchungsbedingungen, die Reihen die Versuchspersonengruppen enthalten.

	Tageszeiten				
Vpn-Gruppen	8.00	12.00	16.00	Li	Li^2
1	0	1	0	1	1
2	1	0	0	1	1
3	1	0	1	2	4
4	0	1	1	2	4
5	1	1	1	3	9
6	1	0	0	1	1
7	1	0	1	2	4
8	0	0	1	1	1
9	1	1	0	2	4
10	1	0	1	2	4
	$G_1 = 7$	$G_2 = 4$	$G_3 = 6$	$\sum_{i=1}^{N} Li = 17$	$33 = \sum_{i=1}^{N} Li^2$

d) Berechnung

Um die Nullhypothese zu überprüfen, verwendet man folgende Formel:

$$Q = \frac{(k-1)\left[k \sum_{j=1}^{k} G_j^2 - \left(\sum_{j=1}^{k} G_j\right)^2\right]}{k \sum_{i=1}^{N} Li - \sum_{i=1}^{N} Li^2}; \quad df = k - 1$$

G_j = Gesamtzahl der „Erfolge" (richtigen Lösungen) in der j-ten Spalte, wobei j von 1 bis k läuft (Untersuchungsbedingungen).

L_i = Gesamtzahl der „Erfolge" (richtigen Lösungen) in der i-ten Reihe, wobei i von 1 bis N läuft (Vpn-Gruppen).

k = Anzahl der Untersuchungsbedingungen.

N = Anzahl der Vpn(-Gruppen).

Setzen wir nun unsere Werte in die Formel ein, dann ergibt sich:

$$Q = \frac{(3-1)\left[3\left(7^2 + 4^2 + 6^2\right) - 17^2\right]}{3 \cdot 17 - 33} = 1.55$$

e) Signifikanzprüfung

Da sich die Q-Verteilung einer Chi-Quadrat-Verteilung mit $df = k - 1$ annähert, können die kritischen Werte in der Tab. XI nachgeschlagen werden. Ist der erhaltene Q-Wert gleich oder größer als der abgelesene Wert, ist die Nullhypothese zurückzuweisen. Aus Tab. XI entnehmen wir, daß der erhaltene Wert kleiner ist als der bei $df = 2$ und Signifikanzniveau .05 abgelesene. H_0 kann nicht zurückgewiesen werden. Die Tageszeit hat keinen Einfluß auf die richtige Lösung der Denksportaufgabe.

(3) *Anwendungsregeln*

a) Jedem „Erfolg" wird eine 1, jedem „Mißerfolg" eine 0 zugeordnet.
b) Die Daten werden in eine Tabelle eingetragen, die k Spalten enthält – Zahl der Untersuchungsbedingungen – und N Reihen, wobei N der Zahl der Vpn in jeder der k Bedingungsgruppen entspricht.
c) Q wird nach der angegebenen Formel berechnet.
d) Der erhaltene Q-Wert wird mit dem in der Tab. XI unter Berücksichtigung der Freiheitsgrade ($df = k - 1$) und des Signifikanzniveaus abgelesenen Wert verglichen.
Ist der errechnete Wert gleich oder größer als der abgelesene, dann ist H_0 abzulehnen.

(4) *Voraussetzungen und Einschränkungen*

a) Die Zahl der Vpn unter jeder Untersuchungsbedingung muß gleich groß sein.
b) Der Test ist anwendbar bei Nominaldaten und dichotomisierten Ordinaldaten.

(5) *Übungsbeispiel*

Nr. 5/27
Jede von 20 Vpn wurde in zufälliger Reihenfolge 4 experimentellen Bedingungen unterworfen. Es war jedesmal ein mathematisches Problem zu lösen. Gelang die Lösung innerhalb einer Minute, dann wurde für jede Vpn eine 1 notiert, gelang die Lösung nicht, eine 0. Unterscheiden sich die Lösungshäufigkeiten unter den verschiedenen Bedingungen signifikant?

	Bedingung			
Vpn	1	2	3	5
1	1	1	1	0
2	0	1	1	1
3	0	0	1	0
4	1	1	1	1
5	0	1	0	0
6	0	0	1	0
7	1	0	0	0
8	0	0	1	1
9	0	0	0	0
10	1	0	0	0
11	1	0	1	0

	Bedingung			
Vpn	1	2	3	5
12	0	0	1	1
13	0	1	0	1
14	1	0	0	0
15	0	1	0	0
16	1	0	1	1
17	0	1	0	0
18	0	0	1	0
19	0	1	1	0
20	0	0	1	1

5.4.7. Wilcoxon-Test zum Vergleich zweier abhängiger Stichproben

(1) *Untersuchungsproblem*

Unterscheiden sich zwei abhängige Stichproben hinsichtlich eines auf Rangskalenniveau gemessenen Merkmals?

5.4.7.1. Wilcoxon-Test für Stichproben N < 25

(2) *Beispiel* (fiktiv)

a) Versuchsplan

Es soll untersucht werden, ob der Kindergartenbesuch einen Einfluß auf die Kontaktfreudigkeit der Kinder hat. Als Vpn dienen 8 Paare eineiiger Zwillinge. Per Zufall wird von jedem Zwillingspaar ein Kind zum Kindergartenbesuch zugeteilt, das andere Kind bleibt zu Hause. Nach einigen Monaten werden die Kinder auf einer Skala (die keine Intervallskala ist) hinsichtlich ihrer Kontaktbereitschaft eingestuft.

Stichproben: 2 abhängige Stichproben zu je 8 Kindern

Merkmal: Kontaktbereitschaft

b) Hypothesen

H_0: Es besteht kein Unterschied in der Kontaktbereitschaft von Kindern, die den Kindergarten besuchten, und solchen, die zu Hause blieben.

H_1: Die Kontaktbereitschaft von Kindern, die einen Kindergarten besuchten, unterscheidet sich signifikant von der Kontaktbereitschaft zu Hause gebliebener Kinder.

Signifikanzniveau: .05.

c) Ergebnisse

Zwillingspaar	Kontaktbereitschaft Kindergarten	zu Hause	d	Ränge von d	Ränge mit positivem Vorzeichen —	Ränge mit negativem Vorzeichen —	Ränge mit dem weniger häufigen Vorzeichen
a	53	54	− 1	1		− 1	1
b	62	48	14	7	+ 7		
c	41	57	− 16	8		− 8	8
d	74	68	6	4	+ 4		
e	88	75	13	6	+ 6		
f	47	58	11	5	+ 5		
g	91	88	3	3	+ 3		
h	72	74	− 2	2		− 2	2
					$T_1 = 25$	$T_2 = 11$	$T = 11$

d) Berechnung

Zunächst werden die Differenzen (d) zwischen den Meßwertpaaren gebildet, also $53 - 54 = -1$, $62 - 48 = 14$ usw. Diesen Differenzen werden ihrer absoluten Größe nach Rangplätze zugeordnet, die kleinste Differenz erhält Rang 1, die größte den

Rang N (N = Anzahl der Vpn, im Beispiel = 8). Das Vorzeichen einer Differenz ist für die Zuordnung der Ränge ohne Bedeutung. Allerdings werden nach der Rangzuteilung die Vorzeichen der Differenzen auf die ihnen zugeordneten Ränge übertragen. Eine Differenz von 0 wird nicht berücksichtigt, das entsprechende Meßwertpaar wird in der weiteren Analyse nicht mehr berücksichtigt (N ist dann = Anzahl der Meßwertpaare minus der Zahl der Meßwertpaare mit der Differenz 0). Treten gleich große Differenzen auf, dann werden ihnen, wie üblich (s. U-Test), mittlere Ränge zugeordnet.
Der nächste Schritt ist die Berechnung des T-Wertes. Wie beim U-Test kann man auch beim Wilcoxon-Test zwei T-Werte berechnen. Der erste T-Wert ergibt sich aus der Summe der Rangzahlen mit positivem Vorzeichen, der zweite T-Wert ergibt sich aus der Summe der Rangzahlen mit den negativen Vorzeichen. Der kleinere T-Wert wird für die Signifikanzprüfung berücksichtigt.
Vor allem bei größeren Stichproben kann man eine einfachere Methode zur Bestimmung des kleineren T-Wertes anwenden.
Man schreibt diejenigen Rangzahlen mit dem weniger häufigen Vorzeichen in eine besondere Spalte und addiert diese Zahlen. Die sich ergebende Summe ist der T-Wert.

e) Signifikanzprüfung
Zur Signifikanzprüfung ziehen wir Tab. XV heran.
Unter Berücksichtigung des Signifikanzniveaus und der Größe von N suchen wir den kritischen T-Wert auf. Ist der errechnete T-Wert gleich oder *kleiner* als der in der Tabelle aufgefundene Wert, dann kann die Nullhypothese zurückgewiesen werden.
Für N = 8 und Signifikanzniveau .05 lesen wir den kritischen T-Wert 4 ab. Unser errechneter T-Wert = 11 ist also größer. Die Nullhypothese, daß zwischen Kindern, die den Kindergarten besuchen, und solchen, die zu Hause bleiben, keine unterschiedliche Kontaktbereitschaft besteht, muß beibehalten werden.

5.4.7.2. Wilcoxon-Test für Stichproben N > 25

Bei Stichproben mit N > 25 kann Tab. XV nicht mehr benutzt werden. T ist bei großen Stichproben nahezu normal verteilt.
Die Signifikanzprüfung kann dann über z erfolgen:

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}$$

Welches T man einsetzt, ist gleichgültig.
Die Berechnung des T-Wertes entspricht den Darstellungen in Kap. 5.4.7.1. Ebenso gelten die übrigen dort gemachten Ausführungen. Nur die Signifikanzprüfung erfolgt in der gewohnten Weise über die z-Tabelle (Tab. II).

(3) *Anwendungsregeln*

a) Für jedes Meßwertpaar wird die Differenz ermittelt.
b) Diesen Differenzen werden Rangzahlen zugeordnet, und zwar nach der absoluten Größe der Differenzen. Die kleinste Differenz erhält den Rang 1, die größte den Rang N.
c) Den Rängen werden die Vorzeichen der entsprechenden Differenzen zugeordnet.
d) Der T-Wert wird berechnet. Er ergibt sich aus der kleineren Summe der Rangzahlen mit gleichem Vorzeichen.
e) N = die Gesamtzahl der Differenzen mit einem Vorzeichen (0-Differenzen entfallen also).
f) Für die Signifikanzprüfung ist
bei $N < 25$ die Tab. XV heranzuziehen,
bei $N > 25$ bedient man sich der z-Tabelle (Tab. II).

(4) *Voraussetzungen und Einschränkungen*

a) Es müssen Rangdaten vorliegen.
b) Die Differenzen müssen in eine Rangordnung ihrer absoluten Größe nach gebracht werden können.

(5) *Übungsbeispiel*

Nr. 5/28
In einem Experiment wurden 10 Paare von je zwei vergleichbaren Vpn gebildet *(matched pairs)*. Per Zufall wurde nun je ein Mitglied eines Paares der Bedingung A, das andere der Bedingung B unterworfen. Es ist gefragt, ob sich die Leistungen der Vpn unter Bedingung A von denen der Vpn unter Bedingung B signifikant unterscheiden.

Paar	Bedingung A	Bedingung B
1	83	75
2	80	78
3	81	66
4	74	77
5	79	80
6	78	68
7	72	75
8	84	90
9	85	81
10	88	83

5.4.8. Die Friedman-Rangvarianzanalyse zum Vergleich mehrerer (k) abhängiger Stichproben

(1) *Untersuchungsproblem*

a) Es werden N Gruppen mit jeweils k vergleichbaren Individuen gebildet, und zwar vergleichbar in bezug auf bestimmte relevante Kontrollmerkmale (z.B. Intelligenz, Schulbildung usw.). Jede Gruppe enthält so viel Individuen, wie Untersuchungsbedingungen vorliegen (k). Jedes Individuum einer Gruppe wird per Zufall einer der k Versuchsbedingungen zugeteilt (sog. Blockpläne). Wir haben also k abhängige Stichproben mit jeweils N Versuchspersonen.

b) Jedes von N Individuen wird in zufälliger Reihenfolge jeder der k Untersuchungsbedingungen unterworfen (sog. „wiederholte Messungen"). Wir haben auch hier k abhängige Stichproben mit N Versuchspersonen.

Problem: Unterscheiden sich die den k verschiedenen Untersuchungsbedingungen unterworfenen Individuen bzw. Gruppen von anderen Individuen hinsichtlich des gemessenen Merkmals?

5.4.8.1. Friedman-Rangvarianzanalyse für kleine Stichproben (k = 3, N = 2 bis 9; und k = 4, N = 2 bis 4)

(2) *Beispiel*

a) Versuchsplan
Es soll der unterschiedliche Lernerfolg von 4 Lehrmethoden überprüft werden. Dazu werden 3 Gruppen mit jeweils 4 Schülern gebildet, wobei die Schüler innerhalb jeder Gruppe in bezug auf eine Reihe von Merkmalen (z.B. Alter, Intelligenz usw.) vergleichbar sind. Aus jeder Gruppe wird per Zufall ein Schüler einer der vier Lehrmethoden zugeordnet.
Stichproben: 4 (k) abhängige Stichproben mit jeweils 3 (N) Vpn
Merkmal: Lernerfolg (auf einer Skala eingeschätzt)

b) Hypothesen
H_0: Die 4 verschiedenen Lehrmethoden haben keinen Einfluß auf den Lernerfolg.
H_1: Der Lernerfolg bei den 4 Lehrmethoden wird unterschiedlich sein.
Signifikanzniveau: .05.

c) Ergebnisse

Vpn Gruppe	Lehrmethoden I	II	III	IV
1	8	11	5	9
2	7	10	11	8
3	9	7	8	10

Wie dargestellt, werden die Werte in eine zweidimensionale Matrix eingetragen, die Spalten repräsentieren die verschiedenen Bedingungen (Lehrmethoden), die Reihen die verschiedenen Gruppen von Vpn.

d) Berechnung
Die Meßwerte in jeder Reihe werden mit Rangzahlen versehen. Der niedrigste Wert einer Reihe erhält den Rang 1, der höchste den Rang k (k = Zahl der Bedingungen). In jeder *Spalte* werden die Rangsummen gebildet.
Es ergibt sich:

	I	II	III	IV
1	2	4	1	3
2	1	3	4	2
3	3	1	2	4
	$R_j = 6$	8	7	9

$\sum_{j=1}^{k} R_j = 30$

Ob die Ränge korrekt zugeteilt werden, läßt sich über die Beziehung

$$\sum_{j=1}^{k} R_j = \frac{Nk(k+1)}{2}$$

überprüfen. Bei korrekter Rangzuteilung ergeben sich auf beiden Seiten der Gleichung identische Werte: $6 + 8 + 7 + 9 = \frac{3 \cdot 4(4+1)}{2}$; $30 = 30$

Würde die Nullhypothese gelten, dann dürften sich die Rangsummen R_j nur per Zufall voneinander unterscheiden. Um dies zu überprüfen, wird der von Friedman mit Chi_r^2 bezeichnete Wert berechnet:

$$Chi_r^2 = \frac{12}{N\,k(k+1)} \left(\sum_{j=1}^{k} (R_j)^2 \right) - 3N(k+1)$$

N = Zahl der Reihen
k = Zahl der Spalten
R_j = Summe der Rangzahlen in der j-ten Spalte
Setzen wir unsere Werte in die Formel ein, so erhalten wir:

$$Chi_r^2 = \frac{12}{3 \cdot 4\,(4+1)} (6^2 + 8^2 + 7^2 + 9^2) - 3 \cdot 3\,(4+1) = 1.0$$

e) Signifikanzprüfung
Für kleine Stichproben (k = 3, N von 2 bis 9; und k = 4, N von 2 bis 4) kann zur Signifikanzprüfung die Tab. XVI mit exakten Wahrscheinlichkeiten herangezogen werden.
Wir suchen in Tab. XVI für k = 4 die Auftretenswahrscheinlichkeit für $Chi_r^2 = 1$ und erhalten .910. Da dieser Wert größer ist als .05, ist die Nullhypothese beizubehalten.

5.4.8.2. Friedman-Rangvarianzanalyse für Stichproben mit großem N und k

Wenn die Zahl der Reihen und/oder der Spalten nicht zu klein ist, verteilt sich Chi_r^2 wie Chi-Quadrat mit $df = k - 1$. Die Berechnung von Chi_r^2 erfolgt in der gleichen Weise wie in diesem und Kap. 5.4.8.1 beschrieben. Die Signifikanzprüfung wird mit Hilfe der Tab. XI vorgenommen. Ist der errechnete Chi_r^2-Wert gleich oder größer als der aufgefundene Tabellenwert bei $df = k - 1$, dann können wir die Nullhypothese auf dem entsprechenden Signifikanzniveau zurückweisen.
Außer dem Einfluß der verschiedenen Untersuchungsbedingungen können wir auch überprüfen, ob zwischen den Gruppen bzw. Individuen über alle Untersuchungsbedingungen hinweg ein Unterschied besteht (ob z.B. etwa intelligentere Schüler bei allen Lehrmethoden besser abschneiden).
Dazu werden Meßwerte in jeder Spalte mit Rangzahlen versehen. Für jede Reihe werden die Rangsummen gebildet.
Es ergibt sich:

Vpn	Lehrmethoden				
	I	II	III	IV	R_j
1	2	3	1	2	8
2	1	2	3	1	7
3	3	1	2	3	1

Ob die Ränge korrekt zugeteilt wurden, läßt sich wie oben über die Beziehung

$$\sum_{j=1}^{N} R_j = \frac{k\,N\,(N+1)}{2}, \quad 8 + 7 + 9 = \frac{4 \cdot 3\,(3+1)}{2}$$

$$24 = 24$$

überprüfen.
Zur Berechnung von Chi_r^2 wird umgeformt:

$$Chi_r^2 = \frac{12}{k\,N\,(N+1)} \left(\sum_{j=1}^{N} (R_j)^2 \right) - 3\,k\,(N+1)$$

Die weitere Auswertung entspricht den oben dargestellten Vorgehensweisen. Chi_r^2 hat hier den Wert $Chi_r^2 = 0.50$.
Bei $k = 3$ und $N = 4$ lesen wir ab: .931
Zwischen den Gruppen besteht also kein Unterschied.

(3) *Anwendungsregeln*

a) Die Meßwerte werden in eine zweidimensionale Matrix eingetragen, die k Spalten (Bedingungen) und N Reihen (Vpn o. Gruppen) hat.
b) Den Werten in jeder Reihe werden Rangzahlen von 1 bis k zugeordnet.
c) Die Ränge in jeder Spalte werden addiert: R_j.
d) Chi_r^2 wird nach der angegebenen Formel berechnet.
e) Je nach Stichprobengröße wird Tab. XVI oder Tab. XI zur Signifikanzprüfung herangezogen.

(4) *Voraussetzungen und Einschränkungen*

Die Daten müssen Rangskalenniveau haben.

(5) *Übungsbeispiel*

Nr. 5/29
11 Gruppen von je 4 Schülern wurden gebildet, wobei die 4 Schüler untereinander vergleichbar waren. Jeder der 4 Schüler jeder Gruppe wurde einer von vier Lehrmethoden per Zufall zugeordnet. Ist der Lernerfolg bei den 4 Lehrmethoden unterschiedlich?

	Lehrmethode			
Gruppe	I	II	III	IV
1	1	4	8	0
2	2	3	13	1
3	10	0	11	3
4	12	11	13	10
5	1	3	10	0
6	10	3	11	9
7	4	12	10	11
8	10	4	5	3
9	10	4	9	3
10	14	4	7	2
11	3	2	4	13

Anhang

Anhang 1

Lösungen zu den Übungsbeispielen in den Kapiteln 3 bis 5

Aufgaben in Kapitel 3.

