Sonderdruck aus:

Mitteilungen
aus der
Arbeitsmarkt- und
Berufsforschung

Franz Egle, Marlis Eichinger

Die Kontrastgruppenanalyse

9. Jg./1976

2

Mitteilungen aus der Arbeitsmarkt- und Berufsforschung (MittAB)

Die MittAB verstehen sich als Forum der Arbeitsmarkt- und Berufsforschung. Es werden Arbeiten aus all den Wissenschaftsdisziplinen veröffentlicht, die sich mit den Themen Arbeit, Arbeitsmarkt, Beruf und Qualifikation befassen. Die Veröffentlichungen in dieser Zeitschrift sollen methodisch, theoretisch und insbesondere auch empirisch zum Erkenntnisgewinn sowie zur Beratung von Öffentlichkeit und Politik beitragen. Etwa einmal jährlich erscheint ein „Schwerpunktheft", bei dem Herausgeber und Redaktion zu einem ausgewählten Themenbereich gezielt Beiträge akquirieren.

Hinweise für Autorinnen und Autoren

Das Manuskript ist in dreifacher Ausfertigung an die federführende Herausgeberin
Frau Prof. Jutta Allmendinger, Ph. D.
Institut für Arbeitsmarkt- und Berufsforschung
90478 Nürnberg, Regensburger Straße 104
zu senden.

Die Manuskripte können in deutscher oder englischer Sprache eingereicht werden, sie werden durch mindestens zwei Referees begutachtet und dürfen nicht bereits an anderer Stelle veröffentlicht oder zur Veröffentlichung vorgesehen sein.

Autorenhinweise und Angaben zur formalen Gestaltung der Manuskripte können im Internet abgerufen werden unter http://doku.iab.de/mittab/hinweise_mittab.pdf. Im IAB kann ein entsprechendes Merkblatt angefordert werden (Tel.: 09 11/1 79 30 23, Fax: 09 11/1 79 59 99; E-Mail: ursula.wagner@iab.de).

Herstellung

Satz und Druck: Tümmels Buchdruckerei und Verlag GmbH, Gundelfinger Straße 20, 90451 Nürnberg

Verlag

W. Kohlhammer GmbH, Postanschrift: 70549 Stuttgart: Lieferanschrift: Heßbrühlstraße 69, 70565 Stuttgart: Telefon 07 11/78 63-0;
Telefax 07 11/78 63-84 30: E-Mail: waltraud.metzger@kohlhammer.de, Postscheckkonto Stuttgart 163 30.
Girokonto Städtische Girokasse Stuttgart 2 022 309.
ISSN 0340-3254

Bezugsbedingungen

Die „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" erscheinen viermal jährlich. Bezugspreis: Jahresabonnement 52,- €
inklusive Versandkosten: Einzelheft 14,- € zuzüglich Versandkosten. Für Studenten, Wehr- und Ersatzdienstleistende wird der Preis
um 20 % ermäßigt. Bestellungen durch den Buchhandel oder direkt beim Verlag. Abbestellungen sind nur bis 3 Monate vor Jahresende möglich.

Zitierweise:

MittAB = „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" (ab 1970)
Mitt(IAB) = „Mitteilungen" (1968 und 1969)
In den Jahren 1968 und 1969 erschienen die „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" unter dem Titel
„Mitteilungen", herausgegeben vom Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit.

Internet: http://www.iab.de

Die Kontrastgruppenanalyse

Theoretische Beschreibung und empirische Anwendung am Beispiel einer Untersuchung zur Erwerbstätigkeit verheirateter Frauen

Franz Egle, Marlis Eichinger

Die Aufdeckung komplexer Zusammenhänge zwischen verschiedenen Merkmalen aus dem sozialwissenschaftlichen Forschungsbereich erfordert ein statistisches Analyseverfahren, das nicht nur die Vielzahl der erklärenden Merkmale (Datenvielfalt), sondern auch den Einfluß berücksichtigt, der von der Kombination dieser erklärenden Größen (Wechselwirkungen) auf ein bestimmtes, zu erklärendes Merkmal ausgeht.

Die von Morgan und Sonquist entwickelte und inzwischen im IAB häufig angewandte Kontrastgruppenanalyse (KGA) zur Untersuchung von Merkmalszusammenhängen ist ein Verfahren, das dem Anspruch der gleichzeitigen Berücksichtigung der Datenvielfalt und der Wechselwirkungen entgegenkommt.

In der vorliegenden Arbeit werden die für die Entwicklung der KGA maßgebenden Probleme bei der klassischen Analyse sozialwissenschaftlicher Daten dargestellt und aufgezeigt, inwieweit die KGA diesem anspruchsvollen Ziel Rechnung trägt. Dabei wird eine Darstellung der statistischen Grundlagen der KGA in der Terminologie des allgemeinen linearen Modells verwendet, die es zudem erlaubt, auf den insbesondere im Forschungsbereich des IAB häufig auftretenden Fall einzugehen, wonach die Bestimmungsgründe für die Variation einer *qualitativen abhängigen* Variablen zu untersuchen sind.

Als begleitendes Beispiel für diesen Fall dient eine empirische Untersuchung der Erwerbstätigkeit verheirateter Frauen.

Gliederung

1. Einleitung
2. Analyse sozialwissenschaftlicher Daten mit den Verfahren des allgemeinen linearen Modells
   - 2.1 Darstellung
   - 2.2 Problemaufriß
     - 2.2.1 Datenvielfalt
     - 2.2.2 Wechselwirkungen (horizontal)
3. Analyse sozialwissenschaftlicher Daten mit dem Modell der Kontrastgruppenanalyse
   - 3.1 Grundlage: Die einfache Varianzanalyse
   - 3.2 Problemlösung
     - 3.2.1 Datenreduktion (Informationsverdichtung)
     - 3.2.2 Wechselwirkungen (vertikal)
     - 3.2.3 Kontrastgruppenanalyse bei qualitativen abhängigen Variablen
4. Anwendung der Kontrastgruppenanalyse zur Untersuchung der Erwerbstätigkeit verheirateter Frauen
   - 4.1 Datenmaterial
   - 4.2 Ergebnisse
5. Zusammenfassung und Ausblick auf Verbesserungen

1. Einleitung

Die Kontrastgruppenanalyse ist ein neueres statistisches Verfahren, das zur Analyse sozialwissenschaftlicher Erhebungsdaten entwickelt wurde und inzwischen im Institut für Arbeitsmarkt- und Berufsforschung immer häufiger zur Untersuchung komplexer Zusammenhänge angewandt wird[1]). Da auch hier — wie bei allen modellgebundenen Analysemethoden — die Ergebnisse von den zugrundeliegenden Prämissen abhängen, erscheint uns eine ausführliche Darstellung der Grundlagen und Voraussetzungen insbesondere im Hinblick auf die Interpretation der mit diesem Verfahren erzielbaren Ergebnisse erforderlich.

Die Aufdeckung der überwiegend komplexen Zusammenhänge im sozialwissenschaftlichen Forschungsbereich erfordert zum einen die Erhebung einer Vielzahl von Merkmalen und Merkmalsausprägungen und zum anderen ein statistisches Instrumentarium, das geeignet ist, statistisch gesicherte Aussagen über die Bestimmungsgründe der Variation interessierender Merkmale zu liefern.

Das einfachste statistische Instrumentarium, die Auszählung der Daten in ein- oder mehrdimensionale Tabellen und ihre Aufbereitung nach bestimmten Kriterien reicht nicht aus, gleichzeitige Einflüsse mehrerer Merkmale auf andere nachzuweisen und Zusammenhänge zwischen Merkmalen aufzudecken. Ein komplizierteres statistisches Instrumentarium, mit dem Zusammenhänge zwischen einer interessierenden und mehreren beeinflussenden Merkmalen analysiert werden können, sind die Verfahren des allgemeinen linearen Modells. Hierzu zählen die Regressions-, die Varianz- und die Covarianzanalyse, die zusammen eine große Klasse von Problemstellungen umfassen. Wie im folgenden dargelegt wird, ist jedoch bei Erhebungsdaten, die nicht unter den Bedingungen eines statistischen Versuchsplanes gewonnen werden, auch die Anwendung dieser Verfahren problematisch. Aus der Analyse dieser Probleme entstand die speziell für sozialwissenschaftliche Erhebungsdaten konzipierte Kontrastgruppenanalyse, die in dieser Arbeit theoretisch dargestellt und zur Untersuchung der Erwerbstätigkeit verheirateter Frauen angewandt wird.

2. Analyse sozialwissenschaftlicher Daten mit den Verfahren des allgemeinen linearen Modells 2.1 Darstellung

Wird das interessierende Merkmal als abhängige Variable und werden die beeinflussenden Merkmale als

[1]) Siehe z.B. Hofbauer, H., P. König, B. Nagel, Betriebszugehörigkeitsdauer bei männlichen deutschen Arbeitnehmern, MittAB 3/1974.

erklärende Variablen bezeichnet, so läßt sich das allgemeine lineare Modell wie folgt beschreiben[2]):

$$(1) \quad Y_i = \sum_{k=1}^{K} \beta_k X_{ki} + \varepsilon_i \qquad (i = 1, \ldots, n)$$

mit Y_i = i-te Realisation der abhängigen Zufallsvariablen

X_{ki} = i-te Realisation der k-ten erklärenden Variablen.