Nr. 3/1

$\bar{X}$	Strichliste	f
40–44	/	1
45–49	///	3
50–54	//	2
55–59	////	4
60–64	////	4
65–69	~~////~~ /	6
70–74	~~////~~ ~~////~~	10
75–79	~~////~~ ///	8
80–84	~~////~~	5
85–89	////	4
90–94	//	2
95–99	/	1
i = 5		N = 50

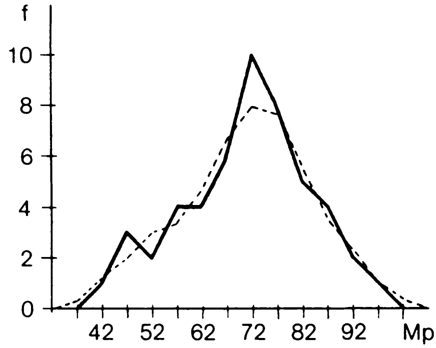
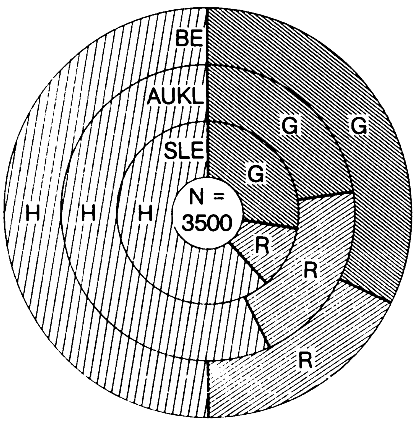


Abb. 33: Adjustierte Frequenzen:
0.3/1.3/2.0/3.0/3.3/4.7/6.7/
8.0/7.7/5.7/3.7/2.3/1.0/0.3

Nr. 3/2 *Abb. 34:*



Nr. 3/3
Mod = 72 RP; Med = 72.0 RP; M = 70.8 RP

Nr. 3/4
Med = 60.80 RP; M = 60.76 RP
cumf-Werte: 1/4/6/10/15/22/28/32/35/37/39
cumf%-Werte: 2.6/10.3/15.4/25.6/38.5/56.4/71.8/82.1/89.7/94.9/100.0

Nr. 3/5
Absolute Dispersion = 42 RP bis 97 RP; IQD = 16.56 RP; Q = 8.28 RP
sowie MV = 10.04 RP; s^2 = 157.24 RP; s = 12.54 RP

Nr. 3/6
Q = 3.375 RP sowie s = 4.715 RP

Nr. 3/7
Ergebnisse der Normalisierung einer anormalen Verteilung (Flächentransformation):

X (Mp)	f	cumf	$\text{cumf} - \frac{f}{2}$	$\frac{\text{cumf} - \frac{f}{2}}{N} 100$	z	T	z*
10	6	6	3	3	− 1.88	31	− 1.60
15	8	14	10	10	− 1.28	37	− 1.08
20	30	44	28	29	− 0.55	44	− 0.62
25	20	64	54	54	0.10	51	− 0.17
30	10	74	69	69	0.50	55	0.29
35	8	82	78	78	0.77	58	0.75
40	6	88	85	85	1.04	60	1.21
45	6	94	91	91	1.34	63	1.67
50	4	98	96	96	1.75	68	2.13
55	2	100	99	99	2.33	73	2.59

N = 100 M = 26.80 RP s = 10.90 RP

Anmerkung: Zum Vergleich werden auch die nichtnormalisierten z-Werte z* (aufgrund von M und s berechnet!) mit angegeben.

Nr. 4/1
r = − .16; nicht signifikant

Nr. 4/2
r = .25; nicht signifikant

Nr. 4/3

$$r_{12.3} = \frac{.78 - (.52)(.54)}{\sqrt{(1 - .52^2)(1 - 54^2)}} = .69$$

Bei Konstanthaltung des Alters erniedrigt sich die Korrelation zwischen Größe und Gewicht.

Nr. 4/4
Es ist der punktbiseriale Korrelationskoeffizient zu berechnen.
$M_p = 3.037$; $M_q = 2.558$; $s_t = .631$; $p = .43$; $q = .57$
$r_{p\text{-bis.}} = .38$; $t = 1.317$; nicht signifikant

Nr. 4/5
Rho = .695; signifikant $p < .01$

Nr. 4/6
$k = 3$; $N = 6$; QUSR = 25.5; $W = .16$; nicht signifikant

Nr. 4/7
Phi = .0769; $Chi^2 = .7316$; nicht signifikant

Nr. 4/8
$Chi^2 = 91.66$; $C = .39$; signifikant $p < .001$
$df = 4$

Nr. 4/9
$C_C = .302$

Nr. 4/10

Vpn Nr.	Y'	Vpn Nr.	Y'
1	68	11	68
2	55	12	64
3	68	13	65
4	70	14	63
5	71	15	60
6	54	16	57
7	77	17	65
8	70	18	57
9	61		
10	63		

Nr. 4/11 a
$$z_1' = .098 \cdot 1.7 + .631 \cdot .9$$
$$= .735$$

Nr. 4/11 b
$$z_1' = .098 \cdot .9 + .631 \cdot 1.7$$
$$= 1.161$$

Der Wortschatztest ist demnach für die Deutschzensur relevanter als der Intelligenztest.

Nr. 5/1
$p = .25$ (25%)

Nr. 5/2
$p = .172$ (17,2%)

Nr. 5/3
$p = .003$ (3%)

Nr. 5/4 a
47.72%

Nr. 5/4 b
95.44%

Nr. 5/5 a
93.32%

Nr. 5/5 b
$z = 0.67$

Nr. 5/5c
$z = -0.84$ $(z = +0.84)$

Nr. 5/6a
74%

Nr. 5/6b
5%

Nr. 5/6c
± 1.15 z (zwischen 11.4 RP und 20.6 RP)

Nr. 5/7
$z_{emp} = 1.73$, $z_{theo} = 1.96$, $p = .05$
$z_{emp} < z_{theo}$

Der Altersunterschied zwischen dem Durchschnitt aller BRD-Volksschullehrer und der Volksschullehrer im Lande Bremen ist nicht bedeutsam bei einer Fehlerwahrscheinlichkeit von 5%.

Nr. 5/8
$z_{emp} = 1{,}2$, $z_{theo} = 1.96$, $p = .05$
$z_{emp} < z_{theo}$

Der Unterschied zwischen dem mittleren Intelligenzgrad der untersuchten Klasse und dem Bundesdurchschnitt ist demnach nicht signifikant, wenn wir eine Unsicherheit von 5% für unsere Aussagen zulassen.

Nr. 5/9
$t_{emp} = 1.45$, $t_{theo} = 2.05$, $df = 28$, $p = .05$ (zweiseitig)
$t_{emp} < t_{theo}$

Wir können mit 95% Sicherheit sagen, daß zwischen den beiden Stadtgebieten hinsichtlich der IQs von Schülern kein signifikanter Unterschied besteht.

Nr. 5/10
Da Varianzheterogenität für die beiden Gruppen besteht, muß der t-Test nach der Formel für Stichproben mit heterogenen Varianzen berechnet werden. Dann ist:
$t_1 = 2.131$, $t_2 = 2.201$, $t_{emp} = 0.84$, $t_{theo} = 2.14$, $df = 26$, $p = .05$, $t_{emp} < t_{theo}$

Die Lernfähigkeit der beiden untersuchten Stichproben unterscheidet sich, bezogen auf den verwendeten Test, mit einer Sicherheit von 95% nicht signifikant.

Nr. 5/11
$t_{emp} = 1.18$, $t_{theo} = 2.262$, $df = 9$, $p = .05$ (einseitige Frage)
$t_{emp} < t_{theo}$

Das Ergebnis ist nicht signifikant auf dem 5%-Niveau, d.h., die Streß-Bedingung beeinflußt die Reaktionsgeschwindigkeit bei Studenten nicht bedeutsam positiv.

Nr. 5/12
$t_{emp} = 2.7$, $t_{theo} = 1.73$, $df = 18$, $p = .05$ (einseitige Frage)
$t_{emp} > t_{theo}$

Das Ergebnis ist signifikant auf dem 5%-Niveau, d.h., die Therapie ist für Tic-Bewegungen tatsächlich erfolgreich.

Nr. 5/13
$F_{emp} = 7.44$, $F_{theo} = 2.55$, $df_1 = 15$, $df_2 = 13$, $p = .05$
$F_{emp} > F_{theo}$

Die beiden Varianzen unterscheiden sich mit 95% Wahrscheinlichkeit signifikant voneinander.

Nr. 5/14
Es ist der F_{max}-Test durchzuführen.

$F_{emp} = 1.6$, $F_{theo} = 3.41$, $df_1 = 19$, $df_2 = 7$, $p = .05$
$F_{emp} < F_{theo}$

Für die 4 Gruppen besteht mit 95% Sicherheit Varianzhomogenität.

Nr. 5/15

	SAQ	df	MAQ
ZW	11	2	5.50
IN	1030	27	35.82
TOT	1041	29	35.09

$F_{emp} = .14$, $F_{theo} = 3.35$, $p = .05$
Die Volksweisheit trifft zumindest für die Gemeinde Abbesbüttel mit einer Wahrscheinlichkeit von 95% nicht zu; der F-Test in der Varianzanalyse erbringt kein signifikantes Ergebnis.

Nr. 5/16

	SAQ	df	MAQ
ZW	50.08	2	25.04
IN	86.88	21	4.14
TOT	136.96	23	5.94

$F_{emp} = 6.05$, $F_{theo} = 3.47$, $p = .05$
Die 3 Kurse sind mit 95% Wahrscheinlichkeit nicht gleich effektiv; der F-Test in der Varianzanalyse erbringt ein signifikantes Ergebnis. – Kurs 3 scheint der wirksamste zu sein (letzteres wird in Beispielaufgabe 11 überprüft). Wie stark der Zusammenhang zwischen abhängiger (Kurs-Art) ist, wird in Beispiel 12 geprüft.

Nr. 5/17
Vergleich zw. Kurs 2 u. 3: $q_{emp} = 4.32$
Vergleich zw. Kurs 1 u. 3: $q_{emp} = 4.16$
Vergleich zw. Kurs 1 u. 2: $q_{emp} = .18$

q_{theo} (für alle) = 2.95 (5%-Niveau)
Die ersten beiden empirischen q-Werte sind größer als der theoretische. Daraus zeigt sich, daß der Kurs 3 im Vergleich zu den beiden anderen eine signifikant höhere Effektivität besitzt. Kurs 1 und 2 unterscheiden sich hinsichtlich der Wirksamkeit für den Erwerb von Kenntnissen nicht bedeutsam.

Nr. 5/18
Omega = .29
Die Gesamtvarianz in unserem Beispiel zur Effektivität von Förderkursen ist demnach zu 29% systematischer Art und beruht zu 71% auf Fehlern.

Nr. 5/19
Prüfung gegen Gleichverteilung
$Chi^2 = 22.5$; $df = 4$; signifikant $p < .001$

Nr. 5/20
$Chi^2 = 4.74$; $df = 1$; signifikant $p < .05$

Nr. 5/21
$Chi^2 = 9.36$; $df = 2$; signifikant $p < .05$

Nr. 5/22
$2\hat{I} = 3345.933 - 3337.334 = 8.599$; $df = 1$; signifikant $p < .01$

Nr. 5/23
$U = 15$; $U = 5$; $p = .143$ nicht signifikant $p > .05$

Nr. 5/24
$R_1 = 321$; $R_2 = 382$; $U = 230$; $U' = 82$; $z = 2.35$; einseitiger Test, signifikant $p < .01$

Nr. 5/25
$\sum R_j = 666$; $H = 13.81$; $df = 2$; signifikant $p < .01$

Nr. 5/26
$Chi^2 = 4.5$; $df = 1$; einseitiger Test 1/2 (.05), signifikant

Nr. 5/27
$Q = 3.40$; $df = 3$; nicht signifikant $p > .05$

Nr. 5/28
$T = 15$; nicht signifikant

Nr. 5/29
$R_I = 30$; $R_{II} = 24$; $R_{III} = 38$; $R_{IV} = 18$
$Chi_r^2 = 11.79$; $df = 3$; signifikant, $p < .01$

Anhang 2

Statistische Tabellen

Tab. I:
Zufallszahlen (für die Stichprobenbildung)

	00–04	05–09	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067
20	75884	12952	84318	95108	72305	64620	91318	89872	45375	85436
21	16777	37116	58550	42958	21460	43910	01175	87894	81378	10620
22	46230	43877	80207	88877	89380	32992	91380	03164	98656	59337
23	42902	66892	46134	01432	94710	23474	20423	60137	60609	13119
24	81007	00333	39693	28039	10154	95425	39220	19774	31782	49037
25	68089	01122	51111	72373	06902	74373	96199	97017	41273	21546
26	20411	67081	89950	16944	93054	87687	96693	87236	77054	33848
27	58212	13160	06468	15718	82627	76999	05999	58680	96739	63700
28	70577	42866	24969	61210	76046	67699	42054	12696	93758	03283
29	94522	74358	71659	62038	79643	79169	44741	05437	39038	13163
30	42626	86819	85651	88678	17401	03252	99547	32404	17918	62880
31	16051	33763	57194	16752	54450	19031	58580	47629	54132	60631
32	08244	27647	33851	44705	94211	46716	11738	55784	95374	72655
33	59497	04392	09419	89964	51211	04894	72882	17805	21896	83864
34	97155	13428	40293	09985	58434	01412	69124	82171	59058	82859
35	98409	66162	95763	47420	20792	61527	20441	39435	11859	41567
36	45476	84882	65109	96597	25930	66790	65706	61203	53634	22557

	00–04	05–09	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49
37	89300	69700	50741	30329	11658	23166	05400	66669	48708	03887
38	50051	95137	91631	66315	91428	12275	24816	68091	71710	33258
39	31753	85178	31310	89642	98364	02306	24617	09609	83942	22716
40	79152	53829	77250	20190	56535	18760	69942	77448	33278	48805
41	44560	38750	83635	56540	64900	42912	13953	79149	18710	68618
42	68328	83378	63369	71381	39564	05615	42451	64559	97501	65747
43	46939	38689	58625	08342	30459	85863	20781	09284	26333	91777
44	83544	86141	15707	96256	23068	13782	08467	89469	93842	55349
45	91621	00881	04900	54224	46177	55309	17852	27491	89415	23466
46	91896	67126	04151	03795	59077	11848	12630	98375	52058	60142
47	55751	62515	21108	80830	02263	29303	37204	96926	30506	09808
48	85156	87689	95493	88842	00664	55017	55539	17771	69448	87530
49	07521	56898	12236	60277	39102	62315	12239	07105	11844	01117

Entnommen aus: Scott & Wertheimer 1962, S. 400.

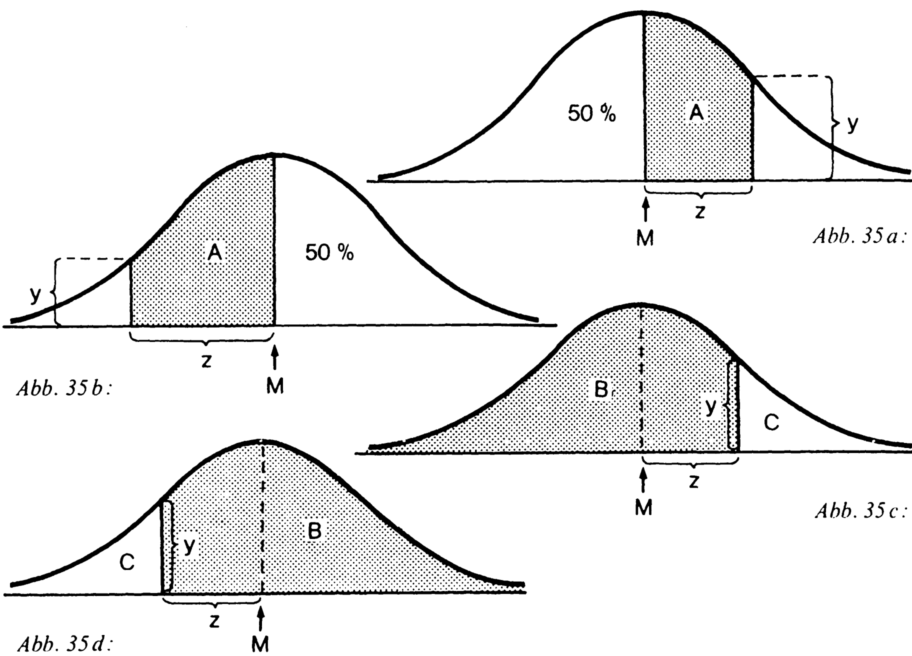


Abb. 35a:

Abb. 35b:

Abb. 35c:

Abb. 35d:

Abb. 35: Erläuterungsskizzen zu Tab. II (nach Guilford 1965, S. 569).
Für die tabellierten und nachstehend abgebildeten Flächenstücke (A, B, C) unter der Normalkurve gelten folgende Beziehungen:

$A = B - 50\%$
$A = 50\% - C$ (A + C = 50%)
$B = A + 50\%$
$B = 100\% - C$ (B + C = 100%)
$C = 50\% - A$
$C = 100\% - B$

Tab. II:
Flächenstücke (in %) und Ordinaten (y) der Gaußschen Normalverteilungskurve, ausgedrückt in Sigma-Einheiten: im Abstand $z = \frac{X - \mu}{\sigma}$ oder $z = \frac{x}{\sigma}$ vom Mittelwert (μ).

	Flächenstücke[1]			
z Standard-Wert $(\frac{x}{\sigma})$	A zwischen μ und z	B im größeren Teil	C im kleineren Teil	y Ordinate bei $\frac{x}{\sigma}$
0.00	.0000	.5000	.5000	.3989
0.01	.0040	.5040	.4960	.3989
0.02	.0080	.5080	.4920	.3989
0.03	.0120	.5120	.4880	.3988
0.04	.0160	.5160	.4840	.3986
0.05	.0199	.5199	.4801	.3984
0.06	.0239	.5239	.4761	.3982
0.07	.0279	.5279	.4721	.3980
0.08	.0319	.5319	.4681	.3977
0.09	.0359	.5359	.4641	.3973
0.10	.0398	.5398	.4602	.3970
0.11	.0438	.5438	.4562	.3965
0.12	.0478	.5478	.4522	.3961
0.13	.0517	.5517	.4483	.3956
0.14	.0557	.5557	.4443	.3951
0.15	.0596	.5596	.4404	.3945
0.16	.0636	.5636	.4364	.3939
0.17	.0675	.5675	.4325	.3932
0.18	.0714	.5714	.4286	.3925
0.19	.0753	.5753	.4247	.3918
0.20	.0793	.5793	.4207	.3910
0.21	.0832	.5832	.4168	.3902
0.22	.0871	.5871	.4129	.3894
0.23	.0910	.5910	.4090	.3885
0.24	.0948	.5948	.4052	.3876
0.25	.0987	.5987	.4013	.3867
0.26	.1026	.6026	.3974	.3857
0.27	.1064	.6064	.3936	.3847
0.28	.1103	.6103	.3897	.3836
0.29	.1141	.6141	.3859	.3825
0.30	.1179	.6179	.3821	.3814
0.31	.1217	.6217	.3783	.3802
0.32	.1255	.6255	.3745	.3790
0.33	.1293	.6293	.3707	.3778
0.34	.1331	.6331	.3669	.3765
0.35	.1368	.6368	.3632	.3752
0.36	.1406	.6406	.3594	.3739

1 Siehe die Erläuterung auf S. 254 in disem Buch.

Fortsetzung von Tab. II

z Standard-Wert $(\frac{x}{\sigma})$	Flächenstücke A zwischen μ und z	B im größeren Teil	C im kleineren Teil	y Ordinate bei $\frac{x}{\sigma}$
0.37	.1443	.6443	.3557	.3725
0.38	.1480	.6480	.3520	.3712
0.39	.1517	.6517	.3483	.3697
0.40	.1554	.6554	.3446	.3683
0.41	.1591	.6591	.3409	.3668
0.42	.1628	.6628	.3372	.3653
0.43	.1664	.6664	.3336	.3637
0.44	.1700	.6700	.3300	.3621
0.45	.1736	.6736	.3264	.3605
0.46	.1772	.6772	.3228	.3589
0.47	.1808	.6808	.3192	.3572
0.48	.1844	.6844	.3156	.3555
0.49	.1879	.6879	.3121	.3538
0.50	.1915	.6915	.3085	.3521
0.51	.1950	.6950	.3050	.3503
0.52	.1985	.6985	.3015	.3485
0.53	.2019	.7019	.2981	.3467
0.54	.2054	.7054	.2946	.3448
0.55	.2088	.7088	.2912	.3429
0.56	.2123	.7123	.2877	.3410
0.57	.2157	.7157	.2843	.3391
0.58	.2190	.7190	.2810	.3372
0.59	.2224	.7224	.2776	.3352
0.60	.2257	.7257	.2743	.3332
0.61	.2291	.7291	.2709	.3312
0.62	.2324	.7324	.2676	.3292
0.63	.2357	.7357	.2643	.3271
0.64	.2389	.7389	.2611	.3251
0.65	.2422	.7422	.2578	.3230
0.66	.2454	.7454	.2546	.3209
0.67	.2486	.7486	.2514	.3187
0.68	.2517	.7517	.2483	.3166
0.69	.2549	.7549	.2451	.3144
0.70	.2580	.7580	.2420	.3123
0.71	.2611	.7611	.2389	.3101
0.72	.2642	.7642	.2358	.3079
0.73	.2673	.7673	.2327	.3056
0.74	.2704	.7704	.2296	.3034
0.75	.2734	.7734	.2266	.3011
0.76	.2764	.7764	.2236	.2989
0.77	.2794	.7794	.2206	.2966
0.78	.2823	.7823	.2177	.2943
0.79	.2852	.7852	.2148	.2920

Fortsetzung von Tab. II

z Standard-Wert ($\frac{x}{\sigma}$)	A zwischen μ und z	Flächenstücke B im größeren Teil	C im kleineren Teil	y Ordinate bei $\frac{x}{\sigma}$
0.80	.2881	.7881	.2119	.2897
0.81	.2910	.7910	.2090	.2874
0.82	.2939	.7939	.2061	.2850
0.83	.2967	.7967	.2033	.2827
0.84	.2995	.7995	.2005	.2803
0.85	.3023	.8023	.1977	.2780
0.86	.3051	.8051	.1949	.2756
0.87	.3078	.8078	.1922	.2732
0.88	.3106	.8106	.1894	.2709
0.89	.3133	.8133	.1867	.2685
0.90	.3159	.8159	.1841	.2661
0.91	.3186	.8186	.1814	.2637
0.92	.3212	.8212	.1788	.2613
0.93	.3238	.8238	.1762	.2589
0.94	.3264	.8264	.1736	.2565
0.95	.3289	.8289	.1711	.2541
0.96	.3315	.8315	.1685	.2516
0.97	.3340	.8340	.1660	.2492
0.98	.3365	.8365	.1635	.2468
0.99	.3389	.8389	.1611	.2444
1.00	.3413	.8413	.1587	.2420
1.01	.3438	.8438	.1562	.2396
1.02	.3461	.8461	.1539	.2371
1.03	.3485	.8485	.1515	.2347
1.04	.3508	.8508	.1492	.2323
1.05	.3531	.8531	.1469	.2299
1.06	.3554	.8554	.1446	.2275
1.07	.3577	.8577	.1423	.2251
1.08	.3599	.8599	.1401	.2227
1.09	.3621	.8621	.1379	.2203
1.10	.3643	.8643	.1357	.2179
1.11	.3665	.8665	.1335	.2155
1.12	.3686	.8686	.1314	.2131
1.13	.3708	.8708	.1292	.2107
1.14	.3729	.8729	.1271	.2083
1.15	.3749	.8749	.1251	.2059
1.16	.3770	.8770	.1230	.2036
1.17	.3790	.8790	.1210	.2012
1.18	.3810	.8810	.1190	.1989
1.19	.3830	.8830	.1170	.1965
1.20	.3849	.8849	.1151	.1942
1.21	.3869	.8869	.1131	.1919
1.22	.3888	.8888	.1112	.1895

Fortsetzung von Tab. II

z Standard-Wert $(\frac{x}{\sigma})$	A zwischen μ und z	Flächenstücke B im größeren Teil	C im kleineren Teil	y Ordinate bei $\frac{x}{\sigma}$
1.23	.3907	.8907	.1093	.1872
1.24	.3925	.8925	.1075	.1849
1.25	.3944	.8944	.1056	.1826
1.26	.3962	.8962	.1038	.1804
1.27	.3980	.8980	.1020	.1781
1.28	.3997	.8997	.1003	.1758
1.29	.4015	.9015	.0985	.1736
1.30	.4032	.9032	.0968	.1714
1.31	.4049	.9049	.0951	.1691
1.32	.4066	.9066	.0934	.1669
1.33	.4082	.9082	.0918	.1647
1.34	.4099	.9099	.0901	.1626
1.35	.4115	.9115	.0885	.1604
1.36	.4131	.9131	.0869	.1582
1.37	.4147	.9147	.0853	.1561
1.38	.4162	.9162	.0838	.1539
1.39	.4177	.9177	.0823	.1518
1.40	.4192	.9192	.0808	.1497
1.41	.4207	.9207	.0793	.1476
1.42	.4222	.9222	.0778	.1456
1.43	.4236	.9236	.0764	.1435
1.44	.4251	.9251	.0749	.1415
1.45	.4265	.9265	.0735	.1394
1.46	.4279	.9279	.0721	.1374
1.47	.4292	.9292	.0708	.1354
1.48	.4306	.9306	.0694	.1334
1.49	.4319	.9319	.0681	.1315
1.50	.4332	.9332	.0668	.1295
1.51	.4345	.9345	.0655	.1276
1.52	.4357	.9357	.0643	.1257
1.53	.4370	.9370	.0630	.1238
1.54	.4382	.9382	.0618	.1219
1.55	.4394	.9394	.0606	.1200
1.56	.4406	.9406	.0594	.1182
1.57	.4418	.9418	.0582	.1163
1.58	.4429	.9429	.0571	.1145
1.59	.4441	.9441	.0559	.1127
1.60	.4452	.9452	.0548	.1109
1.61	.4463	.9463	.0537	.1092
1.62	.4474	.9474	.0526	.1074
1.63	.4484	.9484	.0516	.1057
1.64	.4495	.9495	.0505	.1040

Fortsetzung von Tab. II

z Standard-Wert ($\frac{x}{\sigma}$)	Flächenstücke A zwischen μ und z	B im größeren Teil	C im kleineren Teil	y Ordinate bei $\frac{x}{\sigma}$
1.65	.4505	.9505	.0495	.1023
1.66	.4515	.9515	.0485	.1006
1.67	.4525	.9525	.0475	.0989
1.68	.4535	.9535	.0465	.0973
1.69	.4545	.9545	.0455	.0957
1.70	.4554	.9554	.0446	.0940
1.71	.4564	.9564	.0436	.0925
1.72	.4573	.9573	.0427	.0909
1.73	.4582	.9582	.0418	.0893
1.74	.4591	.9591	.0409	.0878
1.75	.4599	.9599	.0401	.0863
1.76	.4608	.9608	.0392	.0848
1.77	.4616	.9616	.0384	.0833
1.78	.4625	.9625	.0375	.0818
1.79	.4633	.9633	.0367	.0804
1.80	.4641	.9641	.0359	.0790
1.81	.4649	.9649	.0351	.0775
1.82	.4656	.9656	.0344	.0761
1.83	.4664	.9664	.0336	.0748
1.84	.4671	.9671	.0329	.0734
1.85	.4678	.9678	.0322	.0721
1.86	.4686	.9686	.0314	.0707
1.87	.4693	.9693	.0307	.0694
1.88	.4699	.9699	.0301	.0681
1.89	.4706	.9706	.0294	.0669
1.90	.4713	.9713	.0287	.0656
1.91	.4719	.9719	.0281	.0644
1.92	.4726	.9726	.0274	.0632
1.93	.4732	.9732	.0268	.0620
1.94	.4738	.9738	.0262	.0608
1.95	.4744	.9744	.0256	.0596
1.96	.4750	.9750	.0250	.0584
1.97	.4756	.9756	.0244	.0573
1.98	.4761	.9761	.0239	.0562
1.99	.4767	.9767	.0233	.0551
2.00	.4772	.9772	.0228	.0540
2.01	.4778	.9778	.0222	.0529
2.02	.4783	.9783	.0217	.0519
2.03	.4788	.9788	.0212	.0508
2.04	.4793	.9793	.0207	.0498

Die z-Werte für die mittleren 95% (z = −1.96 bzw. z = +1.96) markieren das sog. *5%-Niveau* zur Sicherung von Statistiken.

Fortsetzung von Tab. II

z Standard-Wert ($\frac{x}{\sigma}$)	Flächenstücke A zwischen μ und z	B im größeren Teil	C im kleineren Teil	y Ordinate bei $\frac{x}{\sigma}$
2.05	.4798	.9798	.0202	.0488
2.06	.4803	.9803	.0197	.0478
2.07	.4808	.9808	.0192	.0468
2.08	.4812	.9812	.0188	.0459
2.09	.4817	.9817	.0183	.0449
2.10	.4821	.9821	.0179	.0440
2.11	.4826	.9826	.0174	.0431
2.12	.4830	.9830	.0170	.0422
2.13	.4834	.9834	.0166	.0413
2.14	.4838	.9838	.0162	.0404
2.15	.4842	.9842	.0158	.0396
2.16	.4846	.9846	.0154	.0387
2.17	.4850	.9850	.0150	.0379
2.18	.4854	.9854	.0146	.0371
2.19	.4857	.9857	.0143	.0363
2.20	.4861	.9861	.0139	.0355
2.21	.4864	.9864	.0136	.0347
2.22	.4868	.9868	.0132	.0339
2.23	.4871	.9871	.0129	.0332
2.24	.4875	.9875	.0125	.0325
2.25	.4878	.9878	.0122	.0317
2.26	.4881	.9881	.0119	.0310
2.27	.4884	.9884	.0116	.0303
2.28	.4887	.9887	.0113	.0297
2.29	.4890	.9890	.0110	.0290
2.30	.4893	.9893	.0107	.0283
2.31	.4896	.9896	.0104	.0277
2.32	.4898	.9898	.0102	.0270
2.33	.4901	.9901	.0099	.0264
2.34	.4904	.9904	.0096	.0258
2.35	.4906	.9906	.0094	.0252
2.36	.4909	.9909	.0091	.0246
2.37	.4911	.9911	.0089	.0241
2.38	.4913	.9913	.0087	.0235
2.39	.4916	.9916	.0084	.0229
2.40	.4918	.9918	.0082	.0224
2.41	.4920	.9920	.0080	.0219
2.42	.4922	.9922	.0078	.0213
2.43	.4925	.9925	.0075	.0208
2.44	.4927	.9927	.0073	.0203
2.45	.4929	.9929	.0071	.0198
2.46	.4931	.9931	.0069	.0194
2.47	.4932	.9932	.0068	.0189

Fortsetzung von Tab. II

z Standard-Wert $(\frac{x}{\sigma})$	Flächenstücke A zwischen μ und z	B im größeren Teil	C im kleineren Teil	y Ordinate bei $\frac{x}{\sigma}$
2.48	.4934	.9934	.0066	.0184
2.49	.4936	.9936	.0064	.0180
2.50	.4938	.9938	.0062	.0175
2.51	.4940	.9940	.0060	.0171
2.52	.4941	.9941	.0059	.0167
2.53	.4943	.9943	.0057	.0163
2.54	.4945	.9945	.0055	.0158
2.55	.4946	.9946	.0054	.0154
2.56	.4948	.9948	.0052	.0151
2.57	.4949	.9949	.0051	.0147
2.58	.4951	.9951	.0049	.0143
2.59	.4952	.9952	.0048	.0139
2.60	.4953	.9953	.0047	.0136
2.61	.4955	.9955	.0045	.0132
2.62	.4956	.9956	.0044	.0129
2.63	.4957	.9957	.0043	.0126
2.64	.4959	.9959	.0041	.0122
2.65	.4960	.9960	.0040	.0119
2.66	.4961	.9961	.0039	.0116
2.67	.4962	.9962	.0038	.0113
2.68	.4963	.9963	.0037	.0110
2.69	.4964	.9964	.0036	.0107
2.70	.4965	.9965	.0035	.0104
2.71	.4966	.9966	.0034	.0101
2.72	.4967	.9967	.0033	.0099
2.73	.4968	.9968	.0032	.0096
2.74	.4969	.9969	.0031	.0093
2.75	.4970	.9970	.0030	.0091
2.76	.4971	.9971	.0029	.0088
2.77	.4972	.9972	.0028	.0086
2.78	.4973	.9973	.0027	.0084
2.79	.4974	.9974	.0026	.0081
2.80	.4974	.9974	.0026	.0079
2.81	.4975	.9975	.0025	.0077
2.82	.4976	.9976	.0024	.0075
2.83	.4977	.9977	.0023	.0078
2.84	.4977	.9977	.0023	.0071
2.85	.4978	.9978	.0022	.0069
2.86	.4979	.9979	.0021	.0067
2.87	.4979	.9979	.0021	.0065

Die z-Werte für die mittleren 99% (z = −2.58 bzw. z = +2.58) markieren das sog. *1%-Niveau* zur Sicherung von Statistiken.

Fortsetzung von Tab. II

z Standard-Wert ($\frac{x}{\sigma}$)	A zwischen μ und z	Flächenstücke B im größeren Teil	C im kleineren Teil	y Ordinate bei $\frac{x}{\sigma}$
2.88	.4980	.9980	.0020	.0063
2.89	.4981	.9981	.0019	.0061
2.90	.4981	.9981	.0019	.0060
2.91	.4982	.9982	.0018	.0058
2.92	.4982	.9982	.0018	.0056
2.93	.4983	.9983	.0017	.0055
2.94	.4984	.9984	.0016	.0053
2.95	.4984	.9984	.0016	.0051
2.96	.4985	.9985	.0015	.0050
2.97	.4985	.9985	.0015	.0048
2.98	.4986	.9986	.0014	.0047
2.99	.4986	.9986	.0014	.0046
3.00	.4987	.9987	.0013	.0044

Entnommen aus: Guilford 1965, S. 570ff.

Tab. III:
Kritische Werte von t (nach Fisher & Yates) auf verschiedenen Signifikanzniveaus

df	Signifikanzniveaus bei einseitigem Test					
	.10	.05	.025	.01	.005	.0005
	Signifikanzniveaus bei zweiseitigem Test					
	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Entnommen aus: Siegel 1956, S. 248.

Tab. IV:
Kritische Werte von F (nach Fisher & Yates) für drei Signifikanzniveaus: Grenzen für p = .05 (obere Zeile), für p = .01 (mittlere Zeile), für p = .001 (untere Zeile)

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞[1]
	161	200	216	225	230	234	239	244	249	254
1	*4052*	*4999*	*5403*	*5625*	*5724*	*5859*	*5981*	*6106*	*6234*	*6366*
	405284	**500000**	**540379**	**562500**	**576405**	**585937**	**598144**	**610667**	**623497**	**636619**
	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
2	*98.49*	*99.01*	*99.17*	*99.25*	*99.30*	*99.33*	*99.36*	*99.42*	*99.46*	*99.50*
	998.5	**999.0**	**999.2**	**999.2**	**999.3**	**999.3**	**999.4**	**999.4**	**999.5**	**999.5**
	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
3	*34.12*	*30.81*	*29.46*	*28.71*	*28.24*	*27.91*	*27.49*	*27.05*	*26.60*	*26.12*
	167.5	**148.5**	**141.1**	**137.1**	**134.6**	**132.8**	**130.6**	**128.3**	**125.9**	**123.5**
	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
4	*21.20*	*18.00*	*16.69*	*15.98*	*15.52*	*15.21*	*14.80*	*14.37*	*13.93*	*13.46*
	74.14	**61.25**	**56.18**	**53.44**	**51.71**	**50.53**	**49.00**	**47.41**	**45.77**	**44.05**
	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
5	*16.26*	*13.27*	*12.06*	*11.39*	*10.97*	*10.67*	*10.27*	*9.89*	*9.47*	*9.02*
	47.04	**36.61**	**33.20**	**31.09**	**29.75**	**28.84**	**27.64**	**26.42**	**25.14**	**23.78**
	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
6	*13.74*	*10.92*	*9.78*	*9.15*	*8.75*	*8.47*	*8.10*	*7.72*	*7.31*	*6.88*
	35.51	**27.00**	**23.70**	**21.90**	**20.81**	**20.03**	**19.03**	**17.99**	**16.89**	**15.75**
	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
7	*12.25*	*9.55*	*8.45*	*7.85*	*7.46*	*7.19*	*6.84*	*6.47*	*6.07*	*5.65*
	29.22	**21.69**	**18.77**	**17.19**	**16.21**	**15.52**	**14.63**	**13.71**	**12.73**	**11.69**
	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
8	*11.26*	*8.65*	*7.59*	*7.01*	*6.63*	*6.37*	*6.03*	*5.67*	*5.28*	*4.86*
	25.42	**18.49**	**15.83**	**14.39**	**13.49**	**12.86**	**12.04**	**11.19**	**10.30**	**9.34**
	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
9	*10.56*	*8.02*	*6.99*	*6.42*	*6.06*	*5.80*	*5.47*	*5.11*	*4.73*	*4.31*
	22.86	**16.39**	**13.90**	**12.56**	**11.71**	**11.13**	**10.37**	**9.57**	**8.72**	**7.81**
	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
10	*10.04*	*7.56*	*6.55*	*5.99*	*5.64*	*5.39*	*5.06*	*4.71*	*4.33*	*3.91*
	21.04	**14.91**	**12.55**	**11.28**	**10.48**	**9.92**	**9.20**	**8.45**	**7.64**	**6.76**
	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
11	*9.65*	*7.20*	*6.22*	*5.67*	*5.32*	*5.07*	*4.74*	*4.40*	*4.02*	*3.60*
	19.69	**13.81**	**11.56**	**10.35**	**9.58**	**9.05**	**8.35**	**7.63**	**6.85**	**6.00**
	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
12	*9.33*	*6.93*	*5.95*	*5.41*	*5.06*	*4.82*	*4.50*	*4.16*	*3.78*	*3.36*
	18.64	**12.97**	**10.80**	**9.63**	**8.89**	**8.38**	**7.71**	**7.00**	**6.25**	**5.42**

Entnommen aus: Scott & Wertheimer 1962, S. 408ff.

---

1 *Anmerkung:* n_1 ist hier immer df der größeren Varianz im Zähler, also $df_1 = N_1 - 1$; n_2 ist hier immer df der kleineren Varianz im Nenner, also $df_2 = N_2 - 1$.

Fortsetzung von Tab. IV

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
13	*9.07*	*6.70*	*5.74*	*5.20*	*4.86*	*4.62*	*4.30*	*3.96*	*3.59*	*3.16*
	17.81	**12.31**	**10.21**	**9.07**	**8.35**	**7.86**	**7.21**	**6.52**	**5.78**	**4.97**
	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
14	*8.86*	*6.51*	*5.56*	*5.03*	*4.69*	*4.46*	*4.14*	*3.80*	*3.43*	*3.00*
	17.14	**11.78**	**9.73**	**8.62**	**7.92**	**7.43**	**6.80**	**6.13**	**5.41**	**4.60**
	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
15	*8.68*	*6.36*	*5.42*	*4.89*	*4.56*	*4.32*	*4.00*	*3.67*	*3.29*	*2.87*
	16.59	**11.34**	**9.34**	**8.25**	**7.57**	**7.09**	**6.47**	**5.81**	**5.10**	**4.31**
	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
16	*8.53*	*6.23*	*5.29*	*4.77*	*4.44*	*4.20*	*3.89*	*3.55*	*3.18*	*2.75*
	16.12	**10.97**	**9.00**	**7.94**	**7.27**	**6.81**	**6.19**	**5.55**	**4.85**	**4.06**
	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
17	*8.40*	*6.11*	*5.18*	*4.67*	*4.34*	*4.10*	*3.79*	*3.45*	*3.08*	*2.65*
	15.72	**10.66**	**8.73**	**7.68**	**7.02**	**6.56**	**5.96**	**5.32**	**4.63**	**3.85**
	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
18	*8.28*	*6.01*	*5.09*	*4.58*	*4.25*	*4.01*	*3.71*	*3.37*	*3.00*	*2.57*
	15.38	**10.39**	**8.49**	**7.46**	**6.81**	**6.35**	**5.76**	**5.13**	**4.45**	**3.67**
	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
19	*8.18*	*5.93*	*5.01*	*4.50*	*4.17*	*3.94*	*3.63*	*3.30*	*2.92*	*2.49*
	15.08	**10.16**	**8.28**	**7.26**	**6.61**	**6.18**	**5.59**	**4.97**	**4.29**	**3.52**
	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
20	*8.10*	*5.85*	*4.94*	*4.43*	*4.10*	*3.87*	*3.56*	*3.23*	*2.86*	*2.42*
	14.82	**9.95**	**8.10**	**7.10**	**6.46**	**6.02**	**5.44**	**4.82**	**4.15**	**3.38**
	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
21	*8.02*	*5.78*	*4.87*	*4.37*	*4.04*	*3.81*	*3.51*	*3.17*	*2.80*	*2.36*
	14.59	**9.77**	**7.94**	**6.95**	**6.32**	**5.88**	**5.31**	**4.70**	**4.03**	**3.26**
	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
22	*7.94*	*5.72*	*4.82*	*4.31*	*3.99*	*3.76*	*3.45*	*3.12*	*2.75*	*2.31*
	14.38	**9.61**	**7.80**	**6.81**	**6.19**	**5.76**	**5.19**	**4.58**	**3.92**	**3.15**
	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
23	*7.88*	*5.66*	*4.76*	*4.26*	*3.94*	*3.71*	*3.41*	*3.07*	*2.70*	*2.26*
	14.19	**9.47**	**7.67**	**6.69**	**6.08**	**5.65**	**5.09**	**4.48**	**3.82**	**3.05**
	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
24	*7.82*	*5.61*	*4.72*	*4.22*	*3.90*	*3.67*	*3.36*	*3.03*	*2.66*	*2.21*
	14.03	**9.34**	**7.55**	**6.59**	**5.98**	**5.55**	**4.99**	**4.39**	**3.74**	**2.97**