Die unbekannten Koeffizienten β_k werden z.B. nach der Methode der kleinsten Quadrate geschätzt und messen den Einfluß der zugehörigen erklärenden Variablen auf die abhängige Variable.

Für die Störvariablen ε_i, die den Zufallseinfluß der i-ten Beobachtung repräsentieren, wird angenommen, daß ihr Erwartungswert Null ist und daß sie unkorreliert sind, d.h.:

$$E(\varepsilon_i) = 0$$
$$E(\varepsilon_i \varepsilon_{i'}) = \begin{cases} \sigma^2 & \text{für } i = i' \\ 0 & \text{für } i \neq i' \end{cases}$$

Werden die Beobachtungswerte der abhängigen Variablen unabhängig voneinander gewonnen und ist die Störgröße normalverteilt, so können aufgestellte Hypothesen über die Parameter und damit die Einflüsse der erklärenden Variablen auf die abhängige Variable geprüft werden.

Das in (1) angegebene allgemeine lineare Modell[3]) umfaßt die Regressions-, die Varianz- und die Covarianzanalyse, die sich wie folgt unterscheiden:

Sind alle erklärenden Variablen quantitativ (wie z.B. das Merkmal Einkommen), so bezeichnet man diesen Spezialfall des allgemeinen linearen Modells als Regressionsanalyse-Modell. Sind dagegen alle erklärenden Variablen qualitativ (wie z.B. das Merkmal Beruf), so liegt das Modell der Varianzanalyse vor. Treten sowohl quantitative als auch qualitative erklärende Variablen auf, so spricht man von einem Covarianzanalyse-Modell.

Viele Erhebungsdaten in den Sozialwissenschaften repräsentieren Ausprägungen qualitativer Merkmale, so daß hier dem Varianzanalyse-Modell eine große Bedeutung zukommt.

Die erklärenden Variablen X_k haben dabei nur die Aufgabe, eine bestimmte Ausprägung eines Merkmals als vorhanden ($X_{ki} = 1$) oder nicht vorhanden ($X_{ki} = 0$) zu charakterisieren. Zur Darstellung des Varianzanalyse-Modells genügt daher die eindeutige Angabe der Koeffizienten für die jeweiligen Ausprägungen der erklärenden Variablen.

Zur Terminologie und Einteilung der Varianzanalyse ist folgendes festzuhalten:

Die unabhängigen oder erklärenden Variablen heißen Faktoren, ihre Ausprägungen nennt man Stufen oder Gruppen. Die von ihnen ausgehenden Wirkungen auf die abhängige Variable werden auch Effekte genannt. Einer bestimmten Kombination der Faktorstufen entdie den Zellen zugeordnet sind, heißen Beobachtungen oder Ergebnisse.

Nach der Anzahl der Faktoren unterscheidet man die einfache von der mehrfaktoriellen Varianzanalyse. Die Klassifikation der Beobachtungen nach einem Faktor führt zur einfachen Varianzanalyse. Treten gleichzeitig mehrere Faktoren auf, so liegt eine mehrfaktorielle Varianzanalyse vor.

Eine weitere Einteilung unterscheidet zwischen festen und zufälligen Faktorstufen bzw. zwischen festen und zufälligen Effekten. Im Falle fester Stufen gelten die Schlußfolgerungen nur für die in der Analyse befindlichen Faktorstufen, während bei zufälligen Stufen Schlußfolgerungen für die ganze Grundgesamtheit der möglichen Faktorausprägungen gezogen werden können.

Ein wichtiges Einteilungskriterium ist auch die Anzahl der Beobachtungen pro Zelle. Ist die Zellenhäufigkeit gleich, so sind bei der mehrfaktoriellen Varianzanalyse die Faktoren voneinander unabhängig. In diesem Fall, der praktisch nur bei der experimentellen Versuchsplanung eintritt, läßt sich die Gesamtvariation der abhängigen Variablen exakt als Summe der faktorbedingten Teilvariationen darstellen. Bei sozialwissenschaftlichen Erhebungsdaten werden die Zellenhäufigkeiten praktisch immer verschieden sein, so daß hier die exakte Variationszerlegung nicht gilt. Im Gegensatz zur klassischen Varianzanalyse spricht man hier von nichtorthogonaler Varianzanalyse.

2.2 Problemaufriß

Neben dem Problem der Linearität des Modells und dem der Korrelation zwischen den erklärenden Variablen, deren Vorliegen eine eindeutige Zuordnung der Einflüsse auf die einzelnen Faktoren beeinträchtigt, treten bei der Anwendung der Verfahren des allgemeinen linearen Modells zur Untersuchung von Merkmalszusammenhängen bei sozialwissenschaftlichen Erhebungsdaten vor allem zwei im folgenden dargestellte Probleme auf, die für die Entwicklung der Kontrastgruppenanalyse maßgebend waren[4]).

2.2.1 Datenvielfalt

Die Erforschung komplexer sozialer Zusammenhänge erfordert eine große Vielfalt von Informationen über jede Beobachtungseinheit. Die Berücksichtigung dieser Informationsvielfalt in Form von erklärenden Variablen kann insbesondere dann, wenn bei qualitativen Merkmalen jede Ausprägung als Variable betrachtet werden muß, sehr leicht zu einem für viele Rechenprogramme überdimensionierten Modellansatz führen.

2.2.2 Wechselwirkungen (horizontal)

Beziehen sich die Koeffizienten des allgemeinen linearen Modells auf einzelne originäre Variablen (z.B. X_i), so wird der von ihnen gemessene Einfluß *Hauptwirkung* genannt. Charakterisieren die Koeffizienten dagegen den Einfluß von abgeleiteten Variablen, die aus dem Produkt einzelner originärer Variablen (z.B. $X_k = X_i \cdot X_i$) entstanden sind, so wird dieser *Wechselwirkung* genannt.

Nach der Anzahl der Variablen, die zur Bildung der entsprechenden Wechselwirkungsvariablen herangezogen werden, unterscheidet man Wechselwirkung l. Ord-

[2]) Vgl.: Abrens, H., Varianzanalyse, Berlin, 1968.

[3]) Auf die Einfuhrung einer unterschiedlichen Bezeichnung für Zufallsvariablen und ihre Realisationen wurde der Einfachheit wegen verzichtet.

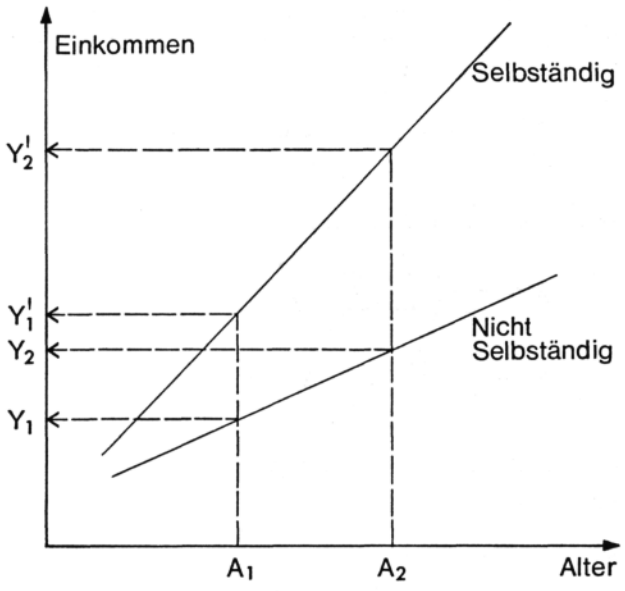
[4]) Vgl.: Morgan und Sonquist, Problems in the Analysis of Survey Data, and a Proposal. Journal of the American Statistical Association, 6/1963.

nung, wenn die abgeleitete Variable durch Multiplikation zweier originärer Variablen entstanden ist, und Wechselwirkung höherer Ordnung, wenn das entsprechende Produkt aus mindestens drei Faktoren besteht.

Zur Unterscheidung von den durch die Kontrastgruppenanalyse aufgedeckten sog. vertikalen Wechselwirkungen sollen die hier beschriebenen Wechselwirkungen horizontal genannt werden.

Sind die Variablen qualitativ, so liegt demnach eine Wechselwirkung erster Ordnung vor, wenn die Wirkung des einen Faktors mit den verschiedenen Stufen des anderen Faktors variiert.

Um dies zu verdeutlichen, soll mit folgendem hypothetischen Beispiel eine Wechselwirkung graphisch dargestellt werden:



Die Variable „Einkommen" sei von den beiden Faktoren „Alter" und „Stellung im Beruf" abhängig. In dem Schaubild gibt die untere Gerade den hypothetischen Zusammenhang zwischen dem erklärenden Faktor Alter und der abhängigen Variablen Einkommen an unter der Bedingung, daß der andere Faktor Stellung im Beruf die Ausprägung „Nichtselbständig" aufweist. Die obere Gerade gibt den entsprechenden Zusammenhang an unter der Bedingung, daß der Faktor Stellung im Beruf die Ausprägung „Selbstständig" aufweist.