Fortsetzung von Tab. IV

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
25	*7.77*	*5.57*	*4.68*	*4.18*	*3.86*	*3.63*	*3.32*	*2.99*	*2.62*	*2.17*
	13.88	**9.22**	**7.45**	**6.49**	**5.88**	**5.46**	**4.91**	**4.31**	**3.66**	**2.89**
	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
26	*7.22*	*5.53*	*4.64*	*4.14*	*3.82*	*3.59*	*3.29*	*2.96*	*2.58*	*2.13*
	13.74	**9.12**	**7.36**	**6.41**	**5.80**	**5.38**	**4.83**	**4.24**	**3.59**	**2.82**
	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
27	*7.68*	*5.49*	*4.60*	*4.11*	*3.78*	*3.56*	*3.26*	*2.93*	*2.55*	*2.10*
	13.61	**9.02**	**7.27**	**6.33**	**5.73**	**5.31**	**4.76**	**4.17**	**3.52**	**2.75**
	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
28	*7.64*	*5.45*	*4.57*	*4.07*	*3.75*	*3.53*	*3.23*	*2.90*	*2.52*	*2.06*
	13.50	**8.93**	**7.19**	**6.25**	**5.66**	**5.24**	**4.69**	**4.11**	**3.46**	**2.70**
	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
29	*7.60*	*5.42*	*4.54*	*4.04*	*3.73*	*3.50*	*3.20*	*2.87*	*2.49*	*2.03*
	13.39	**8.85**	**7.12**	**6.19**	**5.59**	**5.18**	**4.64**	**4.05**	**3.41**	**2.64**
	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
30	*7.56*	*5.39*	*4.51*	*4.02*	*3.70*	*3.47*	*3.17*	*2.84*	*2.47*	*2.01*
	13.29	**8.77**	**7.05**	**6.12**	**5.53**	**5.12**	**4.58**	**4.00**	**3.36**	**2.59**
	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
40	*7.31*	*5.18*	*4.31*	*3.83*	*3.51*	*3.29*	*2.99*	*2.66*	*2.29*	*1.80*
	12.61	**8.25**	**6.60**	**5.70**	**5.13**	**4.73**	**4.21**	**3.64**	**3.01**	**2.23**
	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
60	*7.08*	*4.98*	*4.13*	*3.65*	*3.34*	*3.12*	*2.82*	*2.50*	*2.12*	*1.60*
	11.97	**7.76**	**6.17**	**5.31**	**4.76**	**4.37**	**3.87**	**3.31**	**2.69**	**1.90**
	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
120	*6.85*	*4.79*	*3.95*	*3.48*	*3.17*	*2.96*	*2.66*	*2.34*	*1.95*	*1.38*
	11.38	**7.31**	**5.79**	**4.95**	**4.42**	**4.04**	**3.55**	**3.02**	**2.40**	**1.56**
	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00
∞	*6.64*	*4.60*	*3.78*	*3.32*	*3.02*	*2.80*	*2.51*	*2.18*	*1.79*	*1.00*
	10.83	**6.91**	**5.42**	**4.62**	**4.10**	**3.74**	**3.27**	**2.74**	**2.13**	**1.00**

Tab. V: Kritische Werte für q auf dem 5%- bzw. 1%-Niveau

df des Nenners	$1-\alpha$	Spannweite													
		2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.95	18.0	27.0	32.8	37.1	40.4	43.1	45.4	47.4	49.1	50.6	52.0	53.2	54.3	55.4
	.99	90.0	135	164	186	202	216	227	237	246	253	260	266	272	277
2	.95	6.09	8.3	9.8	10.9	11.7	12.4	13.0	13.5	14.0	14.4	14.7	15.1	15.4	15.7
	.99	14.0	19.0	22.3	24.7	26.6	28.2	29.5	30.7	31.7	32.6	33.4	34.1	34.8	35.4
3	.95	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72	9.95	10.2	10.4	10.5
	.99	8.26	10.6	12.2	13.3	14.2	15.0	15.6	16.2	16.7	17.1	17.5	17.9	18.2	18.5
4	.95	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03	8.21	8.37	8.52	8.66
	.99	6.51	8.12	9.17	9.96	10.6	11.1	11.5	11.9	12.3	12.6	12.8	13.1	13.3	13.5
5	.95	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72
	.99	5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.2	10.5	10.7	10.9	11.1	11.2
6	.95	3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14
	.99	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.49	9.65	9.81	9.95
7	.95	3.34	4.16	4.69	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76
	.99	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12
8	.95	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48
	.99	4.74	5.63	6.20	6.63	6.96	7.24	7.47	7.68	7.87	8.03	8.18	8.31	8.44	8.55
9	.95	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74	5.87	5.98	6.09	6.19	6.28
	.99	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	7.65	7.78	7.91	8.03	8.13
10	.95	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.11
	.99	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.48	7.60	7.71	7.81
11	.95	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.99
	.99	4.39	5.14	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.26	7.36	7.46	7.56
12	.95	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51	5.62	5.71	5.80	5.88
	.99	4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26	7.36

Fortsetzung von Tab. V

df des Nenners	1 − α	Spannweite													
		2	3	4	5	6	7	8	9	10	11	12	13	14	15
13	.95	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79
	.99	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.01	7.10	7.19
14	.95	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	6.64	5.72
	.99	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77	6.87	6.96	7.05
16	.95	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59
	.99	4.13	4.78	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56	6.66	6.74	6.82
18	.95	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50
	.99	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58	6.65
20	.95	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43
	.99	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.29	6.37	6.45	6.52
24	.95	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32
	.99	3.96	4.54	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11	6.19	6.26	6.33
30	.95	2.89	3.49	3.84	4.10	4.30	4.46	4.60	4.72	4.83	4.92	5.00	5.08	5.15	5.21
	.99	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.56	5.76	5.85	5.93	6.01	6.08	6.14
40	.95	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74	4.82	4.91	4.98	5.05	5.11
	.99	3.82	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	5.69	5.77	5.84	5.90	5.96
60	.95	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00
	.99	3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60	5.67	5.73	5.79
120	.95	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.48	4.56	4.64	4.72	4.78	4.84	4.90
	.99	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.38	5.44	5.51	5.56	5.61
∞	.95	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80
	.99	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29	5.35	5.40	5.45

Entnommen aus: Winer 1962, S. 648 f.

Tab. VI:
Kritische Werte von r (Koeffizient der Produ:t-Moment-Korrelation nach Pearson)

df[1]	p = 5%	p = 1%	df	p = 5%	p = 1%
1	0.997	1.000	24	0.388	0.496
2	0.950	0.990	25	0.381	0.487
3	0.878	0.959	26	0.374	0.478
4	0.811	0.917	27	0.367	0.470
5	0.754	0.874	28	0.361	0.463
6	0.707	0.834	29	0.355	0.456
7	0.666	0.798	30	0.349	0.449
8	0.632	0.765	35	0.325	0.418
9	0.602	0.735	40	0.304	0.393
10	0.576	0.708	45	0.288	0.372
11	0.553	0.684	50	0.273	0.354
12	0.532	0.661	60	0.250	0.332
13	0.514	0.641	70	0.232	0.302
14	0.497	0.623	80	0.217	0.283
15	0.482	0.606	90	0.205	0.267
16	0.468	0.590	100	0.195	0.254
17	0.456	0.575	125	0.174	0.228
18	0.444	0.561	150	0.159	0.208
19	0.433	0.549	200	0.138	0.181
20	0.423	0.537	300	0.113	0.148
21	0.413	0.526	400	0.098	0.128
22	0.404	0.515	500	0.088	0.115
23	0.396	0.505	1000	0.062	0.081

Tab. VII:
Transformation von r nach z

r	z	r	z	r	z	r	z	r	z	r	z
.25	.26	.40	.42	.55	.62	.70	.87	.85	1.26	.950	1.83
.26	.27	.41	.44	.56	.63	.71	.89	.86	1.29	.955	1.89
.27	.28	.42	.45	.57	.65	.72	.91	.87	1.33	.960	1.95
.28	.29	.43	.46	.58	.66	.73	.93	.88	1.38	.965	2.01
.29	.30	.44	.47	.59	.68	.74	.95	.89	1.42	.970	2.09
.30	.31	.45	.48	.60	.69	.75	.97	.90	1.47	.975	2.18
.31	.32	.46	.50	.61	.71	.76	1.00	.905	1.50	.980	2.30
.32	.33	.47	.51	.62	.73	.77	1.02	.910	1.53	.985	2.44
.33	.34	.48	.52	.63	.74	.78	1.05	.915	1.56	.990	2.65
.34	.35	.49	.54	.64	.76	.79	1.07	.920	1.59	.995	2.99
.35	.37	.50	.55	.65	.78	.80	1.10	.925	1.62		
.36	.38	.51	.56	.66	.79	.81	1.13	.930	1.66		
.37	.39	.52	.58	.67	.81	.82	1.16	.935	1.70		
.38	.40	.53	.59	.68	.83	.83	1.19	.940	1.74		
.39	.41	.54	.60	.69	.85	.84	1.22	.945	1.78		

Tab. VIII:
Hilfswerte für die Zweizeilenkorrelation

(p od. q)	pq	$\sqrt{pq}$	(p od. q)	pq	$\sqrt{pq}$
.99	.0099	.0995	.74	.1924	.4386
.98	.0196	.1400	.73	.1971	.4440
.97	.0291	.1706	.72	.2016	.4490
.96	.0384	.1960	.71	.2059	.4538
.95	.0475	.2179	.70	.2100	.4583
.94	.0564	.2375	.69	.2139	.4625
.93	.0651	.2551	.68	.2176	.4665
.92	.0736	.2713	.67	.2211	.4702
.91	.0819	.2862	.66	.2244	.4737
.90	.0900	.3000	.65	.2275	.4770
.89	.0979	.3129	.64	.2304	.4800
.88	.1056	.3250	.63	.2331	.4828
.87	.1131	.3363	.62	.2356	.4854
.86	.1204	.3470	.61	.2379	.4877
.85	.1275	.3571	.60	.2400	.4899
.84	.1344	.3666	.59	.2419	.4918
.83	.1411	.3756	.58	.2436	.4936
.82	.1476	.3842	.57	.2451	.4951
.81	.1539	.3923	.56	.2464	.4964
.80	.1600	.4000	.55	.2475	.4975
.79	.1659	.4073	.54	.2484	.4984
.78	.1716	.4142	.53	.2491	.4991
.77	.1771	.4208	.52	.2496	.4996
.76	.1824	.4271	.51	.2499	.4999
.75	.1875	.4330	.50	.2500	.5000

Anmerkung: Wenn p kleiner .50, dann wird in der Tabelle der entsprechende q-Wert statt p aufgesucht. Es müßte in diesem Fall in der obersten Zeile richtig heißen q od. p statt p od. q.

Tab. IX:
Kritische Werte von rho (Koeffizient der Rangreihen-Korrelation nach Spearman) für einseitigen Test

N	Signifikanzniveau .05	Signifikanzniveau .01
4	1.000	
5	.900	1.000
6	.829	.943
7	.714	.893
8	.643	.833
9	.600	.783
10	.564	.746
12	.506	.712
14	.456	.645
16	.425	.601
18	.399	.564
20	.377	.534
22	.359	.508
24	.343	.485
26	.329	.465
28	.317	.448
30	.306	.432

Tab. X:
Kritische Werte von QUSR bei Kendall-Konkordanz-Koeffizienten

k	N					zusätzliche Werte für N = 3	
	3	4	5	6	7	k	QUSR
Werte auf dem 5%-Signifikanzniveau							
3			64.4	103.9	157.3	9	54.0
4		49.5	88.4	143.3	217.0	12	71.9
5		62.6	112.3	182.4	276.2	14	83.8
6		75.7	136.1	221.4	335.2	16	95.8
8	48.1	101.7	183.7	299.0	453.1	18	107.7
10	60.0	127.8	231.2	376.7	571.0		
15	89.8	192.9	349.8	570.5	864.9		
20	119.7	258.0	468.5	764.4	1,158.7		
Werte auf dem 1%-Signifikanzniveau							
3			75.6	122.8	185.6	9	75.9
4		61.4	109.3	176.2	265.0	12	103.5
5		80.5	142.8	229.4	343.8	14	121.9
6		99.5	176.1	282.4	422.6	16	140.2
8	66.8	137.4	242.7	388.3	579.9	18	158.6
10	85.1	175.3	309.1	494.0	737.0		
15	131.0	269.8	475.2	758.2	1,129.5		
20	177.0	364.2	641.2	1,022.2	1,521.9		

Tab. XI:
Kritische Werte der Chi²-Verteilung (χ^2-Werte)

df	Wahrscheinlichkeiten													
	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.00016	.00063	.0039	.016	.064	.15	.46	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	.02	.04	.10	.21	.45	.71	1.39	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	.12	.18	.35	.58	1.00	1.42	2.37	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	.30	.43	.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	11.67	13.28	18.46
5	.55	.75	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	.87	1.13	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	1.24	1.56	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	1.65	2.03	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	2.09	2.53	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	2.56	3.06	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	3.05	3.61	4.58	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	3.57	4.18	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	4.11	4.76	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	4.66	5.37	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	5.23	5.98	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	5.81	6.61	7.96	9.31	11.15	12.62	15.34	18.42	20.46	23.54	26.30	29.63	32.00	39.29
17	6.41	7.26	8.67	10.08	12.00	13.53	16.34	19.51	21.62	24.77	27.59	31.00	33.41	40.75
18	7.02	7.91	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	32.35	34.80	42.31
19	7.63	8.57	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	8.26	9.24	10.85	12.44	14.58	16.27	19.34	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	8.90	9.92	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	36.34	38.93	46.80
22	9.54	10.60	12.34	14.04	16.31	18.10	21.24	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	10.20	11 29	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	10.86	11.99	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	11.52	12.70	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	12.20	13.41	15.38	17.29	19.82	21.79	25.34	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	12.88	14.12	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	13.56	14.85	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	14.26	15.57	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	46.69	49.59	58.30
30	14.95	16.31	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	47.96	50.89	59.70

Entnommen aus: Siegel 1956, S. 249.

Tab. XII:
Kullback's 2 Î-Test: 2Nln N-Tabelle für Werte von N = 1 bis N = 3500

	0	1	2	3	4	5	6	7	8	9	
0	0.000	0.000	2.773	6.592	11.090	16.094	21.501	27.243	33.271	39.550	0
1	46.052	52.754	59.638	66.689	73.894	81.242	88.723	96.329	104.053	111.889	1
2	119.829	127.870	136.006	144.233	152.547	160.944	169.421	177.975	186.603	195.303	2
3	204.072	212.907	221.807	230.770	239.793	248.874	258.013	267.208	276.457	285.758	3
4	295.110	304.513	313.964	323.463	333.009	342.600	352.235	361.914	371.635	381.398	4
5	391.202	401.046	410.929	420.851	430.810	440.807	450.839	460.908	471.011	481.149	5
6	491.321	501.527	511.765	522.035	532.337	542.670	553.034	563.429	573.853	584.307	6
7	594.789	605.301	615.840	626.407	637.002	647.623	658.271	668.946	679.647	690.373	7
8	701.124	711.901	722.702	733.528	744.377	755.251	766.148	777.068	788.011	798.977	8
9	809.966	820.976	832.009	843.064	854.139	865.237	876.355	887.494	898.654	909.834	9
10	921.034	932.254	943.494	954.754	966.033	977.332	988.649	999.985	1011.340	1022.714	10
11	1034.106	1045.516	1056.944	1068.390	1079.853	1091.334	1102.833	1114.349	1125.882	1137.431	11
12	1148.998	1160.581	1172.181	1183.797	1195.430	1207.078	1218.743	1230.424	1242.120	1253.832	12
13	1265.559	1277.302	1289.060	1300.833	1312.621	1324.424	1336.242	1348.075	1359.922	1371.784	13
14	1383.660	1395.550	1407.455	1419.374	1431.306	1443.253	1455.213	1467.187	1479.175	1491.176	14
15	1503.191	1515.219	1527.260	1539.314	1551.381	1563.462	1575.555	1587.661	1599.780	1611.912	15
16	1624.056	1636.212	1648.381	1660.563	1672.756	1684.962	1697.180	1709.410	1721.652	1733.906	16
17	1746.171	1758.449	1770.738	1783.039	1795.351	1807.675	1820.010	1832.357	1844.715	1857.084	17
18	1869.464	1881.856	1894.258	1906.672	1919.096	1931.532	1943.978	1956.435	1968.902	1981.380	18
19	1993.869	2006.368	2018.878	2031.398	2043.929	2056.470	2069.021	2081.582	2094.154	2106.735	19
20	2119.327	2131.929	2144.540	2157.162	2169.793	2182.434	2195.085	2207.746	2220.416	2233.096	20
21	2245.785	2258.484	2271.193	2283.910	2296.638	2309.374	2322.120	2334.875	2347.640	2360.413	21
22	2373.196	2385.988	2398.789	2411.599	2424.417	2437.245	2450.082	2462.927	2475.782	2488.645	22
23	2501.516	2514.397	2527.286	2540.184	2553.090	2566.005	2578.929	2591.861	2604.801	2617.750	23
24	2630.707	2643.672	2656.646	2669.628	2682.618	2695.617	2708.623	2721.638	2734.661	2747.692	24
25	2760.730	2773.777	2786.832	2799.895	2812.966	2826.044	2839.131	2852.225	2865.327	2878.437	25
26	2891.554	2904.680	2917.813	2930.953	2944.101	2957.257	2970.420	2983.591	2996.769	3009.955	26
27	3023.148	3036.348	3049.556	3062.772	3075.994	3089.224	3102.461	3115.706	3128.957	3142.216	27
28	3155.482	3168.755	3182.036	3195.323	3208.617	3221.919	3235.227	3248.543	3261.865	3275.195	28
29	3288.531	3301.874	3315.224	3328.581	3341.945	3355.315	3368.693	3382.077	3395.468	3408.865	29
30	3422.269	3435.680	3449.098	3462.522	3475.953	3489.390	3502.834	3516.285	3529.741	3543.205	30

Entnommen aus: Blöschl, L., 1966. Kullbacks 2Î-Test als ökonomische Alternative zur Chi²-Probe. Psychol. Beitr., 9, S. 392ff.
(Mit freundl. Genehmigung d. Verf. u. d. Verl. A. Hain, Meisenheim a.d. Glan.)