Die Einkommensdifferenz $Y_2 - Y_1$ kann als Wirkung des einen Faktors Alter aufgefaßt werden unter der Bedingung, daß der andere Faktor Stellung im Beruf die Stufe „Nichtselbständig" aufweist, denn der Übergang von der Altersgruppe A_1 zur Altersgruppe A_2 bewirkt entsprechend der unteren hypothetischen Regressionsgeraden einen Übergang vom Einkommen Y_1 zum Einkommen Y_2. Unter der veränderten Bedingung, daß die Stufe „Selbstständig" vorliegt, ergibt sich entsprechend der oberen hypothetischen Regressionsgraden eine altersbedingte Einkommensdifferenz von $Y_2' - Y_1'$, die sich von der ursprünglichen Einkommensdifferenz $Y_2 - Y_1$ deutlich unterscheidet. Somit variiert die Wirkung des Faktors Alter mit den beiden Stufen des Faktors Stellung im Beruf, d.h., es liegt eine Wechselwirkung zwischen den beiden Variablen vor.

Wechselwirkungen würden dann nicht auftreten, wenn beide hypothetischen Regressionsgeraden parallel verliefen. Daß dies in den Sozialwissenschaften eher die Ausnahme als die Regel ist, wird kaum bezweifelt. Häufig tragen Wechselwirkungsvariablen mehr zur Erklärung einer abhängigen Variablen bei als Hauptwirkungsvariablen.

Wesentliche Gründe für das Auftreten von Wechselwirkungen bei Erhebungsdaten sind:

1. Gesellschaftliche Bedingungen, die dafür sorgen können, daß z. B. die oben erwähnten Einkommensunterschiede zwischen zwei beliebigen Ausprägungen des Faktors Alter mit vielen anderen Faktoren variieren, wie z.B. mit den Variablen Stellung im Beruf, Familienstand, Betriebszugehörigkeitsdauer, aber auch mit dem Geschlecht, der Nationalität und der Rassenzugehörigkeit.
2. Einzelne operationalisierbare Variablen messen nur einen Teil von Konstrukten, so daß diese erst durch das gemeinsame Auftreten der gemessenen Variablen repräsentiert werden können, wie z. B. die von Kish und Lansing definierte Variable „Familienzyklus", die durch das Zusammenwirken der einzelnen Variablen Alter, Familienstand, Zahl und Alter der Kinder charakterisiert wird[5]).

Hieran sieht man, daß die Wechselwirkungsvariablen bei der Analyse sozialwissenschaftlicher Daten eine zentrale Rolle spielen, da nur durch sie die in der Realität vorherrschenden nicht additiven Zusammenhänge aufgedeckt werden können.

Ein realitätsnahes Modell erfordert also sowohl die Berücksichtigung der Datenvielfalt durch die Einbeziehung vieler erklärender Faktoren als auch der zwischen ihnen bestehenden Wechselwirkungen.

Dieser Anspruch ist von den Verfahren des allgemeinen linearen Modells im allgemeinen nicht zu verwirklichen, weil die Vielzahl der dann zu berücksichtigenden Variablen die Kapazität jedes z. Z. verfügbaren Rechenprogrammes übersteigen wird. Selbst wenn dies nicht zuträfe, wäre wegen der großen Zahl entstehender nicht besetzter Zellen die Durchführung des Rechenprozesses fragwürdig. Hinzu kommt, daß Wechselwirkungen höherer Ordnung nicht mehr einfach zu interpretieren sind.

Dieses Dilemma zwischen der gleichzeitigen Berücksichtigung der Datenvielfalt und der Wechselwirkungen ist das Hauptproblem bei der Analyse von Merkmalszusammenhängen mit Erhebungsdaten und kann wie folgt charakterisiert werden: Berücksichtigt man zur Erklärung von Merkmalszusammenhängen möglichst alle erklärenden Faktoren, so erfaßt man zwar die Hauptwirkungen; auf die Wechselwirkungen muß jedoch aus oben genannten Kapazitätsgründen im allgemeinen verzichtet werden. Legt man dagegen größeren Wert auf die Berücksichtigung von Wechselwirkungen, so muß man im allgemeinen auf viele Einflußfaktoren verzichten.

Die Lösung dieses Dilemmas kann sinnvollerweise nur in einem Kompromiß zwischen beiden Ansprüchen bestehen. Das Optimierungsproblem besteht darin, zu entscheiden, wieviel von der Datenvielfalt reduziert

[5]) Vgl.: *Kish, L., Lansing, I.*, „Family life cycle as an independent variable", American Sociological Rerview, XXII, 1957.

werden kann, ohne die Aufdeckung der wesentlichen Wechselwirkungen zu verhindern.

Sonquist und Morgan veröffentlichten 1964 eine solche Lösung unter dem Namen „Automatic Interaction Detector" (AID)[5a]. Aus dieser Bezeichnung ist ersichtlich, daß die Autoren insbesondere die Aufdeckung von Wechselwirkungen anstrebten. Wie noch dargelegt wird, wurde dabei die Datenvielfalt so stark reduziert, daß pro erklärendem Merkmal nur noch zwei — gegensätzliche — Ausprägungen auftreten. Aus diesem Grund wurde im deutschen Sprachraum diese Methode zur „automatischen Auffindung von Wechselwirkungen" unter dem Namen Kontrastgruppenanalyse bekannt.

3. Analyse sozialwissenschaftlicher Daten mit dem Modell der Kontrastgruppenanalyse

3.1 Grundlage: Die einfache Varianzanalyse

Bei der Varianzanalyse wird angenommen, daß die Beobachtungen der abhängigen Variablen Realisierungen einer Zufallsvariablen sind, die von den Gruppen der erklärenden Faktoren linear — jedoch stochastisch gestört — abhängen. Im Falle der einfachen Varianzanalyse ergibt sich hieraus folgendes Modell:

$$(2) \quad Y_{ij} = \mu_i + \varepsilon_{ij}$$

wobei Y_{ij} die j-te Beobachtung der i-ten Gruppe darstellt. i ist also der Gruppenindex ($i = 1,\ldots,I$), j der Index innerhalb einer Gruppe ($j = 1,\ldots,n_i$), ε_{ij} sind unabhängige Zufallsvariablen mit dem Erwartungswert 0, der Varianz σ_ε^2 und sind mit dem Faktor unkorreliert.

μ_i = Mittelwert der i-ten Gruppe

n_i = Zahl der Beobachtungen in der i-ten Gruppe

$n = \sum_{i=1}^{I} n_i$ = Gesamtzahl der Beobachtungen

Vielfach ist es zweckmäßig, die Gleichung (2) so zu schreiben, daß der allgemeine Mittelwert in Form einer Konstanten auftritt. Schreibt man den allgemeinen Mittelwert als Summe der gewichteten Gruppenmittelwerte ($\mu = \frac{1}{n}\sum_{i=1}^{n} n_i\mu_i$) und bezeichnet die Abweichung des i-ten Gruppenmittelwertes vom Gesamtmittelwert als $\alpha_i := \mu_i - \mu$, so ergibt sich folgende zu (2) äquivalente Darstellung:

$$(3) \quad Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\text{mit} \sum_{i=1}^{I} n_i\alpha_i = 0^{6})$$

[5a]) *Sonquist* and *Morgan*, The Detection of Interaction Effects, 1964.

[6]) Diese Nebenbedingung ergibt sich wie folgt:

$$\sum_{i=1}^{I} n_i\alpha_i = \sum_{i=1}^{I} n_i(\mu_i - \mu) = \sum_{i=1}^{I} n_i\mu_i - \mu\sum_{i=1}^{I} n_i$$

$$= \sum_{i=1}^{I} n_i\mu_i - n\mu = n\mu - n\mu = 0.$$

Ist die Zahl der Beobachtungen pro Gruppe gleich, so erhält man die Nebenbedingung

$$\sum_{i=1}^{I} \alpha_i = 0.$$

[7]) Die Variation unterscheidet sich von der Varianz dadurch, daß bei ihr nicht durch die Zahl der Freiheitsgrade dividiert wird.

Bildet man den Erwartungswert von (2), so erhält man:

$$(4) \quad E(Y_{ij}) = \mu + \alpha_i$$

Der Koeffizient α_i mißt den Effekt der i-ten Gruppe auf die abhängige Variable. Somit wirkt der Faktor auf die abhängige Variable oder beeinflußt sie, wenn einzelne oder alle α_i von Null verschieden sind. In diesem Fall unterscheiden sich die Gruppenmittelwerte vom Gesamtmittelwert, und die Kenntnis der Stufen ist eine nützliche Information bei der Prognose oder Erklärung der abhängigen Variablen. Sind die α_i dagegen gleich Null oder nur unwesentlich von Null verschieden, so bringt die Einteilung des Faktors in Gruppen keine neue Erkenntnis, und anstelle der Gruppenmittelwerte kann die abhängige Variable genausogut durch den Gesamtmittelwert dargestellt werden.