	0	1	2	3	4	5	6	7	8	9	
31	3556.675	3570.151	3583.634	3597.123	3610.619	3624.121	3637.629	3651.144	3664.665	3678.192	31
32	3691.725	3705.265	3718.811	3732.363	3745.922	3759.486	3773.057	3786.634	3800.217	3813.806	32
33	3827.401	3841.002	3854.610	3868.223	3881.842	3895.467	3909.099	3922.736	3936.379	3950.028	33
34	3963.683	3977.344	3991.011	4004.683	4018.361	4032.046	4045.736	4059.431	4073.133	4086.840	34
35	4100.553	4114.272	4127.996	4141.726	4155.462	4169.204	4182.951	4196.703	4210.462	4224.225	35
36	4237.995	4251.770	4265.550	4279.336	4293.128	4306.925	4320.728	4334.536	4348.349	4362.168	36
37	4375.992	4389.822	4403.657	4417.498	4431.343	4445.195	4459.051	4472.913	4486.780	4500.652	37
38	4514.530	4528.413	4542.301	4556.195	4570.093	4583.997	4597.906	4611.821	4625.740	4639.665	38
39	4653.594	4667.529	4681.469	4695.414	4709.365	4723.320	4737.280	4751.245	4765.216	4779.191	39
40	4793.172	4807.157	4821.147	4835.143	4849.143	4863.149	4877.159	4891.174	4905.194	4919.219	40
41	4933.249	4947.284	4961.323	4975.368	4989.417	5003.471	5017.530	5031.594	5045.662	5059.736	41
42	5073.814	5087.897	5101.984	5116.077	5130.174	5144.276	5158.382	5172.494	5186.609	5200.730	42
43	5214.855	5228.985	5243.120	5257.259	5271.403	5285.551	5299.704	5313.862	5328.024	5342.190	43
44	5356.362	5370.538	5384.718	5398.903	5413.092	5427.286	5441.485	5455.687	5469.895	5484.107	44
45	5498.323	5512.544	5526.769	5540.998	5555.232	5569.471	5583.713	5597.961	5612.212	5626.468	45
46	5640.728	5654.993	5669.262	5683.535	5697.813	5712.095	5726.381	5740.672	5754.966	5769.265	46
47	5783.569	5797.876	5812.188	5826.504	5840.825	5855.149	5869.478	5883.811	5898.148	5912.489	47
48	5926.835	5941.184	5955.538	5969.896	5984.258	5998.624	6012.995	6027.369	6041.748	6056.131	48
49	6070.517	6084.908	6099.303	6113.702	6128.105	6142.512	6156.923	6171.338	6185.758	6200.181	49
50	6214.608	6229.039	6243.475	6257.914	6272.357	6286.804	6301.255	6315.710	6330.169	6344.632	50
51	6359.099	6373.570	6388.044	6402.523	6417.006	6431.492	6445.982	6460.476	6474.974	6489.476	51
52	6503.982	6518.492	6533.005	6547.522	6562.043	6576.568	6591.097	6605.629	6620.166	6634.706	52
53	6649.250	6663.797	6678.349	6692.904	6707.463	6722.025	6736.592	6751.162	6765.736	6780.313	53
54	6794.895	6809.480	6824.068	6838.661	6853.257	6867.857	6882.460	6897.067	6911.678	6926.292	54
55	6940.910	6955.532	6970.157	6984.786	6999.418	7014.055	7028.694	7043.338	7057.985	7072.635	55
56	7087.289	7101.947	7116.608	7131.273	7145.941	7160.613	7175.289	7189.967	7204.650	7219.336	56
57	7234.025	7248.718	7263.415	7278.115	7292.819	7307.526	7322.236	7336.950	7351.667	7366.388	57
58	7381.113	7395.840	7410.572	7425.306	7440.044	7454.786	7469.531	7484.279	7499.031	7513.786	58
59	7528.545	7543.307	7558.072	7572.841	7587.613	7602.388	7617.167	7631.949	7646.735	7661.523	59
60	7676.316	7691.111	7705.910	7720.712	7735.518	7750.326	7765.139	7779.954	7794.773	7809.595	60
61	7824.420	7839.248	7854.080	7868.915	7883.754	7898.595	7913.440	7928.288	7943.140	7957.994	61
62	7972.852	7987.713	8002.577	8017.445	8032.316	8047.190	8062.067	8076.947	8091.830	8106.717	62
63	8121.607	8136.500	8151.396	8166.296	8181.198	8196.104	8211.012	8225.925	8240.840	8255.758	63
64	8270.679	8285.604	8300.531	8315.462	8330.396	8345.333	8360.273	8375.216	8390.162	8405.112	64
65	8420.064	8435.020	8449.978	8464.940	8479.904	8494.872	8509.843	8524.817	8539.794	8554.774	65
66	8569.757	8584.743	8599.732	8614.724	8629.719	8644.717	8659.718	8674.722	8689.729	8704.739	66
67	8719.752	8734.768	8749.787	8764.809	8779.834	8794.862	8809.893	8824.927	8839.964	8855.003	67
68	8870.046	8885.092	8900.140	8915.192	8930.246	8945.304	8960.364	8975.427	8990.493	9005.562	68
69	9020.634	9035.709	9050.787	9065.868	9080.951	9096.037	9111.127	9126.219	9141.314	9156.412	69
70	9171.512	9186.616	9201.722	9216.832	9231.944	9247.059	9262.177	9277.297	9292.421	9307.547	70

71	9322.676	9337.808	9352.943	9368.081	9383.221	9398.364	9413.510	9428.659	9443.810	9458.965	71
72	9474.122	9489.282	9504.444	9519.610	9534.778	9549.949	9565.123	9580.299	9595.478	9610.660	72
73	9625.845	9641.032	9656.223	9671.416	9686.611	9701.810	9717.011	9732.215	9747.421	9762.630	73
74	9777.842	9793.057	9808.274	9823.494	9838.717	9853.942	9869.171	9884.401	9899.635	9914.871	74
75	9930.110	9945.351	9960.595	9975.842	9991.092	10006.344	10021.599	10036.856	10052.116	10067.379	75
76	10082.644	10097.912	10113.183	10128.456	10143.732	10159.010	10174.291	10189.575	10204.861	10220.150	76
77	10235.441	10250.735	10266.032	10281.331	10296.633	10311.938	10327.245	10342.554	10357.866	10373.181	77
78	10388.499	10403.818	10419.141	10434.466	10449.793	10465.123	10480.456	10495.791	10511.129	10526.469	78
79	10541.812	10557.157	10572.505	10587.856	10603.209	10618.564	10633.922	10649.282	10664.645	10680.011	79
80	10695.379	10710.749	10726.122	10741.498	10756.876	10772.256	10787.639	10803.024	10818.412	10833.803	80
81	10849.195	10864.591	10879.989	10895.389	10910.791	10926.197	10941.604	10957.014	10972.427	10987.842	81
82	11003.259	11018.679	11034.101	11049.526	11064.953	11080.283	11095.815	11111.249	11126.686	11142.125	82
83	11157.567	11173.011	11188.457	11203.906	11219.357	11234.811	11250.267	11265.725	11281.186	11296.650	83
84	11312.115	11327.583	11343.054	11358.526	11374.001	11389.479	11404.959	11420.441	11435.926	11451.412	84
85	11466.902	11482.393	11497.887	11513.384	11528.882	11544.384	11559.887	11575.393	11590.901	11606.411	85
86	11621.924	11637.439	11652.956	11668.476	11683.998	11699.522	11715.049	11730.578	11746.109	11761.642	86
87	11777.178	11792.716	11808.257	11823.799	11839.344	11854.892	11870.441	11885.993	11901.547	11917.104	87
88	11932.663	11948.224	11963.787	11979.352	11994.920	12010.490	12026.062	12041.637	12057.214	12072.793	88
89	12088.374	12103.958	12119.544	12135.132	12150.722	12166.314	12181.909	12197.506	12213.105	12228.707	89
90	12244.311	12259.916	12275.525	12291.135	12306.747	12322.362	12337.979	12353.598	12369.220	12384.843	90
91	12400.469	12416.097	12431.727	12447.360	12462.994	12478.631	12494.270	12509.911	12525.554	12541.200	91
92	12556.848	12572.497	12588.149	12603.804	12619.460	12635.118	12650.779	12666.442	12682.107	12697.774	92
93	12713.443	12729.115	12744.788	12760.464	12776.142	12791.822	12807.504	12823.188	12838.875	12854.563	93
94	12870.254	12885.947	12901.642	12917.339	12933.038	12948.740	12964.443	12980.148	12995.856	13011.566	94
95	13027.278	13042.992	13058.708	13074.426	13090.146	13105.869	13121.593	13137.320	13153.048	13168.779	95
96	13184.512	13200.247	13215.984	13231.723	13247.464	13263.207	13278.953	13294.700	13310.449	13326.201	96
97	13341.954	13357.710	13373.468	13389.227	13404.989	13420.753	13436.519	13452.287	13468.057	13483.829	97
98	13499.603	13515.379	13531.157	13546.938	13562.720	13578.504	13594.290	13610.079	13625.869	13641.661	98
99	13657.456	13673.252	13689.051	13704.851	13720.654	13736.458	13752.265	13768.073	13783.884	13799.696	99
100	13815.511	13831.327	13847.146	13862.966	13878.789	13894.613	13910.440	13926.268	13942.098	13957.931	100
101	13973.765	13989.602	14005.440	14021.280	14037.123	14052.967	14068.813	14084.662	14100.512	14116.364	101
102	14132.218	14148.074	14163.932	14179.792	14195.654	14211.518	14227.384	14243.252	14259.122	14274.993	102
103	14290.867	14306.743	14322.620	14338.500	14354.381	14370.264	14386.150	14402.037	14417.926	14433.817	103
104	14449.710	14465.605	14481.502	14497.401	14513.301	14529.204	14545.108	14561.015	14576.923	14592.833	104
105	14608.745	14624.659	14640.575	14656.493	14672.413	14688.335	14704.258	14720.184	14736.111	14752.040	105

Fortsetzung von Tab. XII

	0	1	2	3	4	5	6	7	8	9	
106	14767.971	14783.904	14799.839	14815.776	14831.715	14847.655	14863.597	14879.542	14895.488	14911.436	106
107	14927.386	14943.338	14959.291	14975.247	14991.204	15007.163	15023.124	15039.087	15055.052	15071.019	107
108	15086.987	15102.958	15118.930	15134.904	15150.880	15166.858	15182.837	15198.819	15214.802	15230.787	108
109	15246.774	15262.763	15278.753	15294.746	15310.740	15326.736	15342.734	15358.734	15374.735	15390.739	109
110	15406.744	15422.751	15438.760	15454.771	15470.783	15486.797	15502.813	15518.831	15534.851	15550.873	110
111	15566.896	15582.921	15598.948	15614.977	15631.007	15647.040	15663.074	15679.110	15695.147	15711.187	111
112	15727.228	15743.271	15759.316	15775.363	15791.411	15807.461	15823.513	15839.567	15855.622	15871.680	112
113	15887.739	15903.800	15919.862	15935.927	15951.993	15968.061	15984.130	16000.202	16016.275	16032.350	113
114	16048.426	16064.505	16080.585	16096.667	16112.751	16128.836	16144.923	16161.012	16177.103	16193.195	114
115	16209.290	16225.386	16241.483	16257.583	16273.684	16289.786	16305.891	16321.997	16338.105	16354.215	115
116	16370.327	16386.440	16402.555	16418.671	16434.790	16450.910	16467.032	16483.155	16499.281	16515.407	116
117	16531.536	16547.666	16563.799	16579.932	16596.068	16612.205	16628.344	16644.485	16660.627	16676.771	117
118	16692.917	16709.064	16725.213	16741.364	16757.516	16773.670	16789.826	16805.984	16822.143	16838.304	118
119	16854.466	16870.631	16886.797	16902.964	16919.134	16935.304	16951.477	16967.651	16983.827	17000.005	119
120	17016.184	17032.365	17048.548	17064.732	17080.918	17097.106	17113.295	17129.486	17145.679	17161.873	120
121	17178.069	17194.267	17210.466	17226.667	17242.869	17259.073	17275.279	17291.487	17307.696	17323.907	121
122	17340.119	17356.333	17372.549	17388.766	17404.985	17421.205	17437.428	17453.652	17469.877	17486.104	122
123	17502.333	17518.563	17534.795	17551.029	17567.264	17583.501	17599.739	17615.979	17632.221	17648.464	123
124	17664.709	17680.956	17697.204	17713.454	17729.705	17745.958	17762.213	17778.469	17794.727	17810.986	124
125	17827.247	17843.510	17859.774	17876.040	17892.307	17908.576	17924.847	17941.119	17957.393	17973.668	125
126	17989.945	18006.223	18022.503	18038.785	18055.068	18071.353	18087.640	18103.928	18120.217	18136.509	126
127	18152.801	18169.096	18185.392	18201.689	18217.988	18234.289	18250.591	18266.895	18283.200	18299.507	127
128	18315.815	18332.125	18348.437	18364.750	18381.065	18397.381	18413.699	18430.018	18446.339	18462.662	128
129	18478.986	18495.311	18511.638	18527.967	18544.297	18560.629	18576.962	18593.297	18609.633	18625.971	129
130	18642.311	18658.652	18674.994	18691.338	18707.684	18724.031	18740.380	18756.730	18773.082	18789.435	130
131	18805.790	18822.146	18838.504	18854.863	18871.224	18887.587	18903.951	18920.316	18936.683	18953.052	131
132	18969.422	18985.793	19002.166	19018.541	19034.917	19051.295	19067.674	19084.054	19100.436	19116.820	132
133	19133.205	19149.592	19165.980	19182.369	19198.761	19215.153	19231.547	19247.943	19264.340	19280.739	133
134	19297.139	19313.540	19329.943	19346.348	19362.754	19379.162	19395.571	19411.981	19428.393	19444.807	134
135	19461.222	19477.638	19494.056	19510.475	19526.896	19543.319	19559.743	19576.168	19592.595	19609.023	135
136	19625.453	19641.884	19658.317	19674.751	19691.186	19707.623	19724.062	19740.502	19756.944	19773.386	136
137	19789.831	19806.277	19822.724	19839.173	19855.623	19872.075	19888.528	19904.983	19921.439	19937.896	137
138	19954.355	19970.815	19987.277	20003.741	20020.205	20036.672	20053.139	20069.608	20086.079	20102.551	138
139	20119.024	20135.499	20151.975	20168.453	20184.932	20201.413	20217.895	20234.378	20250.863	20267.349	139
140	20283.837	20300.326	20316.817	20333.309	20349.802	20366.297	20382.793	20399.291	20415.790	20432.291	140

141	20448.793	20465.296	20481.801	20498.307	20514.815	20531.324	20547.834	20564.346	20580.860	20597.374	141
142	20613.891	20630.408	20646.927	20663.447	20679.969	20696.492	20713.017	20729.543	20746.070	20762.599	142
143	20779.129	20795.661	20812.194	20828.728	20845.264	20861.801	20878.339	20894.879	20911.421	20927.963	143
144	20944.507	20961.053	20977.600	20994.148	21010.698	21027.249	21043.801	21060.355	21076.910	21093.467	144
145	21110.025	21126.584	21143.145	21159.707	21176.270	21192.835	21209.401	21225.969	21242.538	21259.108	145
146	21275.680	21292.253	21308.827	21325.403	21341.980	21358.559	21375.139	21391.720	21408.303	21424.887	146
147	21441.472	21458.059	21474.647	21491.236	21507.827	21524.419	21541.013	21557.607	21574.204	21590.801	147
148	21607.400	21624.000	21640.602	21657.205	21673.809	21690.415	21707.022	21723.630	21740.240	21756.851	148
149	21773.464	21790.077	21806.692	21823.309	21839.927	21856.546	21873.166	21889.788	21906.411	21923.035	149
150	21939.661	21956.288	21972.917	21989.546	22006.178	22022.810	22039.444	22056.079	22072.715	22089.353	150
151	22105.992	22122.632	22139.274	22155.917	22172.562	22189.207	22205.854	22222.503	22239.152	22255.803	151
152	22272.455	22289.109	22305.764	22322.420	22339.078	22355.737	22372.397	22389.058	22405.721	22422.385	152
153	22439.050	22455.717	22472.385	22489.054	22505.725	22522.397	22539.070	22555.745	22572.421	22589.098	153
154	22605.776	22622.456	22639.137	22655.819	22672.503	22689.188	22705.874	22722.561	22739.250	22755.940	154
155	22772.632	22789.324	22806.018	22822.714	22839.410	22856.108	22872.807	22889.507	22906.209	22922.912	155
156	22939.616	22956.322	22973.029	22989.737	23006.446	23023.157	23039.869	23056.582	23073.296	23090.012	156
157	23106.729	23123.447	23140.167	23156.888	23173.610	23190.333	23207.058	23223.784	23240.511	23257.239	157
158	23273.969	23290.700	23307.432	23324.166	23340.901	23357.637	23374.374	23391.113	23407.853	23424.594	158
159	23441.336	23458.080	23474.824	23491.571	23508.318	23525.067	23541.816	23558.568	23575.320	23592.074	159
160	23608.829	23625.585	23642.342	23659.101	23675.861	23692.622	23709.384	23726.148	23742.913	23759.679	160
161	23776.446	23793.215	23809.984	23826.756	23843.528	23860.301	23877.076	23893.852	23910.630	23927.408	161
162	23944.188	23960.969	23977.751	23994.534	24011.319	24028.105	24044.892	24061.681	24078.470	24095.261	162
163	24112.053	24128.846	24145.641	24162.437	24179.234	24196.032	24212.831	24229.632	24246.434	24263.237	163
164	24280.041	24296.846	24313.653	24330.461	24347.270	24364.081	24380.892	24397.705	24414.519	24431.334	164
165	24448.151	24464.969	24481.787	24498.607	24515.429	24532.251	24549.075	24565.900	24582.726	24599.553	165
166	24616.382	24633.212	24650.043	24666.875	24683.708	24700.543	24717.379	24734.215	24751.054	24767.893	166
167	24784.734	24801.575	24818.418	24835.262	24852.108	24868.954	24885.802	24902.651	24919.501	24936.352	167
168	24953.205	24970.059	24986.913	25003.770	25020.627	25037.485	25054.345	25071.206	25088.068	25104.931	168
169	25121.795	25138.661	25155.528	25172.395	25189.265	25206.135	25223.006	25239.879	25256.753	25273.628	169
170	25290.504	25307.381	25324.260	25341.140	25358.020	25374.903	25391.786	25408.670	25425.556	25442.442	170
171	25459.330	25476.219	25493.110	25510.001	25526.894	25543.787	25560.682	25577.578	25594.476	25611.374	171
172	25628.274	25645.174	25662.076	25678.979	25695.884	25712.789	25729.696	25746.603	25763.512	25780.422	172
173	25797.333	25814.246	25831.159	25848.074	25864.990	25881.907	25898.825	25915.744	25932.664	25949.586	173
174	25966.509	25983.432	26000.357	26017.284	26034.211	26051.139	26068.069	26085.000	26101.932	26118.865	174
175	26135.799	26152.734	26169.670	26186.608	26203.547	26220.487	26237.428	26254.370	26271.313	26288.258	175

Fortsetzung von Tab. XII

	0	1	2	3	4	5	6	7	8	9	
176	26305. 203	26322. 150	26339. 098	26356. 047	26372. 997	26389. 948	26406. 900	26423. 854	26440. 809	26457. 764	176
177	26474. 721	26491. 679	26508. 638	26525. 599	26542. 560	26559. 523	26576. 486	26593. 451	26610. 417	26627. 384	177
178	26644. 352	26661. 322	26678. 292	26695. 264	26712. 236	26729. 210	26746. 185	26763. 161	26780. 138	26797. 116	178
179	26814. 096	26831. 076	26848. 058	26865. 041	26882. 025	26899. 009	26915. 996	26932. 983	26949. 971	26966. 960	179
180	26983. 951	27000. 943	27017. 935	27034. 929	27051. 924	27068. 920	27085. 917	27102. 916	27119. 915	27136. 916	180
181	27153. 917	27170. 920	27187. 924	27204. 929	27221. 935	27238. 942	27255. 950	27272. 959	27289. 970	27306. 981	181
182	27323. 994	27341. 008	27358. 023	27375. 039	27392. 056	27409. 074	27426. 093	27443. 113	27460. 135	27477. 157	182
183	27494. 181	27511. 205	27528. 231	27545. 258	27562. 286	27579. 315	27596. 345	27613. 376	27630. 409	27647. 442	183
184	27664. 477	27681. 512	27698. 549	27715. 587	27732. 626	27749. 665	27766. 707	27783. 749	27800. 792	27817. 836	184
185	27834. 881	27851. 928	27868. 975	27886. 024	27903. 074	27920. 124	27937. 176	27954. 229	27971. 283	27988. 338	185
186	28005. 394	28022. 451	28039. 510	28056. 569	28073. 629	28090. 691	28107. 753	28124. 817	28141. 882	28158. 948	186
187	28176. 014	28193. 082	28210. 151	28227. 221	28244. 293	28261. 365	28278. 438	28295. 512	28312. 588	28329. 664	187
188	28346. 742	28363. 820	28380. 900	28397. 981	28415. 062	28432. 145	28449. 229	28466. 314	28483. 400	28500. 487	188
189	28517. 575	28534. 665	28551. 755	28568. 846	28585. 938	28603. 032	28620. 126	28637. 222	28654. 318	28671. 416	189
190	28688. 515	28705. 615	28722. 715	28739. 817	28756. 920	28774. 024	28791. 129	28808. 235	28825. 342	28842. 450	190
191	28859. 560	28876. 670	28893. 781	28910. 893	28928. 007	28945. 121	28962. 237	28979. 353	28996. 471	29013. 589	191
192	29030. 709	29047. 830	29064. 951	29082. 074	29099. 198	29116. 323	29133. 449	29150. 576	29167. 704	29184. 833	192
193	29201. 963	29219. 094	29266. 226	29253. 359	29270. 493	29287. 628	29304. 765	29321. 902	29339. 040	29356. 179	193
194	29373. 320	29390. 461	29407. 604	29424. 747	29441. 892	29459. 037	29476. 184	29493. 331	29510. 480	29527. 629	194
195	29544. 780	29561. 932	29579. 085	29596. 238	29613. 393	29630. 549	29647. 706	29664. 863	29682. 022	29699. 182	195
196	29716. 343	29733. 505	29750. 668	29767. 832	29784. 997	29802. 163	29819. 330	29836. 498	29853. 667	29870. 837	196
197	29888. 008	29905. 180	29922. 353	29939. 527	29956. 702	29973. 879	29991. 056	30008. 234	30025. 413	30042. 593	197
198	30059. 774	30076. 957	30094. 140	30111. 324	30128. 509	30145. 696	30162. 883	30180. 071	30197. 260	30214. 451	198
199	30231. 642	30248. 834	30266. 027	30283. 222	30300. 417	30317. 613	30334. 811	30352. 009	30369. 208	30386. 409	199
200	30403. 610	30420. 812	30438. 015	30455. 220	30472. 425	30489. 631	30506. 839	30524. 047	30541. 256	30558. 467	200
201	30575. 678	30592. 890	30610. 103	30627. 318	30644. 533	30661. 749	30678. 966	30696. 185	30713. 404	30730. 624	201
202	30747. 845	30765. 067	30782. 291	30799. 515	30816. 740	30833. 966	30851. 193	30868. 421	30885. 651	30902. 881	202
203	30920. 112	30937. 344	30954. 577	30971. 811	30989. 046	31006. 282	31023. 519	31040. 757	31057. 996	31075. 236	203
204	31092. 477	31109. 719	31126. 962	31144. 205	31161. 450	31178. 696	31195. 943	31213. 191	31230. 439	31247. 689	204
205	31264. 940	31282. 191	31299. 444	31316. 698	31333. 952	31351. 208	31368. 464	31385. 722	31402. 980	31420. 240	205
206	31437. 500	31454. 762	31472. 024	31489. 288	31506. 552	31523. 817	31541. 083	31558. 351	31575. 619	31592. 888	206
207	31610. 158	31627. 429	31644. 701	31661. 974	31679. 248	31696. 523	31713. 799	31731. 076	31748. 354	31765. 633	207
208	31782. 912	31800. 193	31817. 475	31834. 757	31852. 041	31869. 326	31886. 611	31903. 898	31921. 185	31938. 473	208
209	31955. 763	31973. 053	31990. 344	32007. 637	32024. 930	32042. 224	32059. 519	32076. 815	32094. 112	32111. 410	209
210	32128. 709	32146. 009	32163. 310	32180. 611	32197. 914	32215. 281	32232. 522	32249. 828	32267. 135	32284. 442	210

211	32301.750	32319.060	32336.370	32353.681	32370.994	32388.307	32405.621	32422.936	32440.252	32457.569	211
212	32474.887	32492.205	32509.525	32526.846	32544.167	32561.490	32578.814	32596.138	32613.463	32630.790	212
213	32648.117	32665.445	32682.774	32700.105	32717.436	32734.768	32752.101	32769.434	32786.769	32804.105	213
214	32821.442	32838.779	32856.118	32873.457	32890.797	32908.139	32925.481	32942.824	32960.168	32977.513	214
215	32994.859	33012.206	33029.554	33046.903	33064.253	33081.603	33098.955	33116.307	33133.661	33151.015	215
216	33168.370	33185.726	33203.084	33220.442	33237.801	33255.161	33272.521	33289.883	33307.246	33324.609	216
217	33341.974	33359.339	33376.706	33394.073	33411.441	33428.810	33446.180	33463.551	33480.923	33498.296	217
218	33515.669	33533.044	33550.420	33567.796	33585.173	33602.552	33619.931	33637.311	33654.692	33672.074	218
219	33689.457	33706.841	33724.225	33741.611	33758.997	33776.385	33793.773	33811.162	33828.553	33845.944	219
220	33863.336	33880.728	33898.122	33915.517	33932.913	33950.309	33967.706	33985.105	34002.504	34019.904	220
221	34037.305	34054.707	34072.110	34089.514	34106.918	34124.324	34141.730	34159.138	34176.546	34193.955	221
222	34211.365	34228.776	34246.188	34263.601	34281.015	34298.429	34315.845	34333.261	34350.678	34368.097	222
223	34385.516	34402.936	34420.356	34437.778	34455.201	34472.624	34490.049	34507.474	34524.900	34542.327	232
224	34559.756	34577.184	34594.614	34612.045	34629.477	34646.909	34664.342	34681.777	34699.212	34716.648	224
225	34734.085	34751.523	34768.961	34786.401	34803.841	34821.283	34838.725	34856.168	34873.612	34891.057	225
226	34908.503	34925.949	34943.397	34960.846	34978.295	34995.745	35013.196	35030.648	35048.101	35065.555	226
227	35083.009	35100.465	35117.921	35135.379	35152.837	35170.296	35187.756	35205.216	35222.678	35240.141	227
228	35257.604	35275.068	35292.534	35310.000	35327.467	35344.934	35362.403	35379.873	35397.343	35414.814	228
229	35432.286	35449.760	35467.233	35484.708	35502.184	35519.660	35537.138	35554.616	35572.095	35589.575	229
230	35607.056	35624.538	35642.021	35659.504	35676.989	35694.474	35711.960	35729.447	35746.935	35764.423	230
231	35781.913	35799.403	35816.895	35834.387	35851.880	35869.374	35886.869	35904.364	35921.861	35939.358	231
232	35956.856	35974.355	35991.855	36009.356	36026.858	36044.360	36061.864	36079.368	36096.873	36114.379	232
233	36131.886	36149.393	36166.902	36184.411	36201.922	36219.433	36236.945	36254.457	36271.971	36289.486	233
234	36307.001	36324.517	36342.034	36359.552	36377.071	36394.591	36412.111	36429.633	36447.155	36464.678	234
235	36482.202	36499.727	36517.252	36534.779	36552.306	36569.834	36587.363	36604.893	36622.424	36639.955	235
236	36657.488	36675.021	36692.555	36710.090	36727.626	36745.162	36762.700	36780.238	36797.778	36815.318	236
237	36832.858	36850.400	36867.943	36885.486	36903.030	36920.575	36938.121	36955.668	36973.216	36990.764	237
238	37008.313	37025.864	37043.415	37060.966	37078.519	37096.072	37113.627	37131.182	37148.738	37166.295	238
239	37183.853	37201.411	37218.970	37236.531	37254.092	37271.653	37289.216	37306.780	37324.344	37341.909	293
240	37359.475	37377.042	37394.610	37412.178	37429.748	37447.318	37464.889	37482.461	37500.033	37517.607	240
241	37535.181	37552.757	37570.333	37587.909	37605.487	37623.066	37640.645	37658.225	37675.806	37693.388	241
242	37710.970	37728.554	37746.138	37763.723	37781.309	37798.896	37816.484	37834.072	37851.661	37869.251	242
243	37886.842	37904.434	37922.026	37939.620	37957.214	37974.809	37992.405	38010.001	38027.599	38045.197	243
244	38062.796	38080.396	38097.997	38115.598	38133.201	38150.804	38168.408	38186.013	38203.618	38221.225	244
245	38238.832	38256.440	38274.049	38291.659	38309.269	38326.881	38344.493	38362.106	38379.720	38397.334	245

Fortsetzung von Tab. XII

	0	1	2	3	4	5	6	7	8	9	
246	38414.950	38432.566	38450.183	38467.801	38485.420	38503.039	38520.659	38538.281	38555.902	38573.525	246
247	38591.149	38608.773	38626.398	38644.024	38661.651	38679.279	38696.907	38714.536	38732.166	38749.797	247
248	38767.429	38785.061	38802.694	38820.328	38837.963	38855.599	38873.235	38890.873	38908.511	38926.149	248
249	38943.789	38961.430	38979.071	38996.713	39014.356	39032.000	39049.644	39067.289	39084.935	39102.582	249
250	39120.230	39137.879	39155.528	39173.178	39190.829	39208.480	39226.133	39243.786	39261.440	39279.095	250
251	39296.751	39314.407	39332.065	39349.723	39367.382	39385.041	39402.702	39420.363	39438.025	39455.688	251
252	39473.351	39491.016	39508.681	39526.347	39544.014	39561.682	39579.350	39597.019	39614.689	39632.360	252
253	39650.031	39667.704	39685.377	39703.051	39720.725	39738.401	39756.077	39773.754	39791.432	39809.111	253
254	39826.790	39844.471	39862.152	39879.833	39897.516	39915.199	39932.884	39950.568	39968.254	39985.941	254
255	40003.628	40021.316	40039.005	40056.695	40074.385	40092.076	40109.768	40127.461	40145.155	40162.849	255
256	40180.544	40198.240	40215.937	40233.634	40251.333	40269.032	40286.731	40304.432	40322.133	40339.836	256
257	40357.538	40375.242	40392.947	40410.652	40428.358	40446.065	40463.772	40481.481	40499.190	40516.900	257
258	40534.611	40552.322	40570.034	40587.747	40605.461	40623.176	40640.891	40658.607	40676.324	40694.042	258
259	40711.760	40729.479	40747.199	40764.920	40782.642	40800.364	40818.087	40835.811	40853.535	40871.261	259
260	40888.987	40906.714	40924.442	40942.170	40959.899	40977.629	40995.360	41013.092	41030.824	41048.557	260
261	41066.291	41084.025	41101.761	41119.497	41137.234	41154.971	41172.710	41190.449	41208.189	41225.930	261
262	41243.671	41261.413	41279.156	41296.900	41314.645	41332.390	41350.136	41367.883	41385.630	41403.379	262
263	41421.128	41438.878	41456.628	41474.380	41492.132	41509.885	41527.638	41545.393	41563.148	41580.904	263
264	41598.661	41616.418	41634.176	41651.935	41669.695	41687.455	41705.217	41722.979	41740.741	41758.505	264
265	41776.269	41794.034	41811.800	41829.566	41847.334	41865.102	41882.870	41900.640	41918.410	41936.181	265
266	41953.953	41971.726	41989.499	42007.273	42025.048	42042.823	42060.600	42078.377	42096.154	42113.933	266
267	42131.712	42149.492	42167.273	42185.055	42202.837	42220.620	42238.404	42256.188	42273.973	42291.760	267
268	42309.546	42327.334	42345.122	42362.911	42380.701	42398.491	42416.283	42434.075	42451.867	42469.661	268
269	42487.455	42505.250	42523.046	42540.842	42558.639	42576.437	42594.236	42612.035	42629.836	42647.636	269
270	42665.438	42683.240	42701.044	42718.847	42736.652	42754.457	42772.263	42790.070	42807.878	42825.686	270
271	42843.495	42861.305	42879.115	42896.927	42914.739	42932.551	42950.365	42968.179	42985.994	43003.810	271
272	43021.626	43039.443	43057.261	43075.080	43092.899	43110.719	43128.540	43146.362	43164.184	43182.007	272
273	43199.831	43217.655	43235.480	43253.306	43271.133	43288.960	43306.788	43324.617	43342.447	43360.277	273
274	43378.108	43395.940	43413.773	43431.606	43449.440	43467.275	43485.110	43502.946	43520.783	43538.621	274
275	43556.459	43574.298	43592.138	43609.978	43627.820	43645.662	43663.504	43681.348	43699.192	43717.037	275
276	43734.882	43752.729	43770.576	43788.424	43806.272	43824.121	43841.971	43859.822	43877.673	43895.526	276
277	43913.378	43931.232	43949.086	43966.941	43984.797	44002.653	44020.511	44038.368	44056.227	44074.086	277
278	44091.946	44109.807	44127.669	44145.531	44163.394	44181.258	44199.122	44216.987	44234.853	44252.719	278
279	44270.587	44288.454	44306.323	44324.193	44342.063	44359.933	44377.805	44395.677	44413.550	44431.424	279
280	44449.298	44467.173	44485.049	44502.926	44520.803	44538.681	44556.560	44574.439	44592.319	44610.200	280

281	44628.081	44645.964	44663.847	44681.730	44699.615	44717.500	44735.386	44753.272	44771.159	44789.047	281
282	44806.936	44824.825	44842.715	44860.606	44878.497	44896.390	44914.282	44932.176	44950.070	44967.965	282
283	44985.861	45003.757	45021.655	45039.552	45057.451	45075.350	45093.250	45111.151	45129.052	45146.954	283
284	45164.857	45182.760	45200.665	45218.569	45236.475	45254.381	45272.288	45290.196	45308.104	45326.014	284
285	45343.923	45361.834	45379.745	45397.657	45415.570	45433.483	45451.397	45469.312	45487.227	45505.143	285
286	45523.060	45540.977	45558.896	45576.814	45594.734	45612.654	45630.575	45648.497	45666.419	45684.343	286
287	45702.266	45720.191	45738.116	45756.042	45773.968	45791.896	45809.824	45827.752	45845.682	45863.612	287
288	45881.542	45899.474	45917.406	45935.339	45953.272	45971.207	45989.142	46007.077	46025.013	46042.950	288
289	46060.888	46078.826	46096.765	46114.705	46132.646	46150.587	46168.529	46186.471	46204.414	46222.358	289
290	46240.303	46258.248	46276.194	46294.141	46312.088	46330.036	46347.985	46365.934	46383.884	46401.835	290
291	46419.787	46437.739	46455.692	46473.645	46491.599	46509.554	46527.510	46545.466	46563.423	46581.381	291
292	46599.339	46617.298	46635.258	46653.218	46671.179	46689.141	46707.104	46725.067	46743.030	46760.995	292
293	46778.960	46796.926	46814.893	46832.860	46850.828	46868.796	46886.765	46904.735	46922.706	46940.677	293
294	46958.649	46976.622	46994.595	47012.569	47030.544	47048.519	47066.496	47084.472	47102.450	47120.428	294
295	47138.407	47156.386	47174.366	47192.347	47210.329	47228.311	47246.294	47264.277	47282.261	47300.246	295
296	47318.232	47336.218	47354.205	47372.192	47390.181	47408.170	47426.159	47444.149	47462.140	47480.132	296
297	47498.124	47516.117	47534.111	47552.105	47570.100	47588.096	47606.092	47624.089	47642.087	47660.085	297
298	47678.084	47696.084	47714.084	47732.085	47750.087	47768.089	47786.093	47804.096	47822.101	47840.106	298
299	47858.111	47876.118	47894.125	47912.133	47930.141	47948.150	47966.160	47984.170	48002.181	48020.193	299
300	48038.205	48056.218	48074.232	48092.247	48110.262	48128.277	48146.294	48164.311	48182.329	48200.347	300
301	48218.366	48236.386	48254.406	48272.427	48290.449	48308.471	48326.494	48344.518	48362.542	48380.567	301
302	48398.593	48416.619	48434.646	48452.674	48470.703	48488.732	48506.761	48524.792	48542.822	48560.854	302
303	48578.886	48596.919	48614.953	48632.987	48651.022	48669.058	48687.094	48705.131	48723.169	48741.207	303
304	48759.246	48777.285	48795.326	48813.366	48831.408	48849.450	48867.493	48885.536	48903.581	48921.625	304
305	48939.671	48957.717	48975.764	48993.811	49011.859	49029.908	49047.957	49066.007	49084.058	49102.110	305
306	49120.162	49138.214	49156.268	49174.322	49192.376	49210.431	49228.487	49246.544	49264.601	49282.659	306
307	49300.718	49318.777	49336.837	49354.897	49372.958	49391.020	49409.083	49427.146	49445.209	49463.274	307
308	49481.339	49499.404	49517.471	49535.538	49553.605	49571.674	49589.743	49607.812	49625.883	49643.953	308
309	49662.025	49680.097	49698.170	49716.243	49734.318	49752.392	49770.468	49788.544	49806.620	49824.698	309
310	49842.776	49860.854	49878.934	49897.014	49915.094	49933.175	49951.257	49969.340	49987.423	50005.507	310
311	50023.591	50041.676	50059.762	50077.848	50095.935	50114.023	50132.111	50150.200	50168.290	50186.380	311
312	50204.471	50222.562	50240.654	50258.747	50276.841	50294.935	50313.029	50331.125	50349.221	50367.317	312
313	50385.415	50403.513	50421.611	50439.710	50457.810	50475.910	50494.012	50512.113	50530.216	50548.319	313
314	50566.422	50584.527	50602.631	50620.737	50638.843	50656.950	50675.058	50693.166	50711.274	50729.384	314
315	50747.494	50765.604	50783.716	50801.827	50819.940	50838.053	50856.167	50874.281	50892.397	50910.512	315

Fortsetzung von Tab. XII

	0	1	2	3	4	5	6	7	8	9	
316	50928. 629	50946. 746	50964. 863	50982. 981	51001. 100	51019. 220	51037. 340	51055. 461	51073. 582	51091. 704	316
317	51109. 827	51127. 959	51146. 074	51164. 198	51182. 324	51200. 449	51218. 576	51236. 703	51254. 831	51272. 959	317
318	51291. 088	51309. 218	51327. 348	51345. 479	51363. 610	51381. 742	51399. 875	51418. 008	51436. 142	51454. 277	318
319	51472. 412	51490. 548	51508. 684	51526. 822	51544. 959	51563. 098	51581. 237	51599. 376	51617. 517	51635. 657	319
320	51653. 799	51671. 941	51690. 084	51708. 227	51726. 371	51744. 516	51762. 661	51780. 807	51798. 953	51817. 101	320
321	51835. 248	51853. 397	51871. 546	51889. 695	51907. 845	51925. 996	51944. 148	51962. 300	51980. 453	51998 606	321
322	52016. 760	52034. 915	52053. 070	52071. 226	52089. 382	52107. 539	52125. 697	52143. 855	52162. 014	52180. 174	322
323	52198. 334	52216. 494	52234. 656	52252. 818	52270. 981	52289. 144	52307. 308	52325. 472	52343. 637	52361. 803	323
324	52379. 969	52398. 136	52416. 304	52434. 472	52452. 641	52470. 810	52488. 980	52507. 151	52525. 322	52543. 494	324
325	52561. 667	52579. 840	52598. 014	52616. 188	52634. 363	52652. 539	52670. 715	52688. 892	52707. 069	52725. 247	325
326	52743. 426	52761. 605	52779. 785	52797. 965	52816. 146	52834. 328	52852. 511	52870. 693	52888. 877	52907. 061	326
327	52925. 246	52943. 431	52961. 617	52979. 804	52997. 991	53016. 179	53034. 368	53052. 557	53070. 746	53088. 937	327
328	53107. 127	53125. 319	53143. 511	53161. 704	53179. 897	53198. 091	53216. 286	53234. 481	53252. 677	53270. 873	328
329	53289. 070	53307. 267	53325. 466	53343. 664	53361. 864	53380. 064	53398. 265	53416. 466	53434. 668	53452. 870	329
330	53471. 073	53489. 277	53507. 481	53525. 686	53543. 891	53562. 097	53580. 304	53598. 511	53616. 719	53634. 928	330
331	53653. 137	53671. 347	53689. 557	53707. 768	53725. 979	53744. 192	53762. 404	53780. 618	53798. 832	53817. 046	331
332	53835. 261	53853. 477	53871. 693	53889. 910	53908. 128	53926. 346	53944. 565	53962. 784	53981. 004	53999. 225	332
333	54017. 446	54035. 667	54053. 890	54072. 113	54090. 336	54108. 560	54126. 785	54145. 011	54163. 237	54181. 463	333
334	54199. 690	54217. 918	54236. 146	54254. 375	54272. 605	54290. 835	54309. 066	54327. 297	54345. 529	54363. 762	334
335	54381. 995	54400. 228	54418. 463	54436. 698	54454. 933	54473. 169	54491. 406	54509. 643	54527. 881	54546. 120	335
336	54564. 359	54582. 598	54600. 839	54619. 080	54637. 321	54655. 563	54673. 806	54692. 049	54710. 293	54728. 537	336
337	54746. 782	54765. 028	54783. 274	54801. 521	54819. 769	54838. 017	54856. 265	54874. 514	54892. 764	54911. 014	337
338	54929. 265	54947. 517	54965. 769	54984. 022	55002. 275	55020. 529	55038. 784	55057. 039	55075. 294	55093. 551	338
339	55111. 808	55130. 065	55148. 323	55166. 582	55184. 841	55203. 101	55221. 361	55239. 622	55257. 884	55276. 146	339
340	55294. 409	55312. 672	55330. 936	55349. 201	55367. 466	55385. 731	55403. 998	55422. 265	55440. 532	55458. 800	340
341	55477. 069	55495. 338	55513. 608	55531. 878	55550. 149	55568. 421	55586. 693	55604. 966	55623. 239	55641. 513	341
342	55659. 787	55678. 063	55696. 338	55714. 614	55732. 891	55751. 169	55769. 447	55787. 725	55806. 004	55824. 284	342
343	55842. 565	55860. 846	55879. 127	55897. 409	55915. 692	55933. 975	55952. 259	55970. 543	55988. 828	56007. 114	343
344	56025. 400	56043. 687	56061. 974	56080. 262	56098. 550	56116. 840	56135. 129	56153. 419	56171. 710	56190. 002	344
345	56208. 294	56226. 586	56244. 879	56263. 173	56281. 467	56299. 762	56318. 058	56336. 354	56354. 650	56372. 947	345
346	56391. 245	56409. 543	56427. 842	56446. 142	56464. 442	56482. 743	56501. 044	56519. 346	56537. 648	56555. 951	346
347	56574. 254	56592. 559	56610. 863	56629. 169	56647. 474	56665. 781	56684. 088	56702. 395	56720. 703	56739. 012	347
348	56757. 321	56775. 631	56793. 942	56812. 253	56830. 564	56848. 877	56867. 189	56885. 503	56903. 816	56922. 131	348
349	56940. 446	56958. 762	56977. 078	56995. 394	57013. 712	57032. 030	57050. 348	57068. 667	57086. 987	57105. 307	349
350	57123. 628	57141. 949	57160. 271	57178. 593	57196. 916	57215. 240	57233. 564	57251. 889	57270. 214	57288. 540	350

Tab. XIIIa:
Auftretenswahrscheinlichkeiten von U im Mann-Whitney-Test

$n_2 = 3$

U \ n_1	1	2	3
0	.250	.100	.050
1	.500	.200	.100
2	.750	.400	.200
3		.600	.350
4			.500
5			.650

$n_2 = 4$

U \ n_1	1	2	3	4
0	.200	.067	.028	.014
1	.400	.133	.057	.029
2	.600	.267	.114	.057
3		.400	.200	.100
4		.600	.314	.171
5			.429	.243
6			.571	.343
7				.443
8				.557

$n_2 = 5$

U \ n_1	1	2	3	4	5
0	.167	.047	.018	.008	.004
1	.333	.095	.036	.016	.008
2	.500	.190	.071	.032	.016
3	.667	.286	.125	.056	.028
4		.429	.196	.095	.048
5		.571	.286	.143	.075
6			.393	.206	.111
7			.500	.278	.155
8			.607	.365	.210
9				.452	.274
10				.548	.345
11					.421
12					.500
13					.579

$n_2 = 6$

U \ n_1	1	2	3	4	5	6
0	.143	.036	.012	.005	.002	.001
1	.286	.071	.024	.010	.004	.002
2	.428	.143	.048	.019	.009	.004
3	.571	.214	.083	.033	.015	.008
4		.321	.131	.057	.026	.013
5		.429	.190	.086	.041	.021
6		.571	.274	.129	.063	.032
7			.357	.176	.089	.047
8			.452	.238	.123	.066
9			.548	.305	.165	.090
10				.381	.214	.120
11				.457	.268	.155
12				.545	.331	.197
13					.396	.242
14					.465	.294
15					.535	.350
16						.409
17						.469
18						.531

Entnommen aus: Siegel 1956, S. 271 ff.