Die Koeffizienten μ und α_i sind im allgemeinen unbekannt und müssen aus den vorhandenen Beobachtungen geschätzt werden. Nach der Methode der kleinsten Quadrate:

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i} (Y_{ij} - \mu - \alpha_i)^2 \rightarrow \min$$

erhält man die Schätzwerte:

$$\hat{\mu} = \frac{1}{n}\left(\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}\right)$$

$$\hat{\alpha}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} - \hat{\mu}$$

und somit die Schätzgleichung:

$$(5) \quad \hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i$$

Y_{ij} und $\hat{Y}_{ij}$ unterscheiden sich durch die stochastische Störung und stehen demnach in folgender Beziehung:

$$(6) \quad Y_{ij} = \hat{Y}_{ij} + \hat{\varepsilon}_{ij}$$

Da die Schätzfunktionen der Koeffizienten linear von der Zufallsvariablen Y abhängen, sind auch sie Zufallsvariablen. Deshalb kann mit diesen Schätzungen allein nicht entschieden werden, ob die Faktorgruppen auf die abhängige Variable wirken oder nicht. Um dies zu ermöglichen, muß das ausgeführt werden, was der Varianzanalyse den Namen gegeben hat: Die Zerlegung oder Aufspaltung der Gesamtvarianz (eigentlich der Gesamtvariation)[7]) in einen Teil, der mit dem Faktor zusammenhängt, und in einen anderen, der durch die stochastische Störung bedingt ist. Dominiert der erste Teil (Variation *zwischen* den Gruppen), so beeinflussen die Faktorgruppen die abhängige Variable, dominiert dagegen der zweite Teil (Variation *innerhalb* der Gruppen), so wirken die Faktorgruppen nicht wesentlich linear auf die abhängige Variable ein, und die Unterschiede in den Beobachtungen zwischen den Gruppen werden als zufällig angesehen.

Das Verhältnis der Zwischengruppenvariation zur Gesamtvariation wird Bestimmtheitsmaß genannt und gibt an, wieviel der in der abhängigen Variablen beobachteten Unterschiedlichkeit, gemessen durch die Gesamtvariation, auf die Faktorgruppierung zurückzuführen ist. Beträgt dieser Anteil z. B. 0,8, so sagt man, die Unterschiede in der abhängigen Variablen können

zu 80 % durch den Faktor erklärt oder vorhergesagt werden.

Ist die Störvariable stochastisch unabhängig von den erklärenden Variablen und normalverteilt, so läßt sich mit den beiden Variationsteilen und den zugehörigen Freiheitsgraden eine F-verteilte Testgröße bilden, mit der man entscheiden kann, ob die Faktorgruppierungen einen signifikanten Einfluß auf die abhängige Variable haben oder, anders ausgedrückt, ob die α-Effekte von Null verschieden sind, die Gruppenmittelwerte sich also vom Gesamtmittelwert signifikant unterscheiden.

Bei der Zerlegung der Gesamtvariation geht man von Gleichung (6) aus. Zieht man dort auf beiden Seiten den allgemeinen Mittelwert ab, quadriert und summiert die erhaltenen Ausdrücke, so erhält man nach einigen Umformungen[8]) folgende Zerlegungsformel:

$$(7)\quad \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\mu)^2 = \sum_{i=1}^{I} n_i(\hat{\mu}_i-\mu)^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu}_i)^2$$

Die Gesamtvariation der abhängigen Variablen auf der linken Seite wird demnach additiv zerlegt in eine Komponente, welche ausschließlich die Variation zwischen den Gruppen angibt:

$$\sum_{i=1}^{I} n_i(\hat{\mu}_i-\mu)^2$$

und in eine andere, welche ausschließlich die Variation innerhalb der Gruppen mißt:

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu}_i)^2$$

Als Bestimmtheitsmaß ergibt sich somit:

$$R^2 = \frac{\text{Zwischengruppenvariation}}{\text{Gesamtvariation}} = \frac{\sum_{i=1}^{I} n_i(\hat{\mu}_i-\mu)^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\mu)^2}$$

Aus dem Zerlegungssatz sieht man, daß die Minimierung der Variation innerhalb der Gruppen nach der Methode der kleinsten Quadrate äquivalent ist mit der Maximierung der Zwischengruppenvariation und damit auch der Maximierung des Bestimmtheitsmaßes.

3.2 Problemlösung

Die Aufdeckung von linearen Beziehungen zwischen mehreren erklärenden und einer abhängigen Variablen kann bei Erhebungsdaten mittels der oben beschriebenen klassischen Analyseverfahren nicht befriedigend gelöst werden, insbesondere wegen des Dilemmas der gleichzeitigen Berücksichtigung der Datenvielfalt durch zahlreiche erklärende Variablen und der zwischen ihnen bestehenden Wechselwirkungen. Die Kontrastgruppenanalyse versucht dieses Problem zu lösen, indem sie zum einen die Datenvielfalt reduziert und zum anderen anstelle der horizontalen nur vertikale Wechselwirkungen aufdeckt.

3.2.1 Datenreduktion (Informationsverdichtung)

Hier stellt sich das Problem, die Vielzahl der Klassen oder Ausprägungen pro Faktor zu reduzieren und gleichzeitig möglichst wenig Information zu verlieren. Dies ist dann möglich, wenn man die Klasseneinteilung der Faktoren so vornimmt, daß sich die Beobachtungswerte der abhängigen Variablen innerhalb der Klassen nur unwesentlich unterscheiden, zwischen ihnen jedoch große Unterschiede bestehen.

Für die Güte der Einteilung kommen demnach zwei Maße in Frage:

1. Ein Maß für die Kompaktheit oder Homogenität. Dieses bestimmt man numerisch sinnvoll durch die Summe der quadrierten Abweichungen der Beobachtungen in den Klassen von ihren Klassenmittelwerten:

   $\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij}-\mu_i)^2$ wobei K die reduzierte Zahl der Klassen angibt mit $K < I$.

2. Ein Maß für die Unterschiedlichkeit oder Inhomogenität zwischen den Klassen. Dieses wird sinnvollerweise durch die Summe der gewichteten quadratischen Abweichungen der Klassenmittelwerte vom Gesamtmittelwert gemessen:

   $$\sum_{i=1}^{K} n_i(\mu_i-\mu)^2$$

Die Güte der Klasseneinteilung ist am größten, wenn entweder das Inhomogenitätsmaß maximal oder das Homogenitätsmaß minimal ist: Eine solche Klasseneinteilung nennt man eine optimale Gruppierung. Aus der Zerlegungsformel für die einfache Varianzanalyse sieht man, daß die beiden Forderungen äquivalent sind. Aus diesen beiden Maßen abgeleitete Kriterien sind dann die Maximierung des Bestimmtheitsmaßes einerseits und die Maximierung des Verhältnisses von Inhomogenitäts- zu Homogenitätsmaß andererseits.

Die exakte Bestimmung der optimalen Gruppierung erfordert im allgemeinen die Berechnung dieser Gütemaße für alle möglichen Klasseneinteilungen. Werden die Faktoren wie bei der Kontrastgruppenanalyse in nur zwei Klassen eingeteilt ($K=2$), so ergibt das bei I ursprünglichen Ausprägungen $(2^{I-1}-1)$ Möglichkeiten der Gruppierung. Für einige Werte von I und K hat H. H. Bock folgende Zahlen ausgerechnet[9]) (siehe Tabelle, S. 214).

Daran sieht man, daß die exakte Bestimmung der optimalen Gruppierung mit einem zumutbaren Rechenaufwand nur bei kleinen Werten von I und K möglich ist. Häufig wird deshalb die optimale Gruppierung nur näherungsweise gefunden. Ein solches Näherungsverfahren stellt die Clusteranalyse (Automatische Klassifikation) dar.

Solange man sich mit der Einteilung in dichotome Klassen (Klassen mit zwei Ausprägungen) zufriedengibt, kann eine exakte Bestimmung der optimalen Gruppierung auch gefunden werden, ohne daß alle Gruppierungsmöglichkeiten durchgerechnet werden

[8]) Siehe Anhang, Punkt 1.

[9]) Siehe: *Bock, H. H.*, Automatische Klassifikation, in: Statistische Methoden II, Lecture Notes in Operations-Research and Mathematical Systems, Nr. 39, N.Y. 1970, S. 36 ff.

I \ K	2	3	4	5
5	15	25	10	1
10	511	9 330	34 105	42 525
20	$\approx 5 \cdot 10^6$	$\approx 6 \cdot 10^9$	$\approx 4{,}5 \cdot 10^{11}$	$\approx 7{,}5 \cdot 10^{12}$
50	$\approx 10^{15}$	$\approx 10^{23}$	$\approx 10^{29}$	$\approx 10^{33}$
100	$\approx 10^{30}$	$\approx 10^{47}$	$\approx 10^{59}$	$\approx 10^{68}$

I = ursprüngliche Zahl der Ausprägungen pro Faktor

K = Zahl der durch Zusammenfassung der ursprünglichen Ausprägungen entstandenen Klassen

müssen. Charakterisiert man einen beliebigen Faktor A durch die Menge seiner Ausprägungen:

$A = \{A_1, A_2, \ldots, A_I\}$ und ordnet die I Ausprägungen so, daß die entsprechenden Werte der abhängigen Variablen eine auf- oder absteigende Rangfolge bilden, so brauchen nach Ericson nur folgende Teilmengen miteinander verglichen zu werden[10]):

$\{A_1\}$ mit $\{A_2, \ldots, A_I\}$
$\{A_1, A_2\}$ mit $\{A_3, \ldots, A_I\}$
. .
. .
. .
$\{A_1, A_2, \ldots, A_{I-1}\}$ mit $\{A_I\}$

Hiernach sind für wesentlich weniger Gruppierungsmöglichkeiten die entsprechenden Gütemaße auszurechnen. Genau dieses Verfahren wird bei der Kontrastgruppenanalyse gewählt, um die optimale Gruppierung der Beobachtungen in zwei Klassen exakt zu bestimmen.

3.2.2 Wechselwirkungen (vertikal)

Nach dem oben angegebenen Verfahren wird für jeden erklärenden Faktor bezüglich der abhängigen Variablen eine optimale dichotome Gruppierung bestimmt. Die ursprünglich in einer gemeinsamen, inhomogenen Klasse befindlichen Beobachtungswerte der abhängigen Variablen werden dann bezüglich desjenigen Faktors in zwei Klassen zerlegt, dessen optimale Gruppierung den höchsten Wert des Bestimmtheitsmaßes aufweist[11]), also die beste Erklärungs- oder Prognosefähigkeit besitzt. Diese beiden so entstandenen Teilklassen sind homogener als die Ausgangsklasse, weil mit zunehmendem Wert des Bestimmtheitsmaßes die Variation zwischen den Teilklassen größer und entsprechend die Variation innerhalb der Teilklassen wegen des Zerlegungssatzes kleiner wird.

Diese beiden Teilklassen können nun wieder nach demselben Prinzip in je zwei weitere Teilklassen (Teil-Teilklassen) zerlegt werden. Derjenige Faktor, dessen optimale Gruppierung für die jeweilige Teilklasse den höchsten Wert des Bestimmtheitsmaßes aufweist, ist für sie der beste Prediktor.

Dieses Verfahren wird so lange wiederholt, bis die von Schritt zu Schritt homogener werdenden Teilklassen durch keine Gruppierung eines erklärenden Faktors noch kompakter werden.

Das Ergebnis dieses sukzessiven Aufteilens einer inhomogenen Ausgangsklasse in immer homogenere Teilklassen läßt sich übersichtlich mit Hilfe eines sog. Baumdiagrammes darstellen. Als hypothetisches Beispiel sollen hierfür folgende Variablen dienen:

Abhängige Variable: Y = Einkommen

Erklärende Faktoren: A = Ausbildung mit den Ausprägungen:
Hauptschule
Realschule
Fachschule
Universität

B = Stellung im Beruf mit den Ausprägungen:
Selbständig
Beamter
Angestellter
Arbeiter

C = Geschlecht

Die Ergebnisse einer mit diesen Variablen durchgeführten Kontrastgruppenanalyse seien:

1. Die erklärenden Faktoren ergeben folgende optimale Gruppierungen:
   - A: Universität — Hauptschule, Realschule, Fachschule
   - B: Selbständig — Beamter, Angestellter, Arbeiter
   - C: Männlich — Weiblich
2. Die durchschnittlichen Einkommen für die links stehenden Ausprägungen liegen signifikant über den jeweils gegenüberstehenden Ausprägungen.
3. Von den drei erklärenden Faktoren ergibt eine Gruppierung des Faktors A, in Akademiker und Nichtakademiker, den höchsten Wert des Bestimmtheitsmaßes, also die größte Erklärungskraft. Die Ausgangsklasse wird also zuerst nach der Gruppierung dieses Faktors geteilt.
4. Die Trennung der so entstandenen Teilklassen (Akademiker und Nichtakademiker) kann nach jeweils anderen Faktoren erfolgen. Für die Akademikergruppe ist das Merkmal Stellung im Beruf, für die Nichtakademiker dagegen das Merkmal Geschlecht schärfer trennendes Mittel, um Einkommensunterschiede zu erklären.

Eine graphische Darstellung dieser Ergebnisse ist das folgende Baumdiagramm.

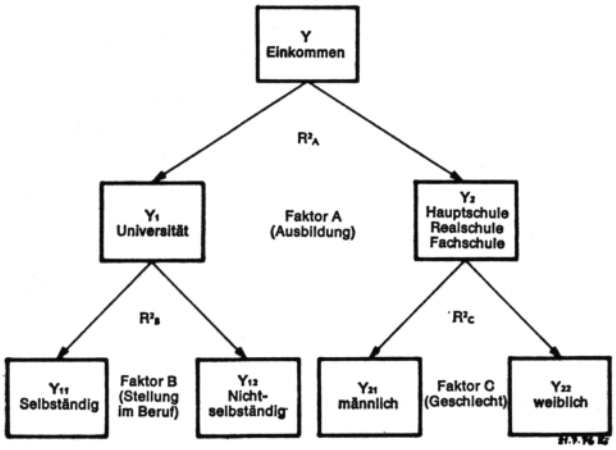
Den entscheidenden Vorteil dieser Darstellung sehen Morgan und Sonquist in der Möglichkeit, Wechselwirkungen zu erkennen. Sie geben hierfür drei Kriterien an:

1. Eine erklärende Variable kann nur auf einer Seite des Baumes oder Zweiges auftreten.

[10]) *Ericson, W. A.*, A Note on Partitioning for Maximum Between Sum of Squares, in: Sonquist and Morgan: The Detection of Interaction Effects, a.a.O., S. 149 – 157.

[11]) Als Trennungskriterium kann anstelle des Bestimmtheitsmaßes auch die Variation zwischen den Gruppen oder der F-Wert (Quotient aus der Variation zwischen den Gruppen zur Variation innerhalb der Gruppen, wobei Zähler und Nenner durch die jeweiligen Freiheitsgrade dividiert werden) verwendet werden.

Baumdiagramm einer hypothetischen Kontrastgruppenanalyse



2. Eine erklärende Variable tritt auf der einen Seite früher als auf der anderen Seite des Baumes oder Zweiges auf.

3. Die Zusammensetzung der optimalen Gruppierung einer erklärenden Variablen unterscheidet sich in den beiden Seiten oder Zweigen des Baumes.

Im gewählten Beispiel treten Wechselwirkungen auf: Während das Einkommen der Akademiker durch den Faktor Stellung im Beruf am stärksten erklärt wird, differenziert der Faktor Geschlecht die Einkommensunterschiede der Nichtakademiker am stärksten.

Entsprechend der sukzessiven Erklärung der Untergruppen einer abhängigen Variablen durch die Faktoren werden die so aufgedeckten Wechselwirkungen *vertikal* genannt. Diese sind jedoch nicht identisch mit den im Problemaufriß dargestellten horizontalen Wechselwirkungen, die aus der gleichzeitigen Erklärung der Gesamtgruppe durch mehrere Faktoren hervorgehen.

Es ist daher zweifelhaft, ob das von Sonquist und Morgan vorgeschlagene Verfahren zur Aufdeckung von Wechselwirkungen („Automatic Interaction Detector") seinen Namen zu Recht trägt. Kempf hat z. B. nachgewiesen, daß bei einem konstruierten Datenmaterial mit Wechselwirkungen die Methode von Morgan und Sonquist diese nicht aufgedeckt, andererseits bei einem Beispiel ohne Wechselwirkungen solche erkannt hat[12]).

Unabhängig davon, ob dieser Vorwurf berechtigt ist, bietet die Kontrastgruppenanalyse durch die sukzessive Unterteilung eines inhomogenen, umfangreichen Datenmaterials in homogenere Teilgruppen eine Möglichkeit, komplexe Zusammenhänge zwischen vielen Merkmalen durchschaubarer zu machen.

3.2.3 Kontrastgruppenanalyse bei qualitativen abhängigen Variablen

Bisher wurde stets angenommen, daß die abhängige Variable quantitativ ist, wie etwa die als Beispiel verwendete Variable Einkommen. In der praktischen Anwendung sind jedoch häufig die zu erklärenden Größen qualitativ, so auch die Variable Erwerbstätigkeit bei der im nächsten Kapitel durchgeführten empirischen Untersuchung zur Erwerbstätigkeit verheirateter Frauen.

Im folgenden soll deshalb der für die Anwendung der Kontrastgruppenanalyse entscheidende Zerlegungssatz der einfachen Varianzanalyse unter dieser veränderten Bedingung hergeleitet und die sich hieraus ergebende Formel für das Bestimmtheitsmaß diskutiert werden.

Ist Y eine dichotome qualitative abhängige Variable, so können ihre Realisierungen durch folgende Werte charakterisiert werden:

$$Y_{ij} = \begin{cases} 1, & \text{falls die j-te Beobachtung in der i-ten Gruppe die interessierende Ausprägung aufweist.} \\ 0, & \text{für alle sonstigen Angaben.} \end{cases}$$

Hieraus folgt:

1. $$\sum_{j=1}^{n_i} Y_{ij} = n_i\,(Y_{ij} = 1) \text{ und}$$

$$\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{n_i\,(Y_{ij} = 1)}{n_i} = p_i$$

2. $$\sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij} = \sum_{i=1}^{I} n_i\,(Y_{ij} = 1) \text{ und}$$

$$\mu = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{I} n_i\,(Y_{ij} = 1) = p$$

Aus dem Gruppenmittelwert μ_i wird also der Gruppenanteilswert p_i, aus dem Gesamtmittelwert μ der Gesamtanteilswert p.