Fortsetzung von Tab. XIIIa

$n_2 = 7$

U \ n_1	1	2	3	4	5	6	7
0	.125	.028	.008	.003	.001	.001	.000
1	.250	.056	.017	.006	.003	.001	.001
2	.375	.111	.033	.012	.005	.002	.001
3	.500	.167	.058	.021	.009	.004	.002
4	.625	.250	.092	.036	.015	.007	.003
5		.333	.133	.055	.024	.011	.006
6		.444	.192	.082	.037	.017	.009
7		.556	.258	.115	.053	.026	.013
8			.333	.158	.074	.037	.019
9			.417	.206	.101	.051	.027
10			.500	.264	.134	.069	.036
11			.583	.324	.172	.090	.049
12				.394	.216	.117	.064
13				.464	.265	.147	.082
14				.538	.319	.183	.104
15					.378	.223	.130
16					.438	.267	.159
17					.500	.314	.191
18					.562	.365	.228
19						.418	.267
20						.473	.310
21						.527	.355
22							.402
23							.451
24							.500
25							.549

Fortsetzung von Tab. XIIIa

$n_2 = 8$

U \ n_1	1	2	3	4	5	6	7	8	t	Normal
0	.111	.022	.006	.002	.001	.000	.000	.000	3.308	.001
1	.222	.044	.012	.004	.002	.001	.000	.000	3.203	.001
2	.333	.089	.024	.008	.003	.001	.001	.000	3.098	.001
3	.444	.133	.042	.014	.005	.002	.001	.001	2.993	.001
4	.556	.200	.067	.024	.009	.004	.002	.001	2.888	.002
5		.267	.097	.036	.015	.006	.003	.001	2.783	.003
6		.356	.139	.055	.023	.010	.005	.002	2.678	.004
7		.444	.188	.077	.033	.015	.007	.003	2.573	.005
8		.556	.248	.107	.047	.021	.010	.005	2.468	.007
9			.315	.141	.064	.030	.014	.007	2.363	.009
10			.387	.184	.085	.041	.020	.010	2.258	.012
11			.461	.230	.111	.054	.027	.014	2.153	.016
12			.539	.285	.142	.071	.036	.019	2.048	.020
13				.341	.177	.091	.047	.025	1.943	.026
14				.404	.217	.114	.060	.032	1.838	.033
15				.467	.262	.141	.076	.041	1.733	.041
16				.533	.311	.172	.095	.052	1.628	.052
17					.362	.207	.116	.065	1.523	.064
18					.416	.245	.140	.080	1.418	.078
19					.472	.286	.168	.097	1.313	.094
20					.528	.331	.198	.117	1.208	.113
21						.377	.232	.139	1.102	.135
22						.426	.268	.164	.998	.159
23						.475	.306	.191	.893	.185
24						.525	.347	.221	.788	.215
25							.389	.253	.683	.247
26							.433	.287	.578	.282
27							.478	.323	.473	.318
28							.522	.360	.368	.356
29								.399	.263	.396
30								.439	.158	.437
31								.480	.052	.481
32								.520		

Tab. XIIIb:
Kritische Werte von U im Mann-Whitney-Test

Einseitige Fragestellung: p = .001 Zweiseitige Fragestellung: p = .002

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2												
3									0	0	0	0
4		0	0	0	1	1	1	2	2	3	3	3
5	1	1	2	2	3	3	4	5	5	6	7	7
6	2	3	4	4	5	6	7	8	9	10	11	12
7	3	5	6	7	8	9	10	11	13	14	15	16
8	5	6	8	9	11	12	14	15	17	18	20	21
9	7	8	10	12	14	15	17	19	21	23	25	26
10	8	10	12	14	17	19	21	23	25	27	29	32
11	10	12	15	17	20	22	24	27	29	32	34	37
12	12	14	17	20	23	25	28	31	34	37	40	42
13	14	17	20	23	26	29	32	35	38	42	45	48
14	15	19	22	25	29	32	36	39	43	46	50	54
15	17	21	24	28	32	36	40	43	47	51	55	59
16	19	23	27	31	35	39	43	48	52	56	60	65
17	21	25	29	34	38	43	47	52	57	61	66	70
18	23	27	32	37	42	46	51	56	61	66	71	76
19	25	29	34	40	45	50	55	60	66	71	77	82
20	26	32	37	42	48	54	59	65	70	76	82	88

Einseitige Fragestellung: p = .01 Zweiseitige Fragestellung: p = .02

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2					0	0	0	0	0	0	1	1
3	1	1	1	2	2	2	3	3	4	4	4	5
4	3	3	4	5	5	6	7	7	8	9	9	10
5	5	6	7	8	9	10	11	12	13	14	15	16
6	7	8	9	11	12	13	15	16	18	19	20	22
7	9	11	12	14	16	17	19	21	23	24	26	28
8	11	13	15	17	20	22	24	26	28	30	32	34
9	14	16	18	21	23	26	28	31	33	36	38	40
10	16	19	22	24	27	30	33	36	38	41	44	47
11	18	22	25	28	31	34	37	41	44	47	50	53
12	21	24	28	31	35	38	42	46	49	53	56	60
13	23	27	31	35	39	43	47	51	55	59	63	67
14	26	30	34	38	43	47	51	56	60	65	69	73
15	28	33	37	42	47	51	56	61	66	70	75	80
16	31	36	41	46	51	56	61	66	71	76	82	87
17	33	38	44	49	55	60	66	71	77	82	88	93
18	36	41	47	53	59	65	70	76	82	88	94	100
19	38	44	50	56	63	69	75	82	88	94	101	107
20	40	47	53	60	67	73	80	87	93	100	107	114

Entnommen aus: Siegel 1956, S. 274ff.

Fortsetzung von Tab. XIIIb

Einseitige Fragestellung: $p = .025$ Zweiseitige Fragestellung: $p = .05$

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

Einseitige Fragestellung: $p = .05$ Zweiseitige Fragestellung: $p = .10$

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

Tab. XIV:
Auftretenswahrscheinlichkeiten von H in der einfachen Rangvarianzanalyse von Kruskal & Wallis

Samplegrößen			H	p
n_1	n_2	n_3		
2	1	1	2.7000	.500
2	2	1	3.6000	.200
2	2	2	4.5714	.067
			3.7143	.200
3	1	1	3.2000	.300
3	2	1	4.2857	.100
			3.8571	.133
3	2	2	5.3572	.029
			4.7143	.048
			4.5000	.067
			4.4643	.105
3	3	1	5.1429	.043
			4.5714	.100
			4.0000	.129
3	3	2	6.2500	.011
			5.3611	.032
			5.1389	.061
			4.5556	.100
			4.2500	.121
3	3	3	7.2000	.004
			6.4889	.011
			5.6889	.029
			5.6000	.050
			5.0667	.086
			4.6222	.100
4	1	1	3.5714	.200
4	2	1	4.8214	.057
			4.5000	.076
			4.0179	.114
4	2	2	6.0000	.014
			5.3333	.033
			5.1250	.052
			4.4583	.100
			4.1667	.105
4	3	1	5.8333	.021
			5.2083	.050
			5.0000	.057
			4.0556	.093
			3.8889	.129
4	3	2	6.4444	.008
			6.3000	.011
			5.4444	.046
			5.4000	.051
			4.5111	.098
			4.4444	.102
4	3	3	6.7455	.010
			6.7091	.013
			5.7909	.046
			5.7273	.050
			4.7091	.092
			4.7000	.101
4	4	1	6.6667	.010
			6.1667	.022
			4.9667	.048
			4.8667	.054
			4.1667	.082
			4.0667	.102
4	4	2	7.0364	.006
			6.8727	.011
			5.4545	.046
			5.2364	.052
			4.5545	.098
			4.4455	.103
4	4	3	7.1439	.010
			7.1364	.011
			5.5985	.049
			5.5758	.051
			4.5455	.099
			4.4773	.102
4	4	4	7.6538	.008
			7.5385	.011
			5.6923	.049
			5.6538	.054
			4.6539	.097
			4.5001	.104
5	1	1	3.8571	.143
5	2	1	5.2500	.036
			5.0000	.048
			4.4500	.071
			4.2000	.095
			4.0500	.119

Samplegrößen			H	p
n_1	n_2	n_3		
5	2	2	6.5333	.008
			6.1333	.013
			5.1600	.034
			5.0400	.056
			4.3733	.090
			4.2933	.122
5	3	1	6.4000	.012
			4.9600	.048
			4.8711	.052
			4.0178	.095
			3.8400	.123
5	3	2	6.9091	.009
			6.8218	.010
			5.2509	.049
			5.1055	.052
			4.6509	.091
			4.4945	.101
5	3	3	7.0788	.009
			6.9818	.011
			5.6485	.049
			5.5152	.051
			4.5333	.097
			4.4121	.109
5	4	1	6.9545	.008
			6.8400	.011
			4.9855	.044
			4.8600	.056
			3.9873	.098
			3.9600	.102
5	4	2	7.2045	.009
			7.1182	.010
			5.2727	.049
			5.2682	.050
			4.5409	.098
			4.5182	.101
5	4	3	7.4449	.010
			7.3949	.011
			5.6564	.049
			5.6308	.050
			4.5487	.099
			4.5231	.103
5	4	4	7.7604	.009
			7.7440	.011
			5.6571	.049
			5.6176	.050
			4.6187	.100
			4.5527	.102
5	5	1	7.3091	.009
			6.8364	.011
			5.1273	.046
			4.9091	.053
			4.1091	.086
			4.0364	.105
5	5	2	7.3385	.010
			7.2692	.010
			5.3385	.047
			5.2462	.051
			4.6231	.097
			4.5077	.100
5	5	3	7.5780	.010
			7.5429	.010
			5.7055	.046
			5.6264	.051
			4.5451	.100
			4.5363	.102
5	5	4	7.8229	.010
			7.7914	.010
			5.6657	.049
			5.6429	.050
			4.5229	.099
			4.5200	.101
5	5	5	8.0000	.009
			7.9800	.010
			5.7800	.049
			5.6600	.051
			4.5600	.100
			4.5000	.102

Entnommen aus: Siegel 1956, S. 282f.

Tab. XV:
Kritische Werte von T im Wilcoxon-Test unter Berücksichtigung verschiedener Signifikanzniveaus und ein- oder zweiseitiger Fragestellung

N	Signifikanzniveaus für einseitigen Test		
	.025	.01	.005
	Signifikanzniveaus für zweiseitigen Test		
	.05	.02	.01
6	0	—	—
7	2	0	—
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	15	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Tab. XVIa:
Friedman-Rangvarianzanalyse: Wahrscheinlichkeiten ausgewählter χ_r^2-Werte für k = 3 Bedingungen

N = 2		N = 3		N = 4		N = 5	
χ_r^2	p	χ_r^2	p	χ_r^2	p	χ_r^2	p
0	1.000	.000	1.000	.0	1.000	.0	1.000
1	.833	.667	.944	.5	.931	.4	.954
3	.500	2.000	.528	1.5	.653	1.2	.691
4	.167	2.667	.361	2.0	.431	1.6	.522
		4.667	.194	3.5	.273	2.8	.367
		6.000	.028	4.5	.125	3.6	.182
				6.0	.069	4.8	.124
				6.5	.042	5.2	.093
				8.0	.0046	6.4	.039
						7.6	.024
						8.4	.0085
						10.0	.00077

N = 6		N = 7		N = 8		N = 9	
χ_r^2	p	χ_r^2	p	χ_r^2	p	χ_r^2	p
.00	1.000	.000	1.000	.00	1.000	.000	1.000
.33	.956	.286	.964	.25	.967	.222	.971
1.00	.740	.857	.768	.75	.794	.667	.814
1.33	.570	1.143	.620	1.00	.654	.889	.865
2.33	.430	2.000	.486	1.75	.531	1.556	.569
3.00	.252	2.571	.305	2.25	.355	2.000	.398
4.00	.184	3.429	.237	3.00	.285	2.667	.328
4.33	.142	3.714	.192	3.25	.236	2.889	.278
5.33	.072	4.571	.112	4.00	.149	3.556	.187
6.33	.052	5.429	.085	4.75	.120	4.222	.154
7.00	.029	6.000	.052	5.25	.079	4.667	.107
8.33	.012	7.143	.027	6.25	.047	5.556	.069
9.00	.0081	7.714	.021	6.75	.038	6.000	.057
9.33	.0055	8.000	.016	7.00	.030	6.222	.048
10.33	.0017	8.857	.0084	7.75	.018	6.889	.031
12.00	.00013	10.286	.0036	9.00	.0099	8.000	.019
		10.571	.0027	9.25	.0080	8.222	.016
		11.143	.0012	9.75	.0048	8.667	.010
		12.286	.00032	10.75	.0024	9.556	.0060
		14.000	.000021	12.00	.0011	10.667	.0035
				12.25	.00086	10.889	.0029
				13.00	.00026	11.556	.0013
				14.25	.000061	12.667	.00066
				16.00	.0000036	13.556	.00035
						14.000	.00020
						14.222	.000097
						14.889	.000054
						16.222	.000011
						18.000	.0000006

Entnommen aus: Siegel 1956, S. 280f.

Tab. XVIb:
Friedman-Rangvarianzanalyse: Wahrscheinlichkeiten ausgewählter χ_r^2-Werte für k = 4 Bedingungen

N = 2		N = 3		N = 4			
χ_r^2	p	χ_r^2	p	χ_r^2	p	χ_r^2	p
.0	1.000	.2	1.000	.0	1.000	5.7	.141
.6	.958	.6	.958	.3	.992	6.0	.105
1.2	.834	1.0	.910	.6	.928	6.3	.094
1.8	.792	1.8	.727	.9	.900	6.6	.077
2.4	.625	2.2	.608	1.2	.800	6.9	.068
3.0	.542	2.6	.524	1.5	.754	7.2	.054
3.6	.458	3.4	.446	1.8	.677	7.5	.052
4.2	.375	3.8	.342	2.1	.649	7.8	.036
4.8	.208	4.2	.300	2.4	.524	8.1	.033
5.4	.167	5.0	.207	2.7	.508	8.4	.019
6.0	.042	5.4	.175	3.0	.432	8.7	.014
		5.8	.148	3.3	.389	9.3	.012
		6.6	.075	3.6	.355	9.6	.0069
		7.0	.054	3.9	.324	9.9	.0062
		7.4	.033	4.5	.242	10.2	.0027
		8.2	.017	4.8	.200	10.8	.0016
		9.0	.0017	5.1	.190	11.1	.00094
				5.4	.158	12.0	.000072

Anhang 3

Übersichtstafeln zu den behandelten Methoden der Forschungsstatistik

Aussagen über Zusammenhänge: *Korrelation*

	*Intervall*skala	*Ordinal*skala	*Nominal*skala
Intervall-skala	Produkt-Moment-Korrelation (s. S. 123ff.) Partial-Korrelation (s. S. 130ff.)		
Ordinal-skala		Rangreihen-Korrelation (s. S. 137ff.) Konkordanz-Koeffizient (s. S. 140ff.)	
Nominal-skala	Zweizeilen-Korrelation (punktbiseriale, biseriale) (s. S. 133ff.)	Biseriale Rangkorrelation (s. S. 143ff.)	Punkt-Vierfelder-Korrelation (s. S. 146ff.) Kontingenz-Koeffizient (s. S. 148ff.) Cramér-Koeffizient (s. S. 151ff.)

Voraussagen: *Regression*

	Intervallskalenniveau	
	Voraussagen aufgrund *einer* Variablen	Voraussagen aufgrund *mehrerer* Variablen
Intervall-skalenniveau	einfache, lineare Regression (s. S. 153ff.)	multiple, lineare Regression (s. S. 159ff.)

Parametrische Verfahren zur Hypothesenprüfung

a) Mittelwertsvergleiche

	unabhängige Stichproben		abhängige Stichproben	
	Vergleich *zweier* Stichpr.	Vergleich *mehrerer* Stichpr.	Vergleich *zweier* Stichpr.	Vergleich *mehrerer* Stichpr.
Intervall-skalenniveau	*z-Test* (wenn der Pop.-Mittelwert bekannt ist) (s. S. 190ff.) *t-Test* für unabhängige Stichproben (s. S. 192ff.)	einfache *Varianzanalyse* (s. S. 198ff.) mit: Newman-Keuls-Test (s. S. 205ff.) und Omega-Test (s. S. 207ff.)	*t-Test* für abhängige Stichproben (s. S. 194ff.)	

b) Varianz-Vergleiche

	unabhängige Stichproben		abhängige Stichproben	
	zwei	mehrere	zwei	mehrere
Intervall-skalenniveau	F-Test (s. S. 197ff.)	F_{max}-Test (s. S. 197ff.)	t-Test nach Ferguson (s. S. 198ff.)	t_{max}-Test nach Ferguson (s. S. 198ff.)

Nonparametrische Verfahren zur Hypothesenprüfung

	unabhängige Stichproben		abhängige Stichproben	
	Vergleich *zweier* Stichpr.	Vergleich *mehrerer* Stichpr.	Vergleich *zweier* Stichpr.	Vergleich *mehrerer* Stichpr.
Häufigkeiten (Nominal-skala)	Chi^2-Test (s. S. 213ff.)	Chi^2-Test (s. S. 217ff.)	McNemar-Test (s. S. 233ff.)	Cochran-Q-Test (s. S. 235ff.)
Rangdaten (Ordinal-skala)	U-Test von Mann-Whitney (s. S. 226ff.)	H-Test von Kruskal-Wallis-Rangvarianz-analyse (s. S. 230ff.)	Wilcoxon-Test (s. S. 238ff.)	Friedman-Rangvarianz-analyse (s. S. 241ff.)

Anhang 4

Verzeichnis der Abbildungen im Text

Anhang 5

Verzeichnis der statistischen Tabellen im Anhang

Anhang 6

Literaturverzeichnis

Allport, G. W., 1942. The use of personal documents in psychological science. New York.
Anger, H., 1969. Befragung und Erhebung. In: *C. F. Graumann* (Hrsg.), Sozialpsychologie, Bd. I. Göttingen. (= Bd. 7 d. Hb. d. Psychol.)

Bales, R. F., 1951. Interaction process analysis. Cambridge.
Bartel, H. u.a., 1971. Statistik I. Stuttgart.
Bartel, H. u.a., 1972. Statistik II. Stuttgart.
Bennett, V. D. C., 1964. Development of a Self Concept & Sort for use with elementary age school children. J. School Psychol., 3, 19–24.
Berelson, B., 1954. Content analysis. In: *G. Lindzey* (Hrsg.), Handbook of Social Psychology, Bd. I., 3. Aufl. 1959, Reading, London.
Blöschl, L., 1966. Kullbacks 2 Î-Test als ökonomische Alternative zur Chi^2-Probe. Psychol. Beitr., 9, 379–406.
Blommers, P. & Lindquist, E. F., 1960. Elementary Statistical Methods in Psychology and Education. Boston.
Bochenski, I. M., 1954. Die zeitgenössischen Denkmethoden. 3. Aufl. 1965, München.
Bredenkamp, J., 1969. Experiment und Feldexperiment. In: *C. F. Graumann* (Hrsg.), Sozialpsychologie, Bd. I. Göttingen. (= Bd. 7 d. Hb. d. Psychol.)
Brown, J. F. & Ghiselli, E. E., 1955. Scientific Methods in Psychology. New York.