Somit geht das varianzanalytische Modell

$$(2') \quad E(Y_{ij}) = \mu + \alpha_i = \mu + (\mu_i - \mu)$$

bei einer qualitativen dichotomen abhängigen Variablen wegen

$$E(Y_{ij}) = P(Y_{ij} = 1) \cdot 1 + P(Y_{ij} = 0) \cdot 0 = P(Y_{ij} = 1)$$

über in

$$(8) \quad P(Y_{ij} = 1) = p + (p_i - p)$$

Der Zerlegungssatz im Falle einer qualitativen abhängigen Variablen lautet[13]):

$$(9) \quad np\,(1 - p) = \left(\sum_{i=1}^{2} n_i p_i^2 - np^2 \right) + \left(np - \sum_{i=1}^{2} n_i p_i^2 \right)$$

Dabei stellt der Ausdruck auf der linken Seite der Gleichung die Gesamtvariation, der erste Ausdruck auf der rechten Seite die Variation zwischen den Gruppen

---

[12]) Vgl.: *Kempf, F.*, The Detection of Interaction Effects according to the Method of Sonquist & Morgan, Kongreß für Multivariate Statistische Verfahren, Wien 1970, unveröffentlicht.

[13]) Siehe Anhang, Punkt 2.

und der letzte Ausdruck die Variation innerhalb der Gruppen dar.

Das Bestimmtheitsmaß wird demnach durch folgende Formeln angegeben:

$$(10)\quad R^2 = \frac{\sum_{i=1}^{2} n_i p_i^2 - np^2}{np\,(1-p)} = \frac{np\,(1-p) + \sum_{i=1}^{2} n_i p_i^2 - np}{np\,(1-p)}$$

$$(11)\quad R^2 = 1 + \frac{\sum_{i=1}^{2} n_i\, p_i^2 - np}{np\,(1-p)}$$

Aufgrund von Formel (11) können folgende Schlußfolgerungen gezogen werden:

1. Damit $R^2 = 1$ wird (vollkommene Erklärung der abhängigen Variablen), muß der Zähler in (11) Null werden, d.h. die Bedingung

   $\sum n_i\, p_i^2 = np$ erfüllen.

   Wegen der Definition der Gesamtwahrscheinlichkeit $p = \frac{1}{n} \sum_{i=1}^{2} n_i p_i$ setzt dies voraus, daß

   $\sum_{i=1}^{2} n_i\, p_i^2 = \sum_{i=1}^{2} n_i\, p_i$ oder $p_i^2 = p_i$, also

   $(12)\quad p_i\,(1 - p_i) = 0$ ist.

   Die quadratische Gleichung in (12) wird nur durch die Werte $p_i = 1$ oder $p_i = 0$ erfüllt. Somit ist für $R^2 = 1$ eine optimale Gruppierung eines Faktors derart erforderlich, daß in der einen Untergruppe nur die interessierende Ausprägung der abhängigen Variablen (z.B. erwerbstätig) und in der anderen nur die nichtinteressierende Ausprägung (also nichterwerbstätig) vorkommt.

   Je mehr sich also p_i^2 von p_i unterscheidet, desto schlechter wird die Erklärung der abhängigen Variablen. Der maximale Unterschied zwischen p_i^2 und p_i ist bei $p_i = 0{,}5$ erreicht. Je mehr sich beide Gruppenwahrscheinlichkeiten also auf diese Intervallmitte zu bewegen, desto kleiner wird das Bestimmtheitsmaß.

2. Damit $R^2 = 0$ wird (die Schwankungen der abhängigen Variablen sind ausschließlich zufallsbedingt), muß in Formel (11) der Bruch -1 ergeben. Dies setzt voraus, daß

   $\sum_{i=1}^{2} n_i\, p_i^2 = np^2$ oder

   $(13)\quad p_1 = p = p_2$ wird.

   Je weniger sich also die beiden Gruppenwahrscheinlichkeiten von der Gesamtwahrscheinlichkeit unterscheiden, desto weniger kann die abhängige Variable durch die optimale Gruppierung eines Faktors erklärt werden.

3. Wegen dieser Tatsache kann folgendes festgehalten werden:

   Ist die Gesamtwahrscheinlichkeit sehr niedrig oder sehr hoch ($p \approx 0$ oder $p \approx 1$), so wird wegen der Bedingung

   $0 \leq p_1 \leq p \leq p_2 \leq 1$

   praktisch erzwungen, daß p_1, p_2 und p sich nicht sehr stark unterscheiden.

   Ein hoher oder niedriger Wert für die Gesamtwahrscheinlichkeit führt deshalb automatisch zu einem niedrigen Bestimmtheitsmaß.

   Nach diesen Ausführungen wird im folgenden Kapitel eine empirische Untersuchung der Erwerbstätigkeit der Frauen nach der Kontrastgruppenanalyse dargestellt, wobei die abhängige Variable die dichotomen Ausprägungen erwerbstätig — nicht erwerbstätig aufweist.

4. Anwendung der Kontrastgruppenanalyse zur Untersuchung der Erwerbstätigkeit verheirateter Frauen

4.1 Datenmaterial

Die hier verwendeten Daten wurden dem Sonderbeitrag des Statistischen Bundesamtes „Erwerbstätigkeit von Frauen und Müttern und ihre berufliche Ausbildung 1964 bis 1966“ entnommen[14]).

Für die Erwerbstätigkeit von Frauen wurden von den erfaßten Merkmalen folgende erklärenden Variablen als entscheidend angesehen: Alter, Ausbildung, Fami-

Tabelle 1:
Erwerbstätige verheiratete Frauen (in 1000)
Gliederungsmerkmale: Alter, Ausbildung und Vorhandensein von Kindern unter 18 Jahren

Kinder	Alter	Ausbildung				
		keine	Lehre	berufsbildende Schule	Hochschule	zusammen
nein	bis unter 25	35	160	100	4	298
	25 bis unter 35	89	187	175	13	464
	35 bis unter 45	93	156	150	9	408
	45 bis unter 55	232	159	163	7	560
	55 bis unter 65	319	85	137	6	547
	65 und mehr	76	19	32	1	128
	zusammen	844	766	757	40	2 405
ja	bis unter 25	31	72	78	1	183
	25 bis unter 35	212	212	357	10	791
	35 bis unter 45	299	217	402	21	932
	45 bis unter 55	166	52	126	7	352
	55 bis unter 65	38	4	14	–	56
	65 und mehr	1	–	–	–	1
	zusammen	747	557	977	39	2 322
zus.	bis unter 25	66	232	178	5	481
	25 bis unter 35	301	399	532	23	1 255
	35 bis unter 45	392	373	552	30	1 347
	45 bis unter 55	398	211	289	14	912
	55 bis unter 65	357	89	151	6	603
	65 und mehr	77	19	32	1	129
insgesamt		1 591	1 323	1 734	79	4 727

[16]) Statistisches Bundesamt, Fachserie A, Reihe 6.

Tabelle 2:
Anteil der erwerbstätigen verheirateten Frauen an den verheirateten Frauen insgesamt

Gliederungsmerkmale: Alter, Ausbildung und Vorhandensein von Kindern unter 18 Jahren

Kinder	Alter	Ausbildung				
		keine	Lehre	berufsbildende Schule	Hochschule	zusammen
nein	bis unter 25	0,50	0,80	0,77	1,00	0,74
	25 bis unter 35	0,57	0,76	0,70	0,76	0,69
	35 bis unter 45	0,27	0,49	0,42	0,64	0,40
	45 bis unter 55	0,23	0,37	0,34	0,39	0,29
	55 bis unter 65	0,19	0,25	0,40	0,33	0,23
	65 und mehr	0,07	0,09	0,17	0,10	0,09
	zusammen	0,20	0,44	0,43	0,49	0,31
ja	bis unter 25	0,22	0,35	0,42	0,33	0,34
	25 bis unter 35	0,24	0,29	0,36	0,33	0,30
	35 bis unter 45	0,34	0,36	0,42	0,41	0,38
	45 bis unter 55	0,49	0,46	0,56	0,64	0,51
	55 bis unter 65	0,28	0,18	0,47	0,00	0,29
	65 und mehr	0,11	0,00	0,00	0,00	0,08
	zusammen	0,31	0,33	0,41	0,40	0,36
zus.	bis unter 25	0,31	0,57	0,57	0,71	0,51
	25 bis unter 35	0,29	0,40	0,43	0,49	0,38
	35 bis unter 45	0,33	0,41	0,42	0,46	0,39
	45 bis unter 55	0,30	0,39	0,41	0,48	0,35
	55 bis unter 65	0,20	0,24	0,40	0,30	0,24
	65 und mehr	0,07	0,09	0,17	0,10	0,09
insgesamt		0,24	0,39	0,42	0,44	0,33

lienstand und Vorhandensein von Kindern unter 18 Jahren. Leider lagen in dieser Aufgliederung keine entsprechenden Kreuztabellen vor. Es wurden daher für die Gruppe der *verheirateten* Frauen unter Ausnutzung des gesamten vorhandenen Datenmaterials in sich konsistente Tabellen erstellt[15]). Dabei wurden folgende Ausprägungen für die erklärenden Variablen gewählt:

Alter: 10-Jahres-Gruppen (bis 25 Jahre, 25 bis 35 Jahre, ..., über 65 Jahre).

Ausbildung: (keine, Lehre, berufsbildende Schule, Hochschule).

Vorhandensein von Kindern unter 18 Jahren: (ja, nein).

Die Altersgliederung in obiger 10-Jahres-Gruppierung wirkt sich bei den 55- bis unter 65jährigen Frauen ungünstig aus, da Frauen ab 60 Jahren Rente beziehen und aus dem Erwerbsleben ausscheiden können. Dadurch, daß sich die Zusatzbefragung im April 1964 nicht auf den gleichen Personenkreis erstreckte wie der Mikrozensus, resultieren geringfügige unterschiedliche Ergebnisse in den beiden Erhebungen[16]).

Obwohl es nicht möglich war, Angaben über das Einkommen des Ehemannes mit in die Analysen einzubeziehen, läßt Tabelle 3 aber vermuten, daß die Höhe des

[15]) Vgl. Tabelle 1 und 2.

[16]) Einzelheiten hierzu siehe Statistisches Bundesamt, a.a.O., S. 12 ff.

Tabelle 3:
Verheiratete Frauen (in 1000) nach Erwerbstätigkeit und Einkommen des Ehemannes

Einkommen des Ehemannes	Verheiratete Frauen					
	erwerbstätig	%	nicht erwerbstätig	%	zusammen	%
bis unter 150	43	40,5	63	59,4	106	100
150 bis unter 300	142	25,9	407	74,1	549	100
300 bis unter 600	1 641	33,6	2 846	66,4	4 887	100
600 bis unter 800	1 071	27,5	2 830	72,5	3 901	100
800 und mehr	660	22,6	2 258	77,4	2 918	100
Landwirtschaft	717	92,5	58	7,5	775	100
kein Einkommen	42	35,3	77	64,7	119	100
ohne Angabe	127	28,5	318	71,5	445	100
nicht erfaßt	291	44,2	367	55,8	658	100

Einkommens des Ehemannes für die Einkommenssituation der Frau eine wesentliche Rolle spielt.

4.2 Ergebnisse

Die in Übersicht 1 (S. 218) dargestellten Ergebnisse der Kontrastgruppenanalyse können wie folgt interpretiert werden:

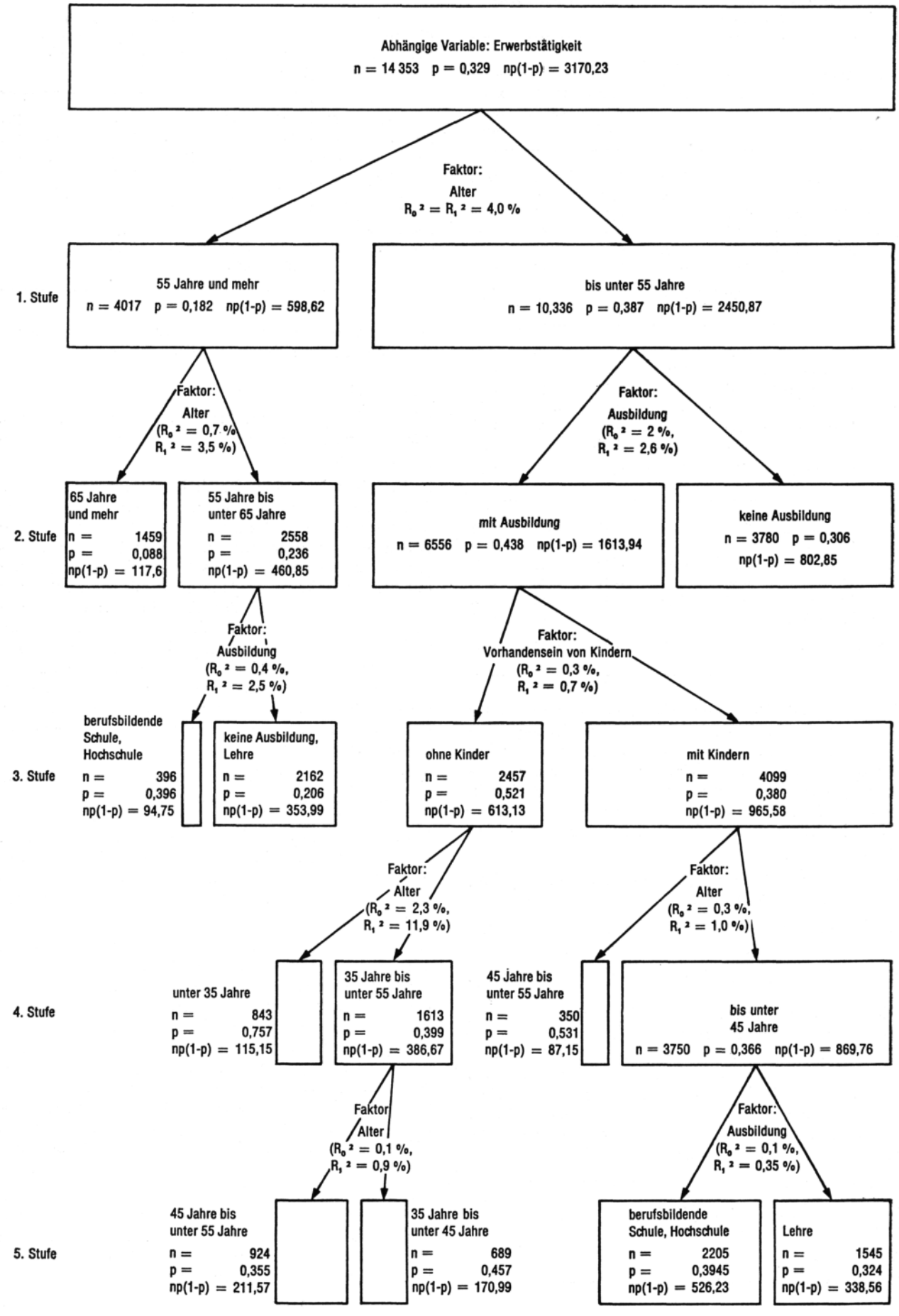
1. Praktisch ist keine der drei erklärenden Faktoren Alter, Ausbildung und Vorhandensein von Kindern unter 18 Jahren in der Lage, die Unterschiede in den Beobachtungswerten der Ausgangsgruppe der abhängigen Variablen allein zu erklären. Die Variable Alter mit der Schranke bei 55 Jahren hat mit rund 4 % noch den größten Anteil an der Variation der abhängigen Variablen. Die Frage, ob eine verheiratete Frau erwerbstätig ist oder nicht, läßt sich also nur zu rund 4 % durch die Kenntnis des Alters beantworten. Daran sieht man, daß auch die Summe der Hauptwirkungen, die aus allen einzelnen erklärenden Variablen resultiert, für die Erklärung der Erwerbstätigkeit verheirateter Frauen nicht ausreicht.

2. Während die Aufteilung der Untergruppe der älteren verheirateten Frauen (50 Jahre und älter) erneut nach dem Alter erfolgt, wird die der jüngeren verheirateten Frauen (unter 55 Jahren) durch die Ausbildung erzeugt. Das Merkmal Ausbildung übt daher bei jüngeren verheirateten Frauen einen stärkeren Einfluß auf die abhängige Variable Erwerbstätigkeit aus als bei den älteren verheirateten Frauen.

3. Das Vorhandensein von Kindern unter 18 Jahren wirkt sich erst in der dritten Analysestufe aus, und zwar bei der Gruppe der jüngeren verheirateten Frauen (unter 55 Jahren) mit Ausbildung. Hier steigt der Erwerbstätigenanteil von 38 % (mit Kindern) auf 52 % (ohne Kinder) an. Analytisch besonders herauszuheben wäre die Gruppe der unter 55jährigen Frauen ohne Ausbildung. Zur Erklärung der Erwerbstätigkeit dieser Gruppe ist keine der Variablen Alter und Vorhandensein von Kindern unter 18 Jahren geeignet. Hier spielen offensichtlich andere Faktoren eine entscheidende Rolle, so z.B. die Stellung des Ehemannes im Beruf oder dessen Einkommen.

4. Vergleicht man zwei Gruppen der 4. und 5. Aufteilungsstufe, so kann folgendes festgestellt werden: Während von den ausgebildeten verheirateten Frauen zwischen 45 und 55 Jahren ohne Kinder nur etwa ein

Übersicht 1:
Anteil der erwerbstätigen verheirateten Frauen an den verheirateten Frauen insgesamt in Abhängigkeit von den Merkmalen Alter, Ausbildung und Vorhandensein von Kindern unter 18 Jahren.
(R_1^2 = Bestimmtheitsmaß, bezogen auf die Variation der jeweiligen Untergruppe.
R_0^2 = Bestimmtheitsmaß, bezogen auf die Variation der Ausgangsgruppe.)