Campbell, D. T. & Stanley, J. C., 1963. Experimental and Quasi-Experimental Designs for Research on Teaching. In: *N. L. Gage* (Hrsg.), Handbook of Research on Teaching. Chicago. – Dt. Bearbeitg. von *E. Schwarz* in: *K. Ingenkamp* (Hrsg.), 1970. Hb. d. Unterrichtsforschung, Teil I. Weinheim/Berlin/Basel.
Carnap, R., 1942. Introduction to Semantics. Cambridge.
Carnap, R., 1956. The methodological character of theoretical concepts. Minnesota Studies in the Philosophy of Science, 1, 38–76.
Carnap, R., 1969. Einführung in die Philosophie der Naturwissenschaft. München.
Carnap, R. & Jeffry, R. C. (Hrsg.), 1971. Studies in inductive logic and probability. London.
Clauß, G. & Ebner, H., 1971. Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Frankfurt/M., Zürich. (1. Aufl. 1967 VEB, Berlin).
Coerper, C.; Hagen, W.; Thomae, H., 1954. Deutsche Nachkriegskinder. Stuttgart.
Cranach, M. v. & Frenz, H. G., 1969. Systematische Beobachtung. In: *C. F. Graumann* (Hrsg.). Sozialpsychologie, Bd. I. Göttingen. (= Bd. 7 d. Hb. d. Psychol.)
Cronbach, L. J., 1960. Essentials of psychological testing. 2. Aufl., New York.
Cube, F. v., 1965. Kybernetische Grundlagen des Lernens und Lehrens. 3. Aufl., 1971, Stuttgart.

Dietrich, R., 1972. Einführung in die methodischen Grundlagen der Pädagogischen Psychologie. München/Basel.
Donat, H., 1970. Persönlichkeitsbeurteilung. 2. Aufl., München.
Drever, J. & Fröhlich, W. D., 1968. Wörterbuch zur Psychologie. München (dtv).

Edwards, A. L., 1954. Experiments: Their planning and execution. In: *G. Lindzey* (Hrsg.), Handbook of social psychology. Cambridge, Mass.

Edwards, A. L., 1960. Experimental design in psychological research. 2. Aufl., New York. – Dt. Übers.: Versuchsplanung in der psychologischen Forschung. Weinheim 1970.
Erlemeier, N. & Tismer, K. G., 1973. Einstellungen und Erwartungen bei Lehrern und ihre Auswirkungen auf die Beurteilung und das Verhalten von Schülern. In: *H. Nickel & E. Langhorst* (Hrsg.), Brennpunkte der Pädagogischen Psychologie. Bern/Stuttgart.

Fenner, J., 1973. Verfahren und Ergebnisse zur Objektivierung des Lehrerverhaltens. In: *H. Nickel & E. Langhorst* (Hrsg.), Brennpunkte der Pädagogischen Psychologie. Bern/Stuttgart.
Ferguson, G. A., 1971. Statistical Analysis in Psychology and Education. 3. Aufl., New York.
Festinger, L. & Katz, D. (Hrsg.), 1953. Research methods in the behavioral science. New York.
French, J. R. P., 1953. Experiments in Field Settings. In: *L. Festinger & D. Katz* (Hrsg.), Research methods in the behavioral science. New York.
Fröhlich, W. D., 1965. Forschungsstatistik. 4. Aufl., Bonn.
Fröhlich, W. D. & Becker, J., 1971. Forschungsstatistik. 5. Aufl., Bonn.
Flesch, R., 1951. How to test readability. New York.

Gadamer, H. G., 1960. Wahrheit und Methode. Tübingen.
Gage, N. L. (Hrsg.), 1963. Handbook of Research on Teaching. Chicago. – Dt. Bearbeitung von *K. Ingenkamp & E. Parey* (Hrsg.), 1970/71. Hb. d. Unterrichtsforschung, 3 Bde. Weinheim/Berlin/Basel.
Gottschalk, L.; Kluckhohn, C.; Angell, R. C., 1945. The use of personal documents in history, anthropology and sociology. New York.
Graumann, C. F., 1960. Grundlagen einer Phänomenologie und Psychologie der Perspektivität. Berlin.
Graumann, C. F., 1964. Eigenschaften als Problem der Persönlichkeitsforschung. In: *P. Lersch & H. Thomae* (Hrsg.), Persönlichkeitsforschung und Persönlichkeitstheorie. 2. Aufl., Göttingen. (= Bd. 4d. Hb. d. Psychol.).
Graumann, C. F., 1965. Methoden der Motivationsforschung. In: *H. Thomae* (Hrsg.), Allgemeine Psychologie II: Motivation. 1. Aufl., Göttingen. (= Bd. 2 d. Hb. d. Psychol.).
Graumann, C. F., 1965. Methoden der Psychologie. Heidelberg (unveröffentl. Vorlesungsskript).
Graumann, C. F., 1966. Grundzüge der Verhaltensbeobachtung. In: *E. Meyer* (Hrsg.), Fernsehen in der Lehrerbildung. München.
Graumann, C. F. (Hrsg.), 1969/1972. Sozialpsychologie, 2 Bde. Göttingen. (= Bd. 7d. Hb. d. Psychol.).
Guilford, J. P., 1954. Psychometric methods. 2. Aufl., New York.
Guilford, J. P., 1965. Fundamental statistics in psychology and education. 4. Aufl., New York.

Hagen, W.; Thomae, H.; Ronge, A., 1962. Zehn Jahre Nachkriegskinder. München.
Hasemann, K., 1964. Verhaltensbeobachtung. In: *R. Heiß* (Hrsg.), Psychologische Diagnostik. Göttingen. (= Bd. 6d. Hb. d. Psychol.).
Hays, W., 1969. Statistics. 2. Aufl., London.
Heller, K., 1966. Voraussetzungen und Möglichkeiten wissenschaftlicher Forschung. N. Blätter f. Taubstummenbildg., 20, 162–174, 290–308.
Heller, K., 1970. Aktivierung der Bildungsreserven. Bern/Stuttgart.
Heller, K., 1972. Sehgeschädigtenpsychologie als Wissenschaftsdisziplin. Sehgeschädigte. Internat. Wiss. Arch., 1, 50–64.
Heller, K., 1973. Intelligenzmessung. Villingen.
Herrmann, T. & Stäcker, K. H., 1969. Sprachpsychologische Beiträge zur Sozialpsychologie. In: *C. F. Graumann* (Hrsg.), Sozialpsychologie, Bd. I. Göttingen. (= Bd. 7 d. Hb. d. Psychol.).
Hilgard, E. R. & Bower, G. H., 1956. Theories of Learning. 2. Aufl. New York. – Dt. 1970/71. Theorien des Lernens. 2 Bde. Stuttgart.

Hörmann, H., 1967. Psychologie der Sprache. Berlin/Heidelberg/New York.
Hofer, M., 1969. Die Schülerpersönlichkeit im Urteil des Lehrers. Weinheim.
Hofstätter, P. R. (Hrsg.), 1957. Psychologie. Frankfurt/M. (= Fischer Lexikon).
Hofstätter, P. R. & Wendt, D., 1967. Quantitative Methoden der Psychologie. 3. Aufl., München.
Holzkamp, K., 1964. Theorie und Experiment in der Psychologie. Berlin.
Holzkamp, K., 1968. Wissenschaft als Handlung. Berlin.
Holzkamp, K., 1970. Zum Problem der Relevanz psychologischer Forschung für die Praxis. Psychol. Rdsch., 21, 1–22.
Husserl, E., 1900/01. Logische Untersuchungen, 2 Bde. 2. Aufl. 1910, Halle.
Husserl, E., 1913. Ideen zu einer reinen Phänomenologie. 2. Aufl. 1950, Haag.

Klauer, K. J., 1973. Das Experiment in der pädagogischen Forschung. Düsseldorf.
König, R., 1956. Beobachtung und Experiment. In: *R. König* (Hrsg.), Praktische Sozialforschung, Bd. II. Köln.
König, R., (Hrsg.), 1962ff. Hb. d. emp. Sozialforschung. Bd. I: 1962, 2. Aufl. 1967; Bd. II: 1969. Stuttgart.
Kullback, S., 1959. Information theory and statistics. New York.
Kullback, S.; Kuppermann, M.; Ku, H. K., 1962. An application of information theory to the analysis of contingency tables of 2n ln n, n = 1 (1) 10000. Journal of Research of the National Bureau of Standards. – B. Math. and Math. Phys., Vol. 66 B, 4, 217ff. (Zit. nach Blöschl 1966).

Landsheere, G. de, 1969. Einführung in die pädagogische Forschung. Weinheim/Berlin/Basel.
Lasswell, H., 1949. Language of politics. 2. Aufl. 1965, New York.
Lewin, K., 1927. Gesetz und Experiment in der Psychologie. Symposion. Phil. Z. f. Forschg. u. Aussprache, 1, 375–421. – Nachdruck 1967, Darmstadt (Wiss. Buchgesellsch.).
Lienert, G. A., 1962. Verteilungsfreie Methoden in der Biostatistik. Meisenheim am Glan.
Lienert, G. A., 1967. Testaufbau und Testanalyse. 2. Aufl., Weinheim, Berlin.
Lindquist, E. F., 1953. Design and analysis of experiments in psychology and education. Boston, Mass.
Löwith, K., 1962. Das Individuum in der Rolle des Mitmenschen. Darmstadt.

Markefka, M. & Nauck, B., 1972. Zwischen Literatur und Wirklichkeit. Neuwied, Berlin.
Marx, M. H., 1963. Theories in contemporary psychology. New York, London.
Maslow, A. H., 1954. Motivation and Personality. New York.
Mc Corquodale, K. & Meehl, P. E., 1948. On a distinction between hypothetical constructs and intervening variables. Psychol. Rev., 55, 95–107.
Mc Nemar, Q., 1962. Psychological statistics. New York.
Medley, D. M. & Mitzel, H. E., 1963. Measuring Classroom-Behavior by Systematic Observation. In: *N. L. Gage* (Hrsg.), Handbook of Research on Teaching. Chicago. – Dt. Bearbtg. von *W. Schulz* u.a. in: *K. Ingenkamp* (Hrsg.), 1970. Hb. d. Unterrichtsforschung, Tl. I. Weinheim/Berlin/Basel.
Meili, R. & Rohracher, H. (Hrsg.), 1963. Lehrbuch der experimentellen Psychologie. 2. Aufl. 1968, Bern/Stuttgart.
Metzger, W., 1952. Das Experiment in der Psychologie. Stud. Gen., 5, 142–163.
Metzger, W., 1954. Psychologie. 2. Aufl., Darmstadt.
Mittenecker, E., 1970. Planung und statistische Auswertung von Experimenten. 8. Aufl., Wien.
Mowrer, O. H., 1953. Q-technique-description, history, and critique. In: *O. H. Mowrer* (Hrsg.), Psychotherapy: Theory and research. New York.

Nickel, H., 1972. Entwicklungspsychologie des Kindes- und Jugendalters. Bd. I. Bern, Stuttgart, Wien.

Norton, D. W., 1952. An empirical Investigation of Some Effects of Non-Normality and Heterogenity on the F-Distribution. (Zit. nach *E. F. Lindquist* 1953).

Osgood, C. E.; Suci, G. J.; Tannenbaum, P. H., 1957. The measurement of meaning. Urbana, I 11.

Pfanzagl, J., 1967/8. Allgemeine Methodenlehre der Statistik. Bd. I (4. Aufl.) u. Bd. II (3. Aufl.). Berlin. (Göschen).
Pongratz, L., 1967. Problemgeschichte der Psychologie. Bern.
Popper, K. R., 1971. Logik der Forschung. 4. Aufl., Tübingen.

Remmers, H. H., 1953. Rating-Methods in Research on Teaching. In: *N. L. Gage* (Hrsg.), Handbook of Research on Teaching. Chicago. – Dt. Bearbtg. von *L. Tent* in *K. Ingenkamp* (Hrsg.), 1970. Hb. d. Unterrichtsforschung, Tl. I. Weinheim/Berlin/Basel.
Röhrs, H., 1968. Forschungsmethoden in der Erziehungswissenschaft. Stuttgart/Berlin/Köln/Mainz.
Rosenthal, R., 1966. Experimenter effects in behavioral research. New York.
Rosenthal, R., 1967. Covert communication in the psychological experiment. Psychol. Bull., 67, 356–367.
Roth, H., 1958. Die Bedeutung der empirischen Forschung für die Pädagogik. In: Hochschule für Internationale Pädagogische Forschung (Hrsg.), Pädagogische Forschung und pädagogische Praxis. Heidelberg.
Ryans, D. G., 1960. Characteristics of Teachers. Washington, D. C.: American Council on Education.

Scott, W. A. & Wertheimer, M., 1962. Introduction to psychological research. New York.
Selg. H., 1966. Einführung in die experimentelle Psychologie. 3. Aufl. 1971, Stuttgart/Berlin/Köln/Mainz.
Sheldon, M. S. & Sorenson, A. G., 1960. On the use of Q-technique in educational evaluation and research. J. exp. Educ., 29, 143–151.
Siegel, S., 1956. Nonparametric Statistics for the behavioral sciences. New York usw.
Sixtl, F., 1967. Meßmethoden der Psychologie. Weinheim/Berlin/Basel.
Stegmüller, W., 1957. Das Wahrheitsproblem und die Idee der Semantik. Wien/Heidelberg/New York.
Stegmüller, W., 1969. Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Bd. I: Wissenschaftliche Erklärung und Begründung. Berlin/Heidelberg/New York.
Stegmüller, W. & Carnap, R., 1972. Induktive Wahrscheinlichkeit. In: *J. Speck* (Hrsg.), Grundprobleme der großen Philosophen. Philosophie der Gegenwart I. Göttingen.
Stephenson, W., 1953. The study of behavior. Q-technique and its methodology. 3. Aufl. 1956, Chicago.
Stern, G. G., 1963. Measuring Noncognitive Variables in Research on Teaching. In: *N. L. Gage* (Hrsg.), Handbook of Research on Teaching. Chicago. – Dt. Bearbeitg. von *E. Parey & K. Ingenkamp* in: *K. Ingenkamp* (Hrsg.), 1970. Hb. d. Unterrichtsforschung, Tl. I. Weinheim/Berlin/Basel.
Stevens S. S., 1939. Psychology and the Science of Science. Psychol. Bull., 36, 221–263.
Süllwold. F., 1969. Theorie und Methodik der Einstellungsmessung. In: *C. F. Graumann* (Hrsg.), Sozialpsychologie, Bd. I Göttingen. (= Bd. 7d. Hb. d. Psychol.).

Tausch, R. & Tausch, A., 1970. Erziehungspsychologie. 5. Aufl., Göttingen.
Taylor, W. L., 1953. "Cloze prozedure": A new tool for measuring readability. J. Quart., 30, 415–433.
Taylor, W. L., 1955. Application of Cloze and Entropy measures to the study of contextual contains in samples of continous prose. Diss. Abstr., 15, 464–465.

Thomae, H., 1954. Beobachtung und Beurteilung von Kindern und Jugendlichen. 9. Aufl. 1970, Basel/New York.
Thomae, H. & Feger, H., 1969. Hauptströmungen der neueren Psychologie. Frankfurt/Bern/Stuttgart. (= Bd. 7 der „Einführung in die Psychologie", Hrsg. *C. F. Graumann*).
Topitsch, E. (Hrsg.), 1960. Probleme der Wissenschaftstheorie. Wien/Heidelberg/New York.
Topitsch, E. (Hrsg.), 1965. Logik der Sozialwissenschaften. Köln/Berlin.
Townsend, J. C., 1953. Introduction to experimental method. New York usw.
Traxel, W., 1964. Einführung in die Methodik der Psychologie. 2. Aufl. 1973, Bern/Stuttgart.
Traxel, W., 1968. Über Gegenstand und Methode der Psychologie. Bern/Stuttgart.

Überla, K., 1968. Faktorenanalyse. Berlin/Heidelberg/New York.

Walker, H. M., 1967. Statistische Methoden für Psychologen und Pädagogen. 3. Aufl., Weinheim.
Walter, H., 1973. Neue Wege zum optimalen Unterricht – Beobachtung und Beurteilung von Schüler- und Lehrerverhalten. München.
Weber, E., 1967. Grundriß der biologischen Statistik. 6. Aufl., Stuttgart.
Westmeyer, H., 1973. Kritik der psychologischen Unvernunft. Stuttgart/Berlin/Köln/Mainz.
White, R. K., 1949. Hitler, Roosevelt, and the nature of war propaganda. J. abnorm. soc. Psychol., 44, 157–174.
Winer, B. J., 1962. Statistical Principles in Experimental Design. New York.
Wundt, W., 1874. Grundzüge der physiologischen Psychologie. 6. Aufl. 1908ff., Leipzig.
Wundt, W., 1896. Grundriß der Psychologie. 15. Aufl., 1922, Leipzig/Stuttgart.

Zimny, G. H., 1961. Method in experimental psychology. New York.

Anhang 7

Stichwortverzeichnis

Die Autoren

Kurt Heller (Jahrgang 1931) studierte Psychologie, Pädagogik, Sonderpädagogik sowie Teilgebiete der Medizin und Philosophie an den Universitäten Freiburg/Br. und Heidelberg. Der Autor ist o. Professor für Psychologie und Direktor am Seminar für Psychologie der Abteilung Bonn der Päd. Hochschule Rheinland sowie Lehrbeauftragter im Fach Erziehungswissenschaft an der Universität Heidelberg.
Arbeitsschwerpunkte: Pädagogische und Sonderpädagogische Psychologie, Diagnostik, Begabungs- und Bildungsforschung.
Wichtigste Veröffentlichungen: Modell eines *Guidance*-Systems für Abiturienten und Studenten (zus. m. E. Demel u. G. Schorre), Villingen 1969; Aktivierung der Bildungsreserven, Bern/Stuttgart 1970; Intelligenzmessung, Villingen 1973; Automatische Klassifikation von psychologischen Untersuchungsbefunden (zus. m. U. Allinger), Villingen 1974; Leistungsbeurteilung in der Schule (Hrsg.) Heidelberg 1974; u.a. – In Kürze erscheinen: Beratung im Bildungswesen (Hrsg.), Stuttgart 1974; Kognitiver Fähigkeitstest für 4. bis 13. Klassen (Adapt. d. Cognitive Abilities Tests von R. L. Thorndike u. E. Hagen), Weinheim 1974.

Bernhard Rosemann (Jahrgang 1940) studierte Psychologie und Soziologie an den Universitäten Heidelberg (Diplom-Prüfung 1969) und Mainz (Promotion 1972). Er ist wiss. Assistent am Psychologischen Seminar der Abteilung Bonn der Päd. Hochschule Rheinland.
Arbeitsschwerpunkte: Sozialpsychologie, Pädagogische Psychologie und Angewandte Psychologie.
Wichtigste Veröffentlichungen: Vorgesetzte und Mitarbeiter – Rollenerwartungen und interpersonales Verhalten, Diss. Mainz 1972; Führungsverhalten und Rollenerwartungen in formellen Gruppen. Psychol. u. Praxis, 1973; Konstruktion und Einsatz von Informellen Tests zur Leistungsbeurteilung (Lernkontrolltests), in: K. Heller (Hrsg.), Leistungsbeurteilung in der Schule, Heidelberg 1974.

Anne-Katrin Gaedike (Jahrgang 1949) studierte Psychologie, Pädagogik sowie Teilgebiete der Medizin und Biologie an den Universitäten in Braunschweig und Hamburg (Institut für experimentelle und angewandte Psychologie, Leiter Prof. Dr. K. Pawlik) und legte 1972 die Diplom-Prüfung ab. Die Mit-Autorin ist wiss. Assistentin am Psychologischen Seminar der Abteilung Bonn der Päd. Hochschule Rheinland. Sie promoviert dort bei Prof. Dr. K. Heller auf dem Gebiet der psychologischen Diagnostik.
Arbeitsschwerpunkte: Experimentelle und Angewandte Psychologie, Psychologische Diagnostik und Methodenlehre.
Veröffentlichung: Determinanten der Schulleistung, in: K. Heller (Hrsg.), Leistungsbeurteilung in der Schule, Heidelberg 1974.