Drittel erwerbstätig sind, sind von den ausgebildeten verheirateten Frauen gleichen Alters mit Kindern mehr als die Hälfte erwerbstätig.

5. Die nach steigendem Erwerbstätigenanteil geordnete Tabelle 4 gibt die hierfür wesentlichen Kombinationen von Ausprägungen der erklärenden Variablen Alter, Ausbildung und Vorhandensein von Kindern an.

Tabelle 4:
Gruppen verheirateter Frauen nach steigendem Erwerbstätigenanteil

Anzahl in 1000	Anteil %	Variation np(1–p)	Merkmale der Gruppen
1 459	8,8	117,60	65 Jahre und älter
2 162	20,6	353,99	55 bis unter 65 Jahre, keine Ausbildung oder Lehre
3 780	30,6	802,85	unter 55 Jahre, keine Ausbildung
1 545	32,4	338,56	unter 45 Jahre, Lehre, Kinder
924	35,5	211,57	45 bis unter 55 Jahre, Ausbildung, ohne Kinder
2 203	39,5	526,23	unter 45 Jahre, berufsbildende Schule oder Hochschule, Kinder
396	39,6	94,75	55 bis unter 65 Jahre, berufsbildende Schule oder Universität
689	45,7	170,99	35 bis unter 45 Jahre, Ausbildung, ohne Kinder
350	53,1	87,15	45 bis unter 55 Jahre, Ausbildung, Kinder
843	75,7	115,15	unter 35 Jahre, Ausbildung, keine Kinder
14 353	32,9	3 170,23	verheiratete Frauen insgesamt

6. Abschließend kann festgehalten werden, daß die Zerlegung der abhängigen Variablen Erwerbstätigkeit in die oben angegebenen homogeneren Untergruppen etwa 10 % der Gesamtvariation der abhängigen Variablen erklärt.

5. Zusammenfassung und Ausblick auf Verbesserungen

Im vorliegenden Aufsatz wird die in letzter Zeit immer häufiger zur Untersuchung komplexer Zusammenhänge im sozialwissenschaftlichen Forschungsbereich verwendete Kontrastgruppenanalyse beschrieben und ihr theoretischer Aufbau dargestellt. Wesentliche Grundlage für die Kontrastgruppenanalyse ist das Modell der einfachen Varianzanalyse. Ausgehend von diesem Modell werden zur Reduktion der Datenvielfalt die Ausprägungen pro Merkmal in zwei gegensätzliche Klassen aufgeteilt und durch sukzessive Anwendung dieses Verfahrens auf die entstehenden Untergruppen der abhängigen Variablen vertikale Wechselwirkungen aufgedeckt. Mit der Kontrastgruppenanalyse wird versucht, das Dilemma der gleichzeitigen Berücksichtigung der Datenvielfalt und der Wechselwirkungen mit einem Kompromiß zu lösen.

Als Beispiel für die Anwendung der Kontrastgruppenanalyse wird in diesem Aufsatz der Anteil erwerbstätiger verheirateter Frauen an den verheirateten Frauen insgesamt in Abhängigkeit von den Merkmalen Alter, Ausbildung und Vorhandensein von Kindern untersucht.

Der von der Kontrastgruppenanalyse eingeschlagene Kompromiß bei der gleichzeitigen Berücksichtigung der Datenvielfalt und der Wechselwirkungen ist u. E. aus zwei Gründen zu restriktiv.

1. Die durch Einteilung der Merkmalsausprägungen in nur zwei Klassen entstehende Datenreduktion kann zu großen Informationsverlusten führen. Um dies zu verhindern, sollte die Zahl der Klassen nicht von vornherein auf zwei fixiert sein. Vorzuschlagen wäre eine trichotome Aufteilung insbesondere bei Merkmalen mit vielen Ausprägungen.

2. Durch die sukzessive Berücksichtigung dichotomer erklärender Variablen können nur vertikale Wechselwirkungen aufgedeckt werden. Die Aufdeckung horizontaler Wechselwirkungen durch die gleichzeitige Berücksichtigung der erklärenden Variablen in jeder Aufteilungsstufe ist jedoch ebenso wichtig.

Dadurch könnte außerdem vermieden werden, daß die bei der Kontrastgruppenanalyse auf einer bestimmten Aufteilungsstufe „beinahe“ verwendeten Faktoren unberücksichtigt bleiben.

In einem späteren Aufsatz wird durch eine Kombination der Kontrastgruppenanalyse mit der Regressionsanalyse ein Lösungsweg für die Aufdeckung von horizontalen Wechselwirkungen aufgezeigt.

Anhang

1. Ableitung der Variationszerlegungsformel für den Fall einer quantitativen abhängigen Variablen.

Ausgehend von Formel (6) erhält man für die Gesamtvariation folgende Darstellung:

$$\text{(I)} \quad \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\mu)^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left[(\hat{Y}_{ij}-\mu)+\hat{\varepsilon}_{ij}\right]^2$$

Die rechte Seite ergibt durch Ausquadrieren folgende Ausdrücke:

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left[(\hat{Y}_{ij}-\mu)+\hat{\varepsilon}_{ij}\right]^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\hat{Y}_{ij}-\mu)^2 +$$

$$+\sum_{i=1}^{I}\sum_{j=1}^{n_i}\hat{\varepsilon}_{ij}^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\hat{Y}_{ij}-\mu)\cdot\hat{\varepsilon}_{ij}$$

Wenn man im gemischten Produkt $\hat{Y}_{ij}$ durch $\hat{\mu}+\hat{\alpha}_i$ ersetzt und ausmultipliziert, so sieht man, daß alle drei Terme des gemischten Produkts verschwinden:

$$\hat{\mu}\cdot\sum_{i=1}^{I}\sum_{j=1}^{n_i}\hat{\varepsilon}_{ij} \quad \text{und} \quad \mu\cdot\sum_{i=1}^{I}\sum_{j=1}^{n_i}\hat{\varepsilon}_{ij}$$

weil die Störkomponenten bei der Methode der kleinsten Quadrate sich ausgleichen und

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}\hat{\alpha}_i\,\hat{\varepsilon}_{ij}$$

weil zwischen den Faktorstufen und der Störvariablen keine stochastische Beziehung zugelassen wird.

Somit läßt sich (I) wie folgt schreiben:

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\mu)^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\hat{Y}_{ij}-\mu)^2 +$$

$$+\sum_{i=1}^{I}\sum_{j=1}^{n_i}\hat{\varepsilon}_{ij}^2$$

Setzt man für

$$\hat{Y}_{ij} = \hat{\mu}+\hat{\alpha}_i = \hat{\mu}_i$$

und für

$$\hat{\varepsilon}_{ij} = Y_{ij}-\hat{Y}_{ij},$$

so erhält man die im Text unter (7) angegebene Zerlegungsformel:

$$\text{(7)} \quad \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\mu)^2 = \sum_{i=1}^{I}n_i\,(\hat{\mu}_i-\mu)^2 +$$

$$+\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu}_i)^2$$

2. Ableitung der Variationszerlegungsformel für den Fall einer qualitativen abhängigen Variablen.

Für die totale Quadratsummenvariation ergibt sich bei der Kontrastgruppenanalyse folgender Ausdruck:

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij}-\mu)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij}^2 -$$

$$-2\sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij}\,\mu + \sum_{i=1}^{2}\sum_{j=1}^{n_i}\mu^2$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij}^2 - 2\,\mu\sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij} + \sum_{i=1}^{2}n_i\,\mu^2$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij} - 2\,\mu\sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij} + \mu^2\sum_{i=1}^{2}n_i$$

$$= np \qquad - 2\,p\,n\,p \qquad + p^2\,n$$

$$= n\,(p-p^2) = np\,(1-p).$$

Für die Zwischengruppenvariation folgt:

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\mu_i-\mu)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(p_i-p)^2$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{n_i}p_i^2 - 2\,p\sum_{i=1}^{2}\sum_{j=1}^{n_i}p_i + \sum_{i=1}^{2}\sum_{j=1}^{n_i}p^2$$

$$= \sum_{i=1}^{2}n_i\,p_i^2 - 2\,p\,n\,p + p^2\cdot n$$

$$= \sum_{i=1}^{2}n_i\,p_i^2 - n\,p^2.$$

Aus der Quadratsumme innerhalb der Gruppen folgt:

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij}-\mu_i)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij}^2 -$$

$$-2\sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij}\,p_i + \sum_{i=1}^{2}\sum_{j=1}^{n_i}p_i^2$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij} - 2\sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij}\,p_i + \sum_{i=1}^{2}n_i\,p_i^2$$

$$= np - 2\sum_{i=1}^{2}p_i\cdot n_i\,p_i + \sum_{i=1}^{2}n_i\,p_i^2$$

$$= np - \sum_{i=1}^{2}n_i\,p_i^2.$$

Somit lautet der Zerlegungssatz im Falle einer qualitativen abhängigen Variablen:

$$\text{(9)} \quad np\,(1-p) = \left(\sum_{i=1}^{2}n_i\,p_i^2 - np^2\right) +$$

$$+\left(np - \sum_{i=1}^{2}n_i\,p_i^2\right)$